International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Fast Simple Linear (FaSL) Unsupervised Feature Extraction Method

Karteeka Pavan Kanadam[1] , G.L.N.JayaPrada[2] , Jeevanajyothi Pujari[3], Hymavathi Thottathyl[4]

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: The increase in volume of high-dimensional data necessitates the use of dimensionality reduction strategies (DRS), which reduce dimensions and extract meaningful insights by eradicating irrelevant features. Linear and nonlinear are the two types in DRS. Nonlinear dimensionality reduction methods have gained considerable popularity in recent years due to their effectiveness in handling real-world datasets with complex nonlinear structures. However, there are some fields where linear data sets are frequently used, including physics, economics, health informatics, social sciences, etc. The major drawback of many existing linear and nonlinear DRS models is their computationally expensive nature. To address this issue, a fast, simple, linear (FaSL) unsupervised feature extraction method is proposed using descriptive statistics. The FaSL performance is evaluated by applying clustering on various benchmark data sets and compared with five linear state-of-the-art methods. The experimental results demonstrate that FaSL outperforms other linear models such as PCA, LDA, LPP, ICA, and FA in terms of accuracy and computation time. The average accuracy improvement of FaSL over PCA, LDA, LPP, ICA, and FA is, in order, 3.4, 9.2, 5.67, 3.97, and 0.075 while reducing computational time by 2.26, 3.1, 1.29, 7.58, and 6.2 times, respectively.

Keywords: *Dimensionality reduction, linear, nonlinear, clustering, PCA, LDA*

1. Introduction

Massive data sets are now a common phenomenon in machine learning problems due to recent developments in data collection techniques. These sets require substantial processing time and resources, and the learning algorithm's efficiency declines with incorrect, irrelevant, and noisy features [1-2]. To increase the effectiveness and efficiency of data mining approaches, dimensionality reduction strategies (DRS) are used to reduce data set dimension by eliminating irrelevant, noisy data. These can be classified into feature extraction and feature selection methods [2-3]. These methods have been used successfully in a variety of real-world applications; including image processing, object detection, video in processing, disease detection, stock analysis etc. [4-7]. Feature selection methods select most relevant attribute subset from the original attributes of the data. Feature extraction methods transform original data into reduced meaningful information which is a combination of all features [8-9]. These can be categorized into supervised, semi supervised and as unsupervised based on whether or not the methods utilize true labels. Among these unsupervised methods are most challenging without the utilization of domain knowledge [10].

Also, DRS can be classified into linear and nonlinear models. Linear DRS assumes linear intrinsic structure, which may not be suitable for real-world datasets with nonlinear intrinsic structures [11]. Nonlinear DRS are become interesting and burning topic in machine learning research for more than a decade, with numerous proposed methods proved to be effective for the selected applications.

Also, DRS can be classified into linear and nonlinear models. Linear DRS assumes linear intrinsic structure, which may not be suitable for real-world datasets with nonlinear intrinsic structures [11]. Nonlinear DRS are become interesting and burning topic in machine learning research for more than a decade, with numerous proposed methods proved to be effective for the selected applications.

In fact, there are many real-world datasets with linear relationshipsbetween the variables, and they may be found in a variety of disciplines, such as physics, economics, finance, and the social sciences [12–19]. In 2017, Meier, A., and Kramer analyzed 29 DR methods on 13 datasets and came to the conclusion that MDS, GPLVM, and PCA performed better among all 29 DRS [20]. Maateen et al., has experimented the linear model PCA performance with 13 nonlinear data models on various datasets and reported that PCA, a linear model outperformed on most of the natural data sets compared to 13 various nonlinear DR models [21]. Though PCA is a traditional linear unsupervised DR method, it is found in all studies of DR methods.

[1]*Professor,Department of Computer Applications, R.V.R. J.C. College of Engineering, Guntur, AP, India kkp@rvrjc.ac.in ORCID: 0000-0002-0784-6928*
[2]*Associate Professor,Department of CSE (AI & ML), Mallareddy College of Engineering and Technology, Hyderabad gjayaprada74@gmail.com*
[3]*Assistant Professor, Department of Database Systems School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, Vellore, jeevanajyothi.pujari@vit.ac.in*
[4]*Assistant Professor, Department of Computer Applications, R.V.R. J.C. College of Engineering, Guntur, AP, India thottathylhyma@gmail.com, ORCID:0000-0003-0480-3356*

To date, PCA, the linear DR is a widely used algorithm in feature extraction applications [22]. Despite the nonlinear model's success; linear models are preferred strongly on some real-world data sets and linear models are better in solving the classification problems compared to nonlinear models [23]. Hence, till date researchers are focusing on linear DR models.

However, when data dimensionality increases, the computational cost of typical dimensionality reduction algorithms, including linear models (PCA and LDA), increases exponentially, making computation increasingly challenging [24]. Also, none of the feature selection methods produce a more informative combination of characteristics than the original data, and feature extraction methods do not transform a combination into a representation of the original features.

In view of these, in this paper, simple statistical features are used to propose a Fast, Simple and Linear unsupervised feature extraction method (FaSL). The FaSL reduces dimensions of a given data set of size m×n into m×4 without any domain knowledge. The method's performance is examined using k-means clustering. Relative performance of the proposed FaSL is studied with various linear methods using different validity measures, error rate, and computational cost on 6 real-life and 7 simulated data sets. The results of the FaSL are enterprising with higher accuracy and lower computation times. The average accuracy improvement of FaSL over PCA, LDA, LPP, ICA, and FA is, in order, 3.4, 9.2, 5.67, 3.97, and 0.075. On average, the reduction in computational time of FaSL over PCA, LDA, LPP, ICA, and FA, respectively, 2.26, 3.1, 1.29, 7.58, and 6.2 times.

The remaining part of the paper is organized as follows, section-2 contains main contributions of the paper, section-3presents the proposed FaSL methodology, experimental results and discussions are discussed in section-4 and conclusions and possible future enhancements are provided in section-5.

2. RELATED WORK

Dimensionality reduction methods are the fundamental requirement in the pre-processing phase of many data analysis models with huge dimensional data sets. DRS are used to reduce dimensions either by selection or transforming overall features into set of features for further analysis. DRS are divided into feature selection and feature extraction methods. The literature contains a large number of feature selection/extraction algorithms.

Featureselection aims to choose a portion of the original features by eliminating redundant / irrelevant information. Feature selection techniques are divided into supervised, unsupervised, and semi-supervised categories according to the labels that are specified in the data. Supervised feature selection algorithms select subset of features by assessing their association with class labels or their performance in prediction [25-26]. Unsupervised feature selection techniques select features based on data variance or distribution [27-28]. Semi-supervised featureselection methods used the labelled data as additional details to improve the performance of the method [29-30].

Feature extraction methods reduce the dimension of the original data set by transforming features into combination of all original features. Among these unsupervised are more challenging due to the ignorance of true labels. Existing unsupervised feature extraction methods are divided into linear and nonlinear models [31]. Since the paper focus on the linear models, the existing few popular linear feature extraction models are:

2.1 Principal Component Analysis (PCA)

Define abbreviations and acronyms the first time they are used in the text, even after they have already been defined in the abstract. Abbreviations such as IEEE, SI, ac, and dc do not have to be defined. Abbreviations that incorporate periods should not have spaces: write "C.N.R.S.," not "C. N. R. S." Do not use abbreviations in the title unless they are unavoidable (for example, "IEEE" in the title of this article).

2.2 Linear Discriminant Analysis (LDA)

The goal of LDA is to identify a feature subspace that maximizes group separability [34]. The objective of classical LDA is to maximize the trace ratio value of the between-class scatter matrix and within-class scatter matrix in the subspace, such that the points within the same class will be drawn together and the points between different classes will be kept as far as possible. It is a supervised linear feature extraction method.

2.3 Factor Analysis (FA)

This method creates a common score by taking the most common variance out of all the variables. We can use this score as an index of all the variables for further analysis [35].

2.4 Locality Preserving Projection (LPP)

Locality Preserving Projection [36] is a linear approximation of LaplacianEmbedding. The goal of LPP is to project the original data while keeping nearest neighbour relationships.

2.5 Independent Component Analysis (ICA)

In addition to PCA and other linear transformations, Independent Component Analysis is a popular technique. The givendata is described by ICA as a combination of unknown and independent sources. [37]. ICA is a higher-

order approach that looks for linear projections that are nearly statistically independent and are not always orthogonal to one another.

Though the aforementioned methods each provide a reduced set with either original or extracted / transformed features, but not completely meet the requirements. Most of the current research is on proposing nonlinear models since most of the real-world datasets are nonlinear by structure. But the linear models are better in solving the classification problems compared to nonlinear models [38]. Hence, researchers are focusing on linear DR models and the most widely used linear models (PCA, LDA) computational complexity increases exponentially with the increase of data dimensions.

Hence the paper confines to develop a fast, simple, linear unsupervised feature extraction method. The proposed method performance is evaluated on various real and synthetic datasets with the state of art of the linear reduction methods.

Contributions

The major contributions of this paper are

i. A new unsupervised feature extraction method using descriptive statistics is proposed for dimensionality reduction.
ii. The method is simple, lower computational cost with improved accuracy.
iii. A drastic reduction in dimension of the high dimensioned data sets is achieved in lower computational time.
iv. The performance of FaSL is evaluated with application of clustering using 5 various linear state-of-the-art methods, 6 different validity measures on 6 real-life and 7 simulated data sets.

3. A Fast, Simple, Linear (Fasl) Unsupervised Feature Extraction Method

Let X = {X_1, . . ., X_m} be a data set in a n-dimensional Euclidean space R_n, where m is no of samples and n is no of dimensions. Each sample X_i is a vector with n features and X_j is a column vector represents j^{th} feature values of all m samples. The proposed method extracts four features from the data set. Each feature is a Squared Euclidean distance of each sample to mean, maximum, minimum and sum of the selected data set. The following are the key steps.

i. Acquire high dimensional benchmark data sets of various domains.

ii. Pre-process the dataset

iii. Extract Features: Each feature is a Squared Euclidean distance of each sample to mean maximum, minimum and sum of the selected data set.

The features are:

- Feature F1 $= \left\| d(X_i, X_{mean)} \right\|$where Xmean = [Xmean1...Xmeanj, .Xmeann];
  $X_{meanj=} \sum_{i=1}^{m} X_{ij}/m$, is the mean of the elements of column j.
- Feature $F2 = \left\| d(X_i, X_{max)} \right\|$, where Xmax=[Xmax1, .., Xmaxj, ..Xmaxn]; Xmaxjis maximum of jth column/feature
- Feature $F3 = \left\| d(X_i, X_{min)} \right\|$, whereXmin=[Xmin1, .., Xminj, ..Xminn]; Xminj is minimum of jth column
- Feature $F4 = \left\| d(X_i, X_{sum)} \right\|$,where Xsum= [Xsum1, .., Xsumj, ..Xsumn]; $X_{sumj=\sum_{i=1}^{m} x_{ij}}$, is the sum of elements of column j.

iv. Apply k-means clustering method

The proposed model transforms $m \times n$ dataset into $m \times 4$ size, where m, n are number of samples and features. The steps of the proposed method depicted in the figure1.

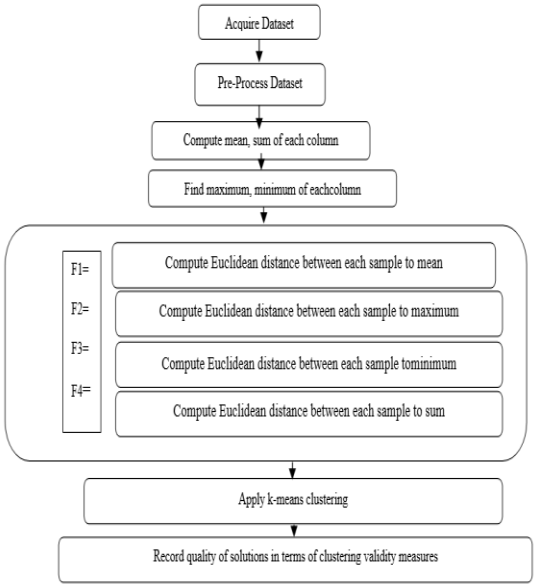

Fig.1. Fast, Simple, Linear (FaSL) unsupervised feature extraction method

For example, assume the following dataset of size 8X6.

For example, assume the following dataset of size 8X6.

Dataset (8X6) =

0.384465 0.968647 0.042657 0.937293 0.250838 0.409011

0.198445 0.960714 0.044687 0.921429 0.122681 0.220098

0.400157	0.968434	0.041942	0.936867	0.262813	0.423814
0.442284	0.971203	0.041103	0.942406	0.295956	0.461889
0.390355	0.968358	0.041729	0.936717	0.254919	0.411407
0.503066	0.974912	0.04193	0.949825	0.347458	0.52008
0.462264	0.972794	0.038835	0.945589	0.312322	0.480125
0.471215	0.972293	0.037657	0.944586	0.32035	0.493014

Mean of the Dataset=

0.406531	0.969669	0.041317	0.939339	0.270917	0.42743

Squared Euclidean distance between each element of dataset to mean (represented in a row) is as follows: New Dataset (8X1)=

0.0352	0.3297	0.0113	0.0557	0.028	0.1546	0.0875	0.1048

4. Experimental Results and Discussions

This section focused on performance assessment of the proposed FaSL method for unsupervised problems. A total of 6 publicly accessible data sets and seven simulated are used to study the performance. The number of features of the datasets is vary from 4 to 1024 and the number of samples is varied from 150 to 5473. The results are compared with linear unsupervised feature extraction methods and LDA, which is supervised linear model. The linear models are: Principal Component Analysis (PCA), Independent Component Analysis (ICA), Factor Analysis (FA), Locality Preserving Projection (LPP), and Linear Discriminant Analysis (LDA). The relative performance of the proposed is evaluated using Clustering (k-means) method.

K-means clustering algorithm is used for assessment of quality after reduction of the dimensions. For each data set, for each DR method k-means is run 30 times and the average values are taken for evaluation. k-value in k-means is a priori and is considered as the number of true classes of the data set in all experiments. The quality of the clustering solutions is measured using the external and internal cluster validity measures: Rand Index (RI), Jaccard coefficient (Jacc), Fowlkes-Mallow's index (FM), Normalized Mutual Information (NMI), Normalized Variation of Information (NVI), Davis Bouldin (DB), Silhouette (Sil) index [39-42]. Error rate is also computed in all experiments. In addition to all these measures, execution time of each dimensionality reduction method for each dataset is considered for evaluation of results.

Table-1 shows details of datasets considered for experimentation. The table contains dataset name, number of samples of the dataset, number of classes and from where the dataset is downloaded in order. Observed mean values of all validity measures in 30 independent runs of k-means on each data set are presented inTable-2. For each validity measure, if higher values are better, then represented with "↑", otherwise i.e., lower values are most preferable then indicates with "↓". Table-3 shows elapsed time noticed for 30 independent runs of k-means by each algorithm for each dataset after reduction of the dimensions.

Table 1. Details of datasets

Dataset	#Samples	#Features	Classes	Source: dataset downloaded
Boston Housing	506	14	3	Kaggle database [43]
Pageblocks	5473	10	5	UCI Machine learning repository [44]
Ionosphere	351	34	2	
Yale	165	1024	15	
iris	150	4	3	
wine	178	13	3	
Demo_Data	1000	21	3	Nguyen X. Vinh et al. (2014) [45]
highdim1	1024	32	16	High dimensional datasets [46-47]
highdim2	1024	64	16	
highdim3	1024	128	16	
highdim4	1024	256	16	
highdim5	1024	512	16	
highdim6	1024	1024	16	

Table 2. Mean Values in 30 independent runs of k-means

Dataset / Method	Metric	Original	PCA	LDA	LPP	ICA	FA	FaSL
	RI ↑	0.56	0.56	0.48	0.58	0.57	0.62	0.54

Boston	JAC↑	0.32	0.32	0.31	0.32	0.32	0.31	0.28
	FM↑	0.49	0.49	0.49	0.48	0.48	0.48	0.44
	DB ↓	0.62	0.6	0.93	0.53	0.66	0.44	0.57
	Sil↑	0.73	0.73	0.54	0.75	0.7	0.76	0.66
	NMI ↑	0.13	0.13	0.05	0.13	0.13	0.18	0.07
	NVI↓	0.92	0.92	0.97	0.93	0.92	0.89	0.96
	Error↓	74.95	74.95	70.84	74.67	74.63	45.68	66.52
	RI ↑	0.33	0.6	0.43	0.59	0.6	0.57	0.59
	JAC↑	0.21	0.58	0.34	0.56	0.58	0.54	0.57
Pageblocks	FM↑	0.43	0.73	0.56	0.72	0.73	0.71	0.73
	DB ↓	1.3	0.54	1.09	0.52	0.54	0.46	0.47
	Sil↑	0.36	0.83	0.37	0.82	0.83	0.84	0.85
	NMI↑	0.13	0.05	0.07	0.05	0.05	0.05	0.05
	NVI↓	0.94	0.97	0.96	0.97	0.97	0.97	0.97
	Error↓	73.79	27.26	54.84	28.26	27.26	29.85	28.04
	RI ↑	0.58	0.58	0.5	0.57	0.58	0.5	0.6
ionosphere	JAC↑	0.43	0.43	0.39	0.42	0.43	0.34	0.44
	FM↑	0.6	0.61	0.56	0.59	0.6	0.59	0.61
	DB↓	1.51	1.5	1.58	1.41	1.36	1.61	0.77
	Sil↑	0.41	0.41	0.43	0.45	0.45	0.52	0.68
	NMI ↑	0.13	0.12	0	0.1	0.13	0.12	0.16
	NVI↓	0.92	0.93	0.99	0.94	0.92	0.97	0.91
	Error↓	28.83	29.29	55.23	30.36	28.86	28.81	27.35
	RI ↑	0.88	0.88	0.81	0.89	0.88	0.89	0.89
Yale	JAC↑	0.14	0.13	0.0526	0.16	0.13	0.15	0.14
	FM↑	0.25	0.24	0.1102	0.28	0.24	0.27	0.25
	DB↓	1.78	1.81	2.107	1.74	1.76	1.87	0.74
	Sil↑	0.16	0.16	0.0348	0.17	0.16	0.26	0.49
	NMI↑	0.46	0.46	0.2469	0.5	0.45	0.49	0.47
	NVI↓	0.69	0.7	0.86	0.66	0.7	0.66	0.68
	Error↓	87.8	87.8	92.3	88.16	87.3	86.02	83.3
	RI ↑	0.68	0.68	0.54	0.7	0.68	0.56	0.69
	JAC↑	0.38	0.38	0.25	0.4	0.38	0.39	0.4
Wine	FM↑	0.55	0.55	0.41	0.57	0.55	0.56	0.57
	DB↓	0.55	0.56	0.87	1.05	0.55	0.52	0.54
	Sil↑	0.69	0.69	0.52	0.49	0.7	0.68	0.72
	NMI↑	0.32	0.34	0.07	0.4	0.34	0.4	0.39
	NVI↓	0.8	0.79	0.96	0.74	0.79	0.72	0.75
	Error↓	33.4	32.5	53.37	35.674	32.35	35.18	30.91
	RI ↑	0.87	0.89	0.89	0.89	0.92	0.87	0.86
	JAC↑	0.73	0.77	0.77	0.77	0.82	0.73	0.65
Iris	FM↑	0.84	0.86	0.86	0.86	0.89	0.83	0.79
	DB↓	0.46	0.46	0.58	0.49	0.31	0.31	0.58
	Sil↑	0.8	0.8	0.75	0.78	0.86	0.85	0.68
	NMI↑	0.79	0.81	0.81	0.81	0.84	0.78	0.66
	NVI↓	0.33	0.3	0.3	0.3	0.27	0.34	0.49
	Error↓	19.13	14.91	15.31	15	8.73	17.97	11.6
	RI ↑	0.67	0.67	0.66	0.67	0.66	0.65	0.57

	JAC↑	0.33	0.34	0.33	0.34	0.33	0.31	0.22
Demo_Data	FM↑	0.5	0.51	0.5	0.51	0.5	0.47	0.36
	DB↓	1.33	1.34	1.25	1.34	1.01	0.49	1
	Sil↑	0.42	0.42	0.46	0.42	0.53	0.73	0.48
	NMI↑	0.37	0.37	0.37	0.37	0.37	0.27	0.03
	NVI↓	0.76	0.76	0.77	0.76	0.77	0.84	0.97
	Error↓	70.98	69.23	70.1	68.91	70.4	55.52	60.2
	RI ↑	0.96	0.96	0.96	0.96	0.96	0.95	0.96
	JAC↑	0.62	0.64	0.64	0.64	0.64	0.56	0.64
highdim1	FM↑	0.78	0.79	0.79	0.79	0.79	0.73	0.79
	DB ↓	0.76	0.81	0.75	0.87	0.78	0.26	0.44
	Sil↑	0.66	0.67	0.66	0.67	0.68	0.83	0.78
	NMI↑	0.92	0.92	0.92	0.92	0.92	0.87	0.91
	NVI↓	0.14	0.13	0.13	0.135	0.13	0.22	0.16
	Error↓	75.13	71.46	71.59	78.27	74.71	71.33	70.65

Table3. Time (Elapsed time) taken in seconds for 30 independent runs of k-means

Method ⇨ / **Dataset**	**Original**	**PCA**	**LDA**	**LPP**	**ICA**	**FA**	**FaSL**
BOSTON	11.57	11.8	15.46	10	13.09	11.3	9.9
Pageblocks	1514.2	1304.17	1807.66	1435.43	1303.06	1035.57	1328.53
ionosphere	9.11	8.59	9.18	23.28	8.22	13.55	5.66
Yale	21.69	4.44	28.18	7.55	269.37	138.18	3.85
wine	3.04	2.27	2.53	2.77	3.15	6.17	2.27
Iris	2.37	1.44	2.2	2.09	1.87	2.93	1.51
Demo_data	50.99	35.5	34.18	34.6	38.11	32.84	53.94
highdim1	64.33	72.7	49.58	43.58	84.02	57.58	61.77
highdim2	89.62	93.2	83.88	65.82	73.53	49.93	50.28
highdim3	92.33	90.74	94.91	50.99	67.07	55.89	52.62
highdim4	145.62	152.48	124.56	53.45	109.24	125.59	57.64
highdim5	273.58	251.3	260.86	45.45	233.82	865.66	56.86
highdim6	608.45	534.5	752.43	39.3	606.82	793.62	52.81

5.1 Discussions

From Table-2, FaSL shows its superiority for Ionosphere, Yale, wine, highdim1, highdim3 and for highdim6 data sets. Fasl is equally good with PCA, and ICA for iris, highdim4 and Page blocks datasets. But, in case of Boston Housing dataset, Demo_data, highdim2, and for highdim5 Factor Analysis perform better compared to all other linear models.

`Friedman's test can be used to compare the quality of some m-algorithms on n-datasets. If the performance of the algorithms varies, applying Friedman's test rejects the null hypothesis that "all algorithms have equal performance"[34]. Then the Nemenyi test can be applied to identify which algorithms differ substantially from one another [35]. The diagrams for each validity measure are shown in Figure 1 along with the mean rank of each algorithm in comparison (lower ranks to the left). As per Nemenyi test, thick line connects groups of algorithms that, are not significantly different from one another. In each subfigure, the critical difference (CD) is also displayed above the axis.

From Fig. 1 (a), it is observed that the proposed FaSL is better than PCA, LPP, and LDA and is comparable with ICA and FA in terms of DB. FaSL is likely good with PCA, ICA, and FA and is better compared to LPP and LDA in terms of error rate from figure 1 (b). From 1(c), (d), (g), (i), FaSL is equally compatible with PCA, LPP, ICA and is better compared to FA and LDA in terms of FM, JAC, RI and time. From 1(i) FaSL is far better than all algorithms in terms of time, compatible with LPP, PCA, ICA but significantly better than LDA, FA. From 1(h) FaSL is compatible with PCA, ICA and FA and better than LPP, LDA.

From Fig. 1 (a), it is observed that the proposed FaSL is better than PCA, LPP, and LDA and is comparable with ICA and FA in terms of DB. FaSL is likely good with PCA, ICA, and FA and is better compared to LPP and LDA in terms of error rate from figure 1 (b). From 1(c) , (d), (g), (i) FaSL is equally compatible with PCA, LPP, ICA and is better compared to FA and LDA in terms of FM, JAC, RI and time. From 1(i) FaSL is far better than all algorithms in terms of time, compatible with LPP, PCA, ICA but significantly better than LDA, FA. From 1(h) FaSL is compatible with PCA, ICA and FA and better than LPP, LDA

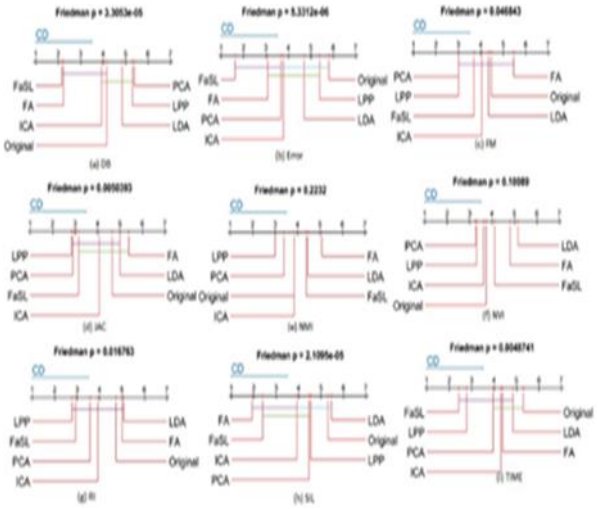

Fig.1.CD diagrams after Nemenyi test for each evaluation criterion in order (a)DB (b)Error (c)FM (d)JACC (e)NMI (f)NVI (g) RI (h) SIL (i)Time

5.1.1. Relative Performance of selected linear models VsFaSL

FaSL is proposed to reduce dimensions (to 4) as much as possible of the high dimensioned dataset(s) such that improves efficiency of the learning algorithm (unsupervised \supervised) in finding appropriate groups in significantly reduced time. Hence, the section focuses on evaluating performance of FaSL in terms of time and error rate.

(a) Performance Evaluation of FaSL in terms of Time

FaSL reduce any dataset dimensions to 4, the time it required to complete 30 independent runs of k-means is presented in Table-3 along with other linear DR methods. Table-3 demonstrates that FaSL finds clusters using k-means in very less time for Ionosphere, Yale, wine, highdim2, highdim3, highdim4, highdim5 and for highdim6. Drastic time reduction can be noticed on datasets with a greater number of dimensions ie., for example Demo_data and highdim6. Details of observations in the difference in mean timings of each algorithm in comparison with the FaSL are presented as follows.

PCA (widely used DR method): Among the data sets PCA find clusters using k-means in less time for only two data sets (Pageblocks, Demo_data) compared to FaSL. Whereas FaSL drastically reduce the time for the datasets with higher dimensions (>25). The difference is very much attractive and significant. In highdim6 (1024 dimensions), FaSL finds results in 10times reduced time compared to PCA. PCA takes four times of FaSL time in case of highdim5.

LDA (Supervised): converge k-means in less time for only two datasets: Demo_data and highdim1, but the difference may be negligible. In case of highdim6, LDA takes 14 times more time compared to FaSL

LPP: Though it takes less time compared to FaSL for high dimensioned datasets, but it may be negligible difference and the FaSL is equally good with LPP. LPP takes 4 times more time for page block dataset compared to FaSL

ICA: It is noticed that ICA recorded lower timing compared to FaSL in case of two datasets, Pageblocks and Demo_data. But the difference is very negligible and not significant. Whereas in case of dataset, highdim6, ICA takes 11 times more time compared to FaSL. In case of Yale, ICA takes 69 times more time compared to FaSL.

FA: FaSL is compatible with FA in terms of time in all datasets except in Yale, highdim4, highdim5 and highdim6. In case of Yale FA takes 35 times more timecompared to FaSL. In case ofhighdim5 and highdim6 FA requires 15 times more time compared to FaSL.

Compared to the above all linear models FaSLis either far better or equally good in terms of time.

(b) Performance Evaluation of FaSL in terms of Error rate:

Any algorithm must have the qualities simplicity, low computing complexity, and ability to provide accurate output. Quality of clustering is measured in this work in number of ways using various evaluation criterion and reported in table-2 and figure-1. In addition to these, accuracy is studied in terms of error rates, and is computed in each experiment of each dataset and mean error rates are reported in table-2.

The following are the few observations in each case of algorithm compared to FaSL in terms of error rate.

PCA (widely used): PCA produce results more accurately in only two cases, Pageblocks and highdim4. But are very negligible difference with FaSL, 0.7857 and -0.027. Hence, for each dataset (Boston Housing, Yale, Demo_data, highdim2, highdim3, highdim6) FaSL is either far better or equally good (Pageblocks, ionosphere, wine, iris, highdim1, highdim4, highdim5) with PCA.

LDA (Supervised): Noticed that FaSL outperforms compared to LDA almost in each dataset

LPP (Local Projection based): Observed that FaSL is far better in each case of dataset compared to LPP in terms of error rate.

ICA: Noticed form table-2 ICA is better in only two cases (Page blocks, iris). But in case of Pageblocks the difference is negligible i.e., 0.7. But remaining all cases FaSL shows its supremacy with lower error rates with increased accuracy.

FA: FA is better in three cases, (Boston Housing, Demo_data, highdim2 and highdim5), but in remaining 9 cases FaSL is better than FA.

The average accuracy improvement of FaSL in terms of error rate over PCA, LDA, LPP, ICA, and FA is 3.4, 9.2, 5.67, 3.97, and 0.075 in order.

5.2 Limitations

Fasl is a simple linear DR method developed using simple statistics methods; represent characteristic of a collection such as mean, maximum, minimum, sum and Euclidean distance, better captures hidden differences. Being the proposed linear or separable bench mark data sets are browsed and selected from well-known public data sites. For the experimentation only popular linear DR feature extraction methods areselected. Experimental results are reported by running k-means 30 times. FaSL shows its superiority compared to the popular, widely used linear DRS on various datasets. However, like all other DRS, FaSL is also may not be suitable to the any kind of dataset.

Conclusion

In view of 'curse in dimensionality' in many data analysis methods numerous dimensionality reduction methods are evolved as linear and nonlinear. As linear methods assume the intrinsic structure of dataset as linear may not be applicable for most of the real-world datasets. But there exists some disciplines Physics, Economics, Medical, social science which generates real word data with linear relationships among attributes. Also, some of the nonlinear models may not satisfy the classification or clustering applications. Hence, still PCA, LDA are the widely used linear unsupervised, supervised models. All these DR models are computationally expensive. Hence, a fast, simple, linear (FaSL) unsupervised feature extraction method using descriptive statistics is proposed. FaSL is very simple and reduce any dataset size to four dimensions. For experimentation 6 real world and 7 synthetic benchmark data sets are taken from popular public data bases. Performance of FaSL is evaluated by applying on k-means clustering, results are evaluated with 5 widely used various linear DR models, PCA, LDA, LPP, ICA, FA. Relative performance is studied and quality of clustering is assessed using 7 different validity measures along with error rate and elapsed time. Experimental results have shown that FaSL is either far better on most of the selected datasets or equally good on few datasets compared to the linear models in the experimentation. FaSL is simple, fast enough and more accurate compared to PCA, LDA, LPP, ICA and FA on the selected datasets. In case of highdim6 (1024 dimensions), FaSL is 10times, 14 times, 11 times, 15 times faster compared to PCA, LDA, ICA and FA in order. In average FaSL is 2.26, 3.1, 1.29, 7.58, 6.2 times faster than PCA, LDA, LPP, ICA and FA in order in terms of time. The average accuracy improvement of FaSL in terms of error rate over PCA, LDA, LPP, ICA, and FA is 3.4, 9.2, 5.67, 3.97, and 0.075 in order.

Being a linear model, the FaSL, linear or separable public datasets and linear DRS are only included in the experimentation. Like any linear DR method, FaSL may not satisfy real world data with nonlinearity. Hence, extending the proposed FaSL to identify the hidden intrinsic local structures of the datasets and make the enhanced model more reliable and to suitable to nonlinear data sets is the future endeavor.

References

[1] H. Liu and H. Motoda, Computational Methods of Feature Selection. Boca Raton, FL, USA: CRC Press, 2007.

[2] H. Li, T. Jiang, K. Zhang, Efficient and robust feature extraction by maximum margin criterion, IEEE Transactions on Neural Networks 17 (1) (2006) 157–165.

[3] J. Yan, B. Zhang, N. Liu, S. Yan, Q. Cheng, W. Fan, Q. Yang, W. Xi, Z. Chen, Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing, IEEE Transactions on Knowledge and Data Engineering 18 (3) (2006) 320–333

[4] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," IEEE Trans. Image Process., vol. 27, no. 1, pp. 38–49, Jan. 2018.

[5] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," IEEE Trans. Neural Netw., vol. 14, no. 1, pp. 195–200, Jan. 2003.

[6] H.Gunduz, An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification, Biomedical Signal Processing and Control 66 (2021) 102452.

[7] J. Han, Z. Ge, Effect of dimensionality reduction on stock selection with cluster analysis in different market situations, Expert Systems with Applications 147 (2020) 113226.

[8] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, vol. 454. Berlin, Germany: Springer, 2012.

[9] J. P. Cunningham and Z. Ghahramani, "Linear dimensionality reduction: Survey, insights, and generalizations," J. Mach. Learn. Res., vol. 16, pp. 2859–2900, 2015.

[10] Wei, Qin Yue, Kai Feng, Junbiao Cui, and Jiye Liang, "Unsupervised Dimensionality Reduction Based on Fusing Multiple Clustering Results", IEEE Transactions on Knowledge and Data Engineering, VOL. 35, NO. 3, 32w11-3223, 2023.

[11] Ruisheng Ran, Ji Feng, Shougui Zhang, and Bin Fang, "A General Matrix Function Dimensionality Reduction Framework and Extension for Manifold Learning", IEEE Transactions on Cybernetics, vol.52, no.4, 2137-2148, 2022.

[12] Smita Rath, Alakananda Tripathy, Alok Ranjan Tripathy, Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model.

[13] Diabetes& Metabolic Syndrome: Clinical Research & Reviews, Volume 14, Issue 5, 2020, Pages 1467-1474, ISSN 1871-4021.

[14] Davide Festa, Alessandro Novellino, Ekbal Hussain, Luke Bateson, Nicola Casagli, Pierluigi Confuorto, Matteo Del Soldato, Federico Raspini, Unsupervised detection of InSAR time series patterns based on PCA and K-means clustering, International Journal of Applied Earth Observation and Geoinformation, Volume 118, 2023, 103276, ISSN 1569-8432.

[15] Qingsong Xiong, Haibei Xiong, Qingzhao Kong, Xiangyong Ni, Ying Li, Cheng Yuan, Machine zearning-driven seismic failure mode identification of reinforced concrete shear walls based on PCA feature extraction, Structures, Volume 44, 2022, Pages 1429-1442, ISSN 2352-0124.

[16] Nai Xue Zhang, Yuzhong Zhong, Songyi Dian, Rethinking unsupervised texture defect detection using PCA, Optics and Lasers in Engineering, Volume 163, 2023, 107470, ISSN 0143-8166.

[17] Ying chao Huang, Abdul Bais, A novel PCA-based calibration algorithm for classification of challenging laser-induced breakdown spectroscopy soil sample data, Spectrochimica Acta Part B: Atomic Spectroscopy, Volume 193, 2022, 106451, ISSN 0584-8547.

[18] Iqbal H. Sarker, Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science (2021) 2:160.

[19] Fa Zhu, Junbin Gao, Jian Yang, Ning Ye, Neighborhood linear discriminant analysis, Pattern Recognition, Volume 123, 2022, 108422, ISSN 0031-3203

[20] Shengkun Xie, Feature extraction of auto insurance size of loss data using functional principal component analysis, Expert Systems with Applications, Volume 198, 2022, 116780, ISSN 0957-4174.

[21] Meier, A., Kramer, O. (2017). An Experimental Study of Dimensionality Reduction Methods. In: Kern-Isberner, G., Fürnkranz, J., Thimm, M. (eds) KI 2017: Advances in Artificial Intelligence. KI 2017. Lecture Notes in Computer Science (), vol 10505. Springer, Cham. https://doi.org/10.1007/978-3-319-67190-1_14.

[22] L. van der Maaten, E. Postma, J. van den Herik, "Dimensionality reduction: A comparative review," tech. rep., Tilburg University, Netherlands, 2009. Tech. report TiCC TR 2009-005.

[23] Rizgar R. Zebari, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Dilovan Asaad Zebari, Jwan Najeeb Saeed, "A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction", Journal of Applied Science and Technology Trends, Vol. 01, No.

[24] 02, pp. 56 –70 (2020).

[25] Haozhe Xie, Jie Li, Qiaosheng Zhang, Yadong Wang, "Comparison among dimensionality reduction techniques based on Random Projection for cancer classification, Computational Biology and Chemistry, Volume 65, 2016, Pages 165-172 ISSN 1476-9271

[26] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," IEEE Trans Knowl. Data Eng., vol. 25, no. 3, pp. 619–632, Mar. 2013.

[27] D. Wang, F. Nie, and H. Huang, "Global redundancy minimization for feature ranking," IEEE Trans.

Knowl. Data Eng., vol. 27, no. 10, pp. 2743–2755, 2015.

[28] D. Cai, C. Zhang, and X. He, “Unsupervised feature selection for multi-cluster data,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2010, pp. 333–342.

[29] M. Qian and C. Zhai, “Robust unsupervised feature selection,” in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 1621–1627.

[30] Z. Xu, I. King, M. R.-T. Lyu, and R. Jin, “Discriminative semi supervised feature selection via manifold regularization,” IEEE Trans. Neural Netw., vol. 21, no. 7, pp. 1033–1047, Jul. 2010.

[31] K. Benabdeslem and M. Hindawi, “Efficient semi-supervised feature selection: Constraint, relevance, and redundancy,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 5, pp. 1131–1143, May 2014.

[32] Mishra D, Sharma S (2021) Performance analysis of dimensionality reduction techniques: a comprehensive review. Adv Mech Eng:639–651.

[33] Moore B (1981) Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans Autom Control 26(1):17–32 49.

[34] Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdiscip Rev Comput Stat 2(4):433–459.

[35] Wang S et al (2016) Semi-supervised linear discriminant analysis for dimension reduction and classification. Pattern Recogn 57:179–189.

[36] T. Jolliffe, “Principal component analysis and factor analysis,” in Principal Component Analysis, pp. 115–128, Springer, 1986.

[37] X. He, P. Niyogi, Locality preserving projections, in: Advances in Neural Information Processing Systems, 2003, pp. 153–160.

[38] Comon P (1994) Independent component analysis, a new concept? Signal Process 36(3):287–314 57. Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw 13(4–5):411–430.

[39] J. Cunningham and Z. Ghahramani, “Linear dimensionality reduction: Survey, insights, and generalizations,” JMLR, vol. 16, pp. 2859–2900, 2015.

[40] Rand WM 1971, “Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association,” vol.66, pp.846-850

[41] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” J. Intell. Inf. Syst., vol. 17, no. 2, pp. 107–145, 2001.

[42] N. X. Vinh, J. Epps, and J. Bailey, “Information theoretic measures for clustering’s comparison: Variants, properties, normalization and correction for chance,” J. Mach. Learn. Res., vol. 11, pp.2837–2854, 2010.

[43] Davies DL and Bouldin DW, 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence 1979, vol.1, pp.224-227.

[44] Kaggle: your home for data sciencehttps://www.kaggle.com/datasets/vikrishnan/boston-house-prices?resource=download.

[45] UCI Machine learning repositoryhttps://archive.ics.uci.edu/ml/datasets.php?format=&task=clu&att=&area=&numAtt=greater100&numIns=&type=&sort=nameUp&view=table.

[46] Nguyen X. Vinh, Jeffrey Chan, Simone Romano and James Bailey, "Effective Global Approaches for Mutual Information based Feature Selection". Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'14), August 24-27, New York City, 2014.

[47] P. Fänti and S. Sieranoja K-means properties on six clustering benchmark datasets Applied Intelligence, 48 (12), 4743-4759, December 2018 https://doi.org/10.1007/s10489-018-1238-7.

[48] P. Fränti, O. Virmajoki and V. Hautamäki, "Fast agglomerative clustering using a k-nearest neighbor graph", IEEE Trans. on Pattern Analysis and Machine Intelligence, 28 (11), 1875-1881, November 2006 https://cs.joensuu.fi/sipu/datasets/.

[49] M. Friedman, “A comparison of alternative tests of significance for the problem of m rankings,” Ann. Math. Statist., vol. 11, no. 1, pp. 86–92, 1940.

[50] P. B. Nemenyi, “Distribution-free multiple comparison,” Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 1963.