Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction

Article *in* Expert Systems with Applications · February 2020

DOI: 10.1016/j.eswa.2020.113277

CITATIONS
22

READS
527

6 authors, including:

Mengmeng Li
Zhengzhou University
42 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Haofeng Wang
7 PUBLICATIONS **36** CITATIONS
SEE PROFILE

Lifang Yang
Zhengzhou University
17 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Zhigang Shang
Zhengzhou University
50 PUBLICATIONS **249** CITATIONS
SEE PROFILE

Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction

Mengmeng Li[a,b,†], Haofeng Wang[a,b,†], Lifang Yang[a,b], You Liang[a,b], Zhigang Shang[a,b,c,*], Hong Wan[a,b,c,*]

[a] School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, Henan, China

[b] Industrial Technology Research Institute, Zhengzhou University, Zhengzhou 450001, Henan, China

[c] Henan Key Laboratory of Brain Science and Brain-Computer Interface Technology, Zhengzhou 450001, Henan, China

* Corresponding authors.

† Mengmeng Li and Haofeng Wang contribute equally to this work.

E-mail address: limengmeng1014@163.com (M. Li), hfwang_zzu@163.com (H. Wang), flyer1014@163.com (L. Yang), YouLiangzzu@163.com (Y. Liang), zhigang_shang@zzu.edu.cn (Z. Shang), wanhong@zzu.edu.cn (H. Wan)

Acknowledgements

This work is supported by the National Natural Science Foundation of China, Grant 61673353 and U1304602.

Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction

Highlights

- A fast hybrid dimensionality reduction method for classification is proposed.
- Multi-strategy based feature selection is used to filter out irrelevant features.
- Grouped feature extraction is used to remove redundancy among features.
- The proposed method shows excellent efficiency and competitive classification performance.

ABSTRACT

Dimensionality reduction is one basic and critical technology for data mining, especially in current "big data" era. As two different types of methods, feature selection and feature extraction each have their pros and cons. In this paper, we combine multi-strategy feature selection and grouped feature extraction and propose a novel fast hybrid dimension reduction method, incorporating their advantages of removing irrelevant and redundant information. Firstly, the intrinsic dimensionality of the data set is estimated by the maximum likelihood estimation method. Fisher Score and Information Gain based feature selection are used as multi-strategy methods to remove irrelevant features. With the redundancy among the selected features as clustering criterion, they are grouped into a certain amount of clusters. In every cluster, Principal Component Analysis (PCA) based feature extraction is carried out to remove redundant information. Four classical classifiers and representation entropy are used to

evaluate the classification performance and information loss of the reduced set. The runtime results of different methods show that the proposed hybrid method is consistently much faster than the other three in almost all of the sets used. Meanwhile, the proposed method shows competitive classification performance, which has no significant difference basically compared with the other methods. The proposed method reduces the dimensionality of the raw data fast and it has excellent efficiency and competitive classification performance compared with the contrastive methods.

KEYWORDS

Dimensionality Reduction, Intrinsic Dimensionality, Feature Selection, Feature Cluster, PCA

1. Introduction

We have entered an era of big data with the typical characteristics of large data set size and high dimensionality (Wang, Wang, & Chang, 2016). It brings us huge challenges to extract useful information from massive data. Compared with the difficulty of data reduction, the curse of dimensionality (Golay & Kanevski, 2017; Maeda, 2014) may be more difficult to solve. Generally, there are a large number of irrelevant and redundant features in high-dimensional data set and it increases the difficulty of data processing, knowledge mining, and pattern classification. As the key method to solve this problem, dimensionality reduction can filter out some noise and redundant information by reducing the original high-dimensional space to the low-dimensional intrinsic space (Golay & Kanevski, 2017; Maeda, 2014). On the premise of effectively reducing the dimensionality, it is an effective and reasonable way of dimension reduction to retain the implicit rules or topological structure in the original data space. It helps to extract meaningful insights from the original data set, reduce the complexity of data processing,

release the computational burden of the computer, and also helps to improve the stability and interpretability of the learning model (Wang, et al., 2016). In addition, dimension reduction also provides useful bases for effective and clear data visualization.

In general, dimensionality reduction methods can be divided into two types: feature extraction (Choi, Shin, Lee, Sheridan, & Lu, 2017; GenaroDaza-Santacoloma, et al., 2009; Subasi & Gursoy, 2010; Sun, Gang, Bo, Zhang, & Zhang, 2017) based type and feature selection (Das, Sengupta, & Bhattacharyya, 2018; Dessì & Pes, 2015; Ferreira & Figueiredo, 2012; Kolhe & Deshkar, 2017) based type. Feature extraction methods are usually based on feature transformation, essentially projecting high-dimensional data into low-dimensional subspace. This kind of dimension reduction methods generally preserve the original relative distance between features and help to cover the potential structure of the original data, so they will not cause a large loss of information. However, when encountering such data sets containing a large number of irrelevant features, the effect of dimensionality reduction is usually poor because almost all features are inevitably taken into account in projection. Principal Component Analysis (PCA) (Hotelling, 1933), Multi-Dimensional Scaling (MDS) (Kruskal & Wish, 1978), Isometric Mapping (ISOMAP) (Tenenbaum, Silva, & Langford, 2000), and Locally Linear Embedding (LLE) (Roweis & Saul, 2000) are typical feature extraction based dimensionality reduction methods. Feature selection methods sort the original features according to specific criteria and select the top-ranked features to form a subset. There are three main models to feature selection: filter, wrapper, and embedded models. The filter models can achieve quick sorting of features to remove a large number of irrelevant or noise features. They usually have good generalization performance and high computational efficiency as they are independent of the classification algorithm. Therefore, these feature select base methods can effectively remove the irrelevant features, but the potential structure of the original sets will be destroyed most likely. Feature selection based on Fisher score (Malina, 1981), information gain (Quinlan,

1986), mutual information (Peng, Long, & Ding, 2005), and Gini index (W. Shang, et al., 2007) are few classical methods.

The coexistence of relevant features usually leads to information redundancy. "The *m* best features are not the best *m* features". Many feature selection methods can remove irrelevant features to obtain "the *m* best features", but there may still be a lot of redundancy among them. To cope with this problem, feature clustering, or feature grouping, has been proposed. Song et al. proposed a fast clustering-based feature selection method, in which features are divided into clusters based on graph-theoretic clustering. For every cluster, the most representative feature strongly related to target classes is selected into the final subset (Song, Ni, & Wang, 2013). Dubey et al. proposed a method based on the k-means clustering and information gain, in which k-means clustering is used for feature clustering and then information gain is employed to select a most relevant feature from each cluster (Dubey, Saxena, & Shrivas, 2017). Dehghan et al. proposed an agglomerative hierarchical clustering based method and mutual information was used to select the representative feature in each cluster (Dehghan & Mansoori, 2018). In these methods, the relationships among features especially the relevant ones are considered to resist the instability caused by the explosion of variable size in high-dimensional set. Highly relevant features will be integrated into groups to reduce redundancy, which helps to improve the generalization performance of the method and reduce the model complexity.

However, some irrelevant features that are correlated to each other will also be integrated into a group, resulting in the existence of irrelevant features in the final subset. To solve this problem, a series of multi-stage or hybrid dimension reduction models were proposed to remove irrelevant, redundant, and noisy features (Bharti & Singh, 2012, 2014, 2015; Zhang, Zou, Zhou, & He, 2018). However, feature selection or extraction operations in all these studies are carried out on the overall feature set or subset to filter out the irrelevant features or information. The mentioned clustering strategy is not combined

further. In fact, feature compression in every single cluster can better help to remove redundant information and cover the latent structure of the set.

Therefore, we design to combine the above clustering strategy and hybrid operation and propose a novel fast hybrid dimension reduction method based on feature selection and grouped feature extraction. Firstly, two different feature selection methods are used to remove irrelevant features. Then, the selected features are grouped into a certain amount of clusters based on the redundancy among them, in which the number of clusters is determined by intrinsic dimensionality estimation method. Finally, PCA based feature extraction is carried out in each cluster to remove redundant information. The proposed method incorporates the advantages of removing both irrelevant and redundant information of the data. Meanwhile, the latent structure of the set is also taken into account. The computational efficiency, classification performance and information loss in supervised learning problems are tested on 20 public datasets considering the diversity of data size and dimensionality. The results show that the proposed method has excellent efficiency and competitive classification performance compared with the contrastive methods.

2. Methods

2.1 Intrinsic dimensionality estimation

The most important factor in dimensionality reduction problems of high-dimensional data is to determine the intrinsic dimensionality, whether to linear or nonlinear approaches. The exact definition of "intrinsic dimensionality" was given by Bennett (Bennett, 1969) in 1969, which is the minimum number of free parameters required to describe all data in a dataset. The accurate estimation of the intrinsic dimensionality provides help to understand the internal structure of the high-dimensional data,

and it will also provide important guides for the subsequent dimensionality reduction and data processing works (Camastra, 2003). The existing intrinsic dimensionality estimation methods can be roughly divided into three categories: projection-based, geometry-based and probability-based methods.

Probability-based methods estimate the intrinsic dimensionality by constructing distribution hypotheses to the data, which make full use of the local information of the datasets and often has high robustness. The most representative one of them is maximum likelihood estimation (MLE) method (Levina & Bickel, 2004). As long as the appropriate maximum likelihood estimation function is used, good results can be obtained. MLE and its improved methods have been applied in most practical applications and show a considerable universality. Therefore we use the MLE method to estimate the intrinsic dimensionality in this paper.

2.2 Multi-strategy combination based feature selection

Feature selection selects a subset of relevant features based on defined criteria and preserves the important information of the original data set. According to the working mechanism of evaluation criteria, feature selection methods can be divided into three main categories: filter method, wrapper method and embedded method (Ladha & Deepa, 2011). Different from wrapper and embedded methods, filter methods depend on only the internal characteristics of the feature subset itself, so they always have high generalization ability and computational efficiency. As feature-ranking methods, filter methods are generally used for preprocessing and independent with specific classifier. Irrelevant and noise features will be filtered out rapidly, decreasing the search scope of the optimal feature subset especially when it comes to high dimensional problems.

A large number of filter methods have been proposed over the years. According to different strategies, different feature subset evaluation functions, such as dependency-based, information-based,

distance-based, consistency-based and so on, are used to select features. Among them, dependency-based functions are used to optimize the feature selection by dependency degree of class information on features, so they are suitable for solving supervised learning problems. Information base ones consider the difference caused by individual feature from the perspective of information theory. Considering to combine these two strategies, we propose multi-strategy feature selection based on Fisher Score (FiS) and Information Gain (IG) to filter irrelevant and noise features. The features having larger between-class distance and lower within-class distance in FiS based method will be selected. IG is based on information entropy to measure the correlation of feature and the class label.

Given a data set $F=\{F_1,\cdots, F_D\}$ containing D features, $FS_1=\{F_{11},\cdots, F_{1D}\}$ is the feature sequence ranked by FiS and $FS_2=\{F_{21},\cdots, F_{2D}\}$ is the one by IG. Thus, to filter low scored features that are present in both feature sequences, we apply the union approach on the lowest $C\%$ of the two sequences and filter out them from the original feature sets. The new feature subset FS after combined feature selection can be defined as follows:

$$FS = F - \{C\%\{FS_1\} \cup C\%\{FS_2\}\} \quad (1)$$

2.3 Hierarchical clustering of the features

According to the aim of dimensionality reduction, the features containing redundant information should be grouped to select/extraction. The maximal information compression index (Mitra, Murthy, & Pal, 2002) can be used as the criterion for measuring the redundancy between two features. Let Σ be the covariance matrix of two variables x and y, the maximal information compression index can be defined as $\delta(x, y) =$ the smallest eigenvalue of Σ, i.e.,

$$2\delta(x,y) = S_x + S_y - \sqrt{\left(S_x + S_y\right)^2 - 4S_xS_y\left(1-\rho^2\right)} \quad (2)$$

where $S_x = \mathrm{var}(x)$, $S_y = \mathrm{var}(y)$, $\rho = \mathrm{cov}(x,y) \big/ \sqrt{S_xS_y}$.

The value of λ is zero when the features are linearly dependent and increases as the dependency decreases. In this paper, features are grouped according to the above criteria. Firstly, the intrinsic dimensionality is calculated and hierarchical clustering (Rokach & Maimon, 2005) is used to integrate the features that have high redundancy into the same group according to the matrix containing λ between all of the feature pairs.

2.4 PCA based dimensionality reduction in feature clusters

PCA is one of the most popular dimensionality reduction method proposed by Hotelling (Hotelling, 1933). With the variance of data as the standard, PCA measures the amount of information contained in the data. Higher variance corresponds to a larger amount of information. The computation of PCA includes singular value decomposition and projection transformation. The original high-dimensional data is mapped to a linear subspace formed by a few number principal components with relatively larger eigenvalues. Thus the correlation among the original dimensions is eliminating and the dimension of the data is reduced. In this paper, PCA is used to reduce the redundancy in every cluster obtained through hierarchical clustering, in which the features share much redundant information and are highly compressible.

2.5 Classifiers

Dimensionality reduction should realize reliable representation of the original data using lower-dimensional extracted data, to explain the data more quickly and more robustly with simpler models. For supervised learning problems, one of the criteria for measuring the effect of dimensionality

reduction is whether the new data can get relatively high classification accuracy and the supervised classifier algorithm will be involved. In this paper, we try different four kinds of algorithms to assess the classification performance of different dimensionality deduction methods, including k-Nearest Neighbors (kNN) (Altman, 1992), Support Vector Machines (SVM) (Chang & Lin, 2011), Bagging (BAG) (Breiman, 1996), and Random Forests (RF) (Breiman, 2001). In this study, the number of nearest neighbors k is set to 1 for kNN and a multi-class SVM model with Radial Basis Function (RBF) kernel function is selected. For the ensemble classifiers BAG and RF, the number of decision trees is set to 50 and 80 respectively.

2.6 Information loss evaluation in dimensional reduction

To evaluate the information loss degree of various dimensionality reduction methods, the concept of representation entropy proposed by Devijve et al. (Devijver & Kittler, 1982) can be introduced.

Given a feature set with dimensionality size d and let Σ be its covariance matrix, thus the eigenvalues of the $d \times d$ matrix can be presented by $\lambda_i \left(i = 1, \cdots, d \right)$. Let $\tilde{\lambda}_i = \lambda_i \Big/ \sum_{i=1}^{d} \lambda_i$, then $\sum_{i=1}^{d} \lambda_i = 1$ and $0 \le \tilde{\lambda}_i \le 1$. The entropy function representation entropy H_R can be defined as:

$$H_R = -\sum_{j=1}^{d} \tilde{\lambda}_j \log \tilde{\lambda}_j \tag{3}$$

Representation entropy can be used to measure the information compression ratio of dimensionality reduction and a higher H_R indicates the less information loss.

2.7 Proposed hybrid dimensionality reduction method

A pictorial summary of the proposed dimensionality reduction method processing is presented in Figure 1. The processing starts with the raw datasets. Next step is intrinsic dimensionality estimation

based on MLE to determine the ultimate dimension. After the above estimation, multi-strategy combination based feature selection is used to filter irrelevant and noise features. We use two kinds of feature ranking methods FiS and IG to assign a score to each feature and a series of low scored features will be filtered out. Next, the remaining features will be clustered into different groups according to the redundant information among them. The number of groups is determined by the results of the above intrinsic dimensionality estimation. Hierarchical clustering based on the maximal information compression index is used to perform this task. Finally, the feature extraction method PCA is applied to further refine the selected feature space. The first principal component of each feature group is used as the corresponding representative feature vector, which not only removes redundant information in the group but also covers the latent structure of the set. To evaluate the proposed method comprehensively, runtime, classification accuracy, and information loss are recorded to be compared with the other conditions or methods.

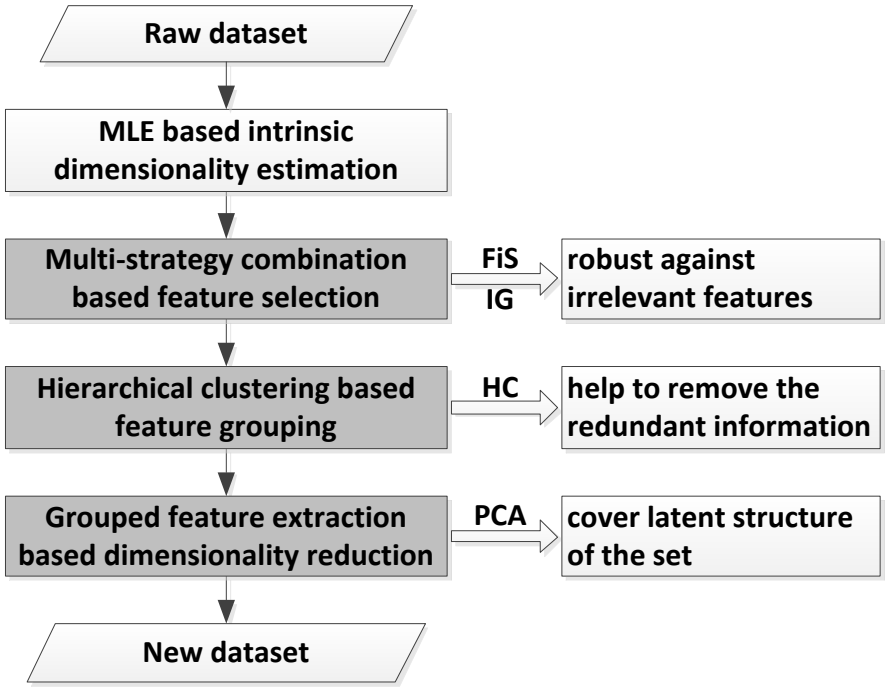

Fig.1 The flowchart of the proposed methodology.

3. Experiments and results

3.1 Datasets and experiment setup

In this paper, 20 real-life publicly available datasets are selected to test the methods taking the diversity of the quantity of class, instance and feature into account [1, 2]. Three categories of datasets are chosen: low-dimensional (dimension less than 100), medium-dimensional (dimension between 100 and 1000), high-dimensional (dimension more than 1000), and there are both small and relatively larger amount of instances respectively. The corresponding statistical information of all datasets is shown in Table 1. In this paper, all related numerical experiments calculations are carried out based on MATLAB software. The experimental condition is: Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz/64 GB/Windows 7/MATLAB R2014a.

3.2 Intrinsic dimensionality estimation and feature selection results

As mentioned above, the MLE method is used to estimate the intrinsic dimensionality in this paper. Multi-strategy based feature selection combing FiS and IG is used to filter irrelevant and noise features. For all 20 datasets, their corresponding intrinsic dimensionality and feature selection results are displayed in Table 1. The last column in the table shows the dimensionality of the feature subset after the multi-strategy based feature selection. In this paper, we set the parameter $C=20$ to filter low scored features.

[1] http://archive.ics.uci.edu/ml/

[2] http://featureselection.asu.edu/datasets.php

Table 1 Summary of the 20 public datasets and intrinsic dimensionality estimation results.

Dataset		# Classes	# Instances	# Features	# Intrinsic Dimensionality	# Dimensionality after Feature Selection
Low-dimensional sets	pima	2	768	8	5	5
	breast	2	648	9	5	6
	wine	3	178	13	6	9
	heart	2	270	13	7	10
	australian	2	690	14	6	10
	waveform3	3	5000	21	14	17
	german	2	1000	24	8	17
	landsat	6	2000	36	10	29
	Satellite	6	6435	36	11	29
	spambase	2	4601	57	7	43
	optdigits	10	5620	64	9	50
Medium-dimensional sets	musk	2	6589	166	4	132
	dna	3	2000	180	43	122
	msplice	3	3175	240	47	167
	MultiFeat	10	2000	649	13	502
	MNIST	10	10000	784	219	591
Large-dimensional sets	COIL20	20	1440	1024	4	767
	lung	5	203	3312	20	2502
	PCMAC	2	1943	3289	90	2167
	BASEHOCK	2	1993	4862	99	3328

3.3 Runtime comparison

For the filtered sets, we use different dimensionality reduction methods including traditional dimension reduction methods PCA, classical feature selection method using minimal-redundancy-maximal-relevance (MRMR) criterion (Peng, et al., 2005), grouped sorting feature selection (GSFS) methods (Z. Shang & Li, 2017) and our proposed fast hybrid method to reduce the dimensionality. We select PCA, GSFS and MRMR because they can meet the needs of comparing the performance of the

algorithms from different aspects we want to carry out. As we know, PCA reduces the dimension of dataset by feature extraction, while MRMR selects features from the original dataset to reduce the dimension and GSFS introduces grouped feature selection strategy to reduce dimension, focusing on selecting features in each feature cluster from the original dataset. Our fast hybrid method combines feature selection and extraction, and introduces grouped feature extraction strategy to reduce dimension. The comparisons between PCA and the proposed method help to explain the advantages of the grouped feature extraction strategy. The introduction of grouped strategy contributes greatly to the fast running time. The comparisons between MRMR feature selection and the proposed method focus on the performance differences between feature selection and feature extraction strategies. The comparisons between GSFS and the proposed method focuses on comparing the performances of grouped feature selection and grouped feature extraction under the common premise of feature clustering. The details of these above comparisons are summarized in the Table 2.

Table 2 Details of the comparisons of the four methods.

Methods	PCA	MRMR	GSFS	Proposed
PCA	——	non-grouped feature **extraction *VS*** non-grouped feature **selection**	**non-grouped** feature **extraction *VS* grouped** feature **selection**	**non-grouped** feature extraction ***VS* grouped** feature extraction
MRMR	non-grouped feature **selection *VS*** non-grouped feature **extraction**	——	**non-grouped** feature selection ***VS* grouped** feature selection	**non-grouped** feature **selection *VS* grouped** feature **extraction**
GSFS	**grouped** feature **selection *VS* non-grouped** feature **extraction**	**grouped** feature selection ***VS* non-grouped** feature selection	——	grouped feature **extraction *VS*** grouped feature **selection**
Proposed	**grouped** feature extraction ***VS* non-grouped** feature extraction	**grouped** feature **extraction *VS* non-grouped** feature **selection**	grouped feature **selection *VS*** grouped feature **extraction**	——

Table 3 records the runtime of the four methods. Best results are marked in boldface.

Table 3 Runtime comparison of different dimensionality reduction methods.

Dataset		Time/ms			
		PCA	MRMR	GSFS	Proposed
Low-dimensional sets	pima	1.03	11.03	6.31	**0.02**
	breast	1.05	3.47	6.89	**0.02**
	wine	**0.90**	2.42	4.17	1.04
	heart	1.00	3.09	6.04	**0.02**
	australian	1.16	3.73	11.76	**0.88**
	waveform3	4.30	52.44	107.18	**0.24**
	german	1.61	10.28	25.72	**0.89**
	landsat	3.76	29.67	88.86	**2.07**
	Satellite	11.65	102.42	254.08	**3.57**
	spambase	12.59	61.35	244.11	**0.05**
	optdigits	20.48	126.96	428.16	**8.34**
Medium-dimensional sets	musk	93.57	147.79	933.97	**34.08**
	dna	18.83	696.45	344.71	**2.26**
	msplice	69.34	1593.88	709.26	**3.87**
	MultiFeat	159.09	852.47	1657.92	**9.41**
	MNIST	774.05	5654.76	8578.16	**3.60**
Large-dimensional sets	COIL20	303.81	331.70	2321.20	**118.59**
	lung	197.04	757.63	1375.45	**8.13**
	PCMAC	2849.45	7679.63	5463.87	**2.23**
	BASEHOCK	5115.46	8726.91	8566.91	**0.05**

From Table 3 we can observe that the proposed hybrid method is consistently much faster than the other three considering feature extraction or feature selection only in almost all of the sets except just one (wine). The runtime of the proposed method is only average 2.70% of that of GSFS, 7.79% of that of MRMR, and 24.39% of that of PCA respectively. Even on big datasets with a large number of samples (such as MNIST) or very high dimensionality (such as PCMAC and BASEHOCK), the computational

efficiency advantage of the proposed method is especially remarkable. In a word, the results show that the proposed hybrid method outperforms the other methods in terms of runtime.

3.4 Classification performance comparison

This section presents the classification results obtained using several methods to compare the performance of the proposed hybrid method with the filtered sets by multi-strategy based feature selection, PCA, MRMR, and GSPS. Classification training are conducted by using repeated hold-out validation, in which the training set (70% of all data) is sampling from the data set randomly, whereas the validation set (the remaining 30%) is used to evaluate the predictive performance for 20 times validation. Accuracy means and standard deviations (std) of kNN, SVM, BAG, and RF are computed and displayed in Figure 2, 3, and 4 for low-, medium-, and large-dimensional sets respectively. And the corresponding average accuracy of four different classifiers by different dimensionality reduction methods is reported in Table 4.

Figure 2 shows the classification performance on 12 low-dimensional datasets. From it, we observe that the accuracy by different dimensionality reduction methods is basically at the same level on almost all of the sets. What's more, on sets heart, landsat, and Satellite, the accuracy by the proposed method is quite close to that of the filtered sets and a little higher than those of PCA, MRMR and GSFS.

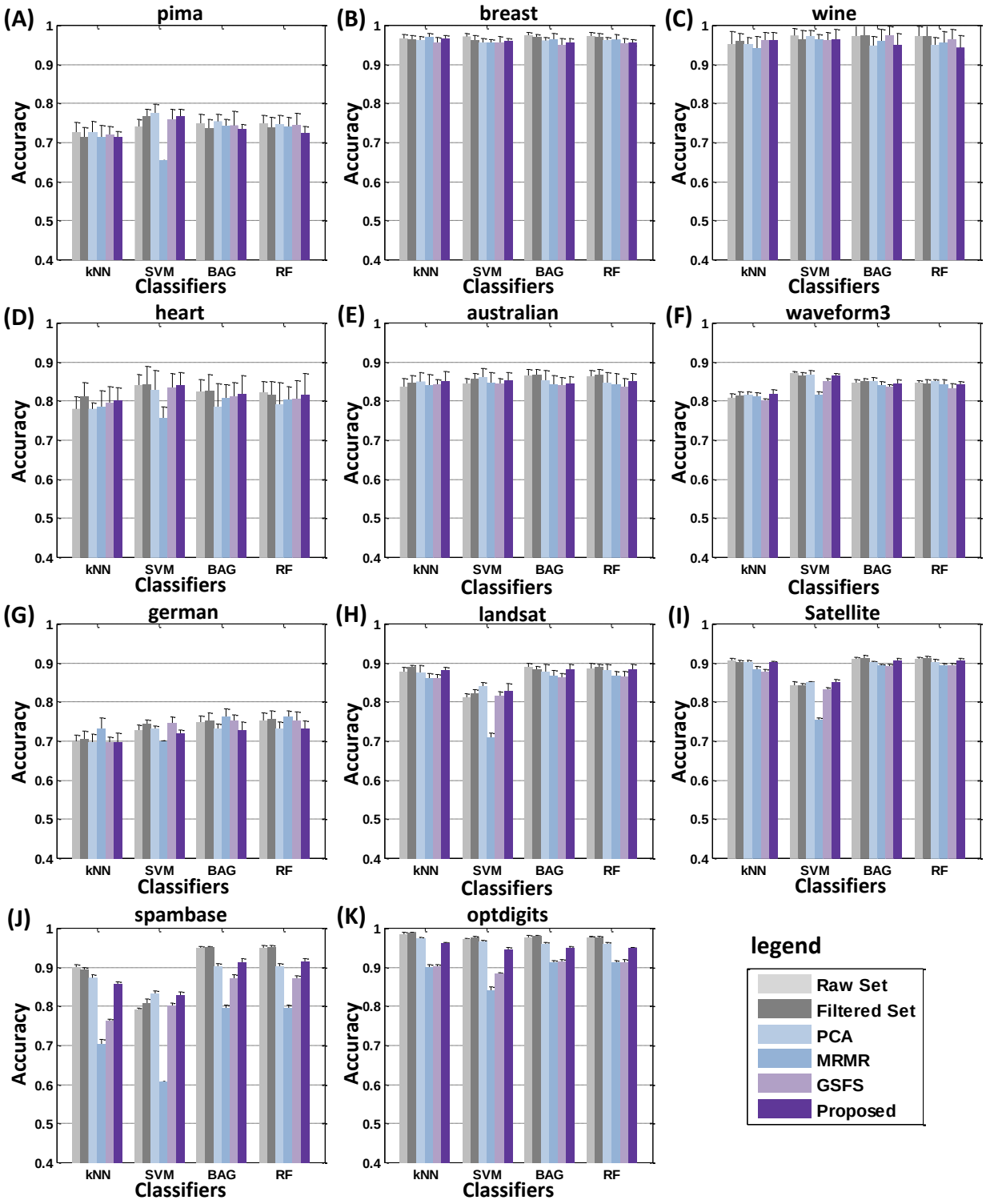

Fig.2 Classification performance comparison of different dimensionality reduction methods using multiple classifiers on 11 low-dimensional datasets.

The classification performance on 5 medium-dimensional datasets is shown in Figure 3. We observe that in most cases the accuracy by the proposed method is competitive compared with those of PCA, MRMR, and GSFS, although sometimes a little lower than that of the filtered sets. However, the overall comparison still shows no significant difference.

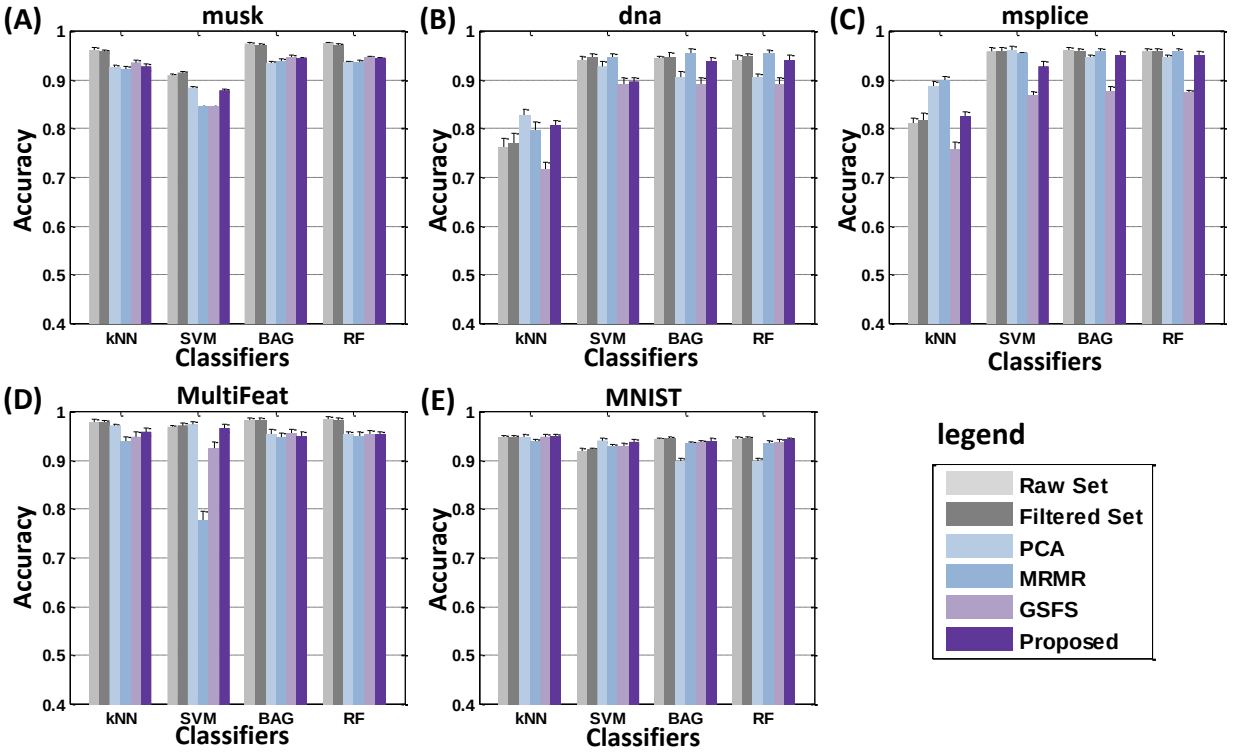

Fig.3 Classification performance comparison of different dimensionality reduction methods using multiple classifiers on 5 medium-dimensional datasets.

Figure 4 reports the classification performance on 4 large-dimensional datasets. It shows that the accuracy by different dimensionality reduction methods is roughly the same. Compared with the filtered set, although the accuracy of the proposed method is obviously lower on COIL20 sets, their performance on the other sets is quite close.

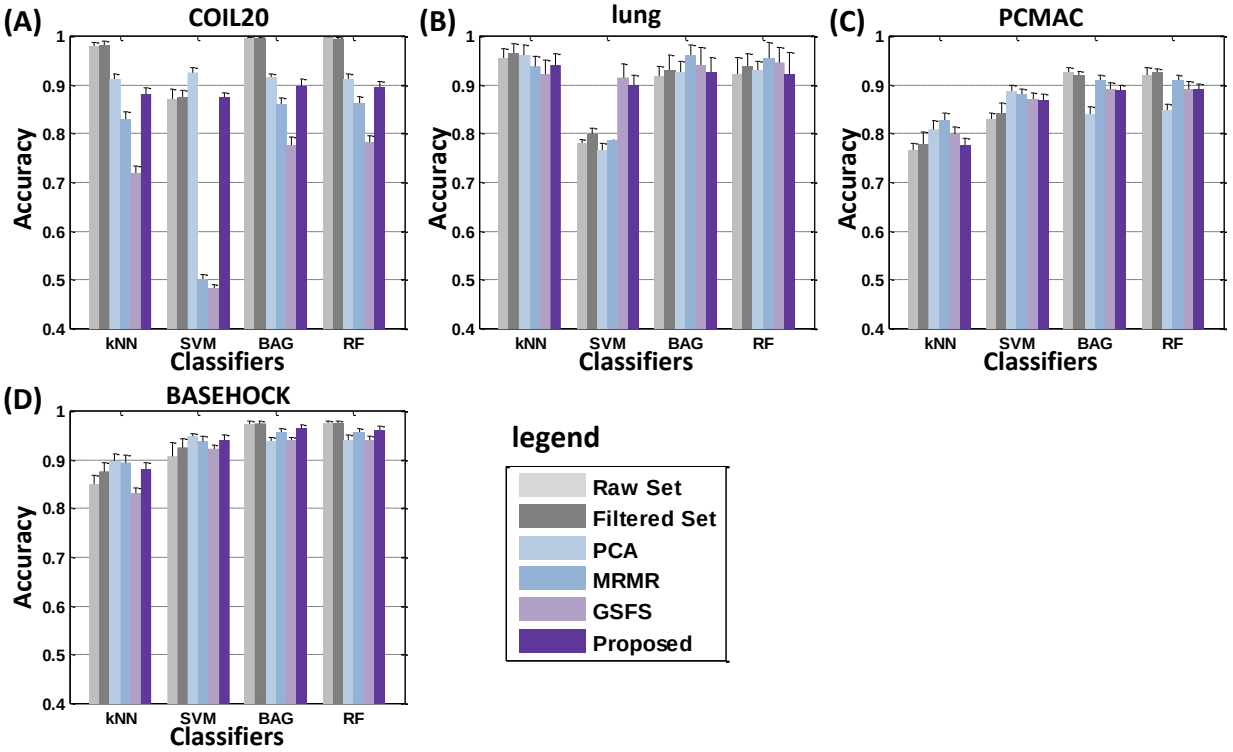

Fig.4 Classification performance comparison of different dimensionality reduction methods using multiple classifiers on 4 large-dimensional datasets.

Generally speaking, the proposed method achieves acceptable classification accuracy on all datasets, which indicates that this kind of hybrid dimensionality reduction method can not only fast reduce the dimension of the feature set, but also ensure the considerable accuracy of classification.

According to the average accuracy reported in Table 4, we observe that on almost all of the sets, the classification performance by different dimensionality reduction methods basically has no significant difference. On a series of sets such as heart, landsat, Satellite, spambase, musk, MNIST, and BASEHOCK, the proposed hybrid method outperforms PCA, MRMR and GSFS, and obtains competitive performance compared with the corresponding filtered set. Furthermore, the proposed method obtained the highest accuracy on MNIST sets. Although the proposed method does not always show the best performance on the remaining other sets, the accuracy results are still acceptable.

Table 4 Average accuracy of four different classifiers by different dimensionality reduction methods.

Dataset		Average accuracy of four different classifiers					
		Raw set	Filtered set	PCA	MRMR	GSFS	Proposed
Low-dimensional sets	pima	0.7383	0.7383	0.7510	0.7121	0.7412	0.7341
	breast	0.9661	0.9661	0.9616	0.9638	0.9545	0.9605
	wine	0.9684	0.9684	0.9557	0.9557	0.9665	0.9552
	heart	0.8247	0.8247	0.7963	0.7889	0.8120	0.8194
	australian	0.8606	0.8606	0.8535	0.8431	0.8409	0.8499
	waveform3	0.8446	0.8446	0.8469	0.8287	0.8303	0.8442
	german	0.7398	0.7398	0.7222	0.7387	0.7370	0.7192
	landsat	0.8707	0.8707	0.8699	0.8276	0.8518	0.8702
	Satellite	0.8929	0.8929	0.8896	0.8573	0.8742	0.8922
	spambase	0.9015	0.9015	0.8782	0.7251	0.8272	0.8798
	optdigits	0.9808	0.9808	0.9661	0.8922	0.9041	0.9522
Medium-dimensional sets	musk	0.9548	0.9548	0.9203	0.9113	0.9188	0.9237
	dna	0.9034	0.9034	0.8914	0.9134	0.8477	0.8960
	msplice	0.9248	0.9248	0.9355	0.9436	0.8444	0.9147
	MultiFeat	0.9799	0.9799	0.9648	0.9056	0.9472	0.9581
	MNIST	0.9410	0.9410	0.9229	0.9365	0.9398	0.9439
Large-dimensional sets	COIL20	0.9619	0.9619	0.9161	0.7640	0.6894	0.8869
	lung	0.9082	0.9082	0.8959	0.9107	0.9320	0.9234
	PCMAC	0.8666	0.8666	0.8453	0.8829	0.8639	0.8566
	BASEHOCK	0.9377	0.9377	0.9322	0.9364	0.9100	0.9372

3.5 Information loss comparison

For the dimensionality reduction method, it is expected that the reduced new set can share as more information as the raw set. Thus we use representative entropy H_R to evaluate the information loss of different dimensionality reduction methods. Their information loss is defined as:

$$Information\_loss = \left(1 - H_R / H_R^{filter}\right) * 100 \quad (4)$$

where H_R^{filter} is the representation entropy of the filtered set and H_R represents the corresponding representation entropy of PCA, MRMR, GSFS, and the proposed method. Table 5 shows the comparative information loss of them. Best results are shown in boldface.

Table 5 Information loss comparison of different dimensionality reduction methods.

Dataset		Information loss /%			
		PCA	MRMR	GSFS	Proposed
Low-dimensional sets	pima	0	0	0	**0**
	breast	10.25	5.47	**4.37**	10.44
	wine	10.65	16.71	**1.36**	11.66
	heart	**8.35**	14.30	11.56	8.45
	australian	**11.63**	15.94	15.94	11.68
	waveform3	9.27	9.82	**5.05**	9.57
	german	**20.45**	26.03	22.54	20.64
	landsat	**9.35**	13.97	22.70	16.82
	Satellite	**8.49**	16.56	20.17	15.72
	spambase	**43.58**	43.65	44.14	43.85
	optdigits	**32.09**	35.92	35.07	34.96
Medium-dimensional sets	musk	55.82	**48.68**	49.97	60.57
	dna	19.41	22.46	**19.33**	19.64
	msplice	21.01	25.33	**20.67**	20.97
	MultiFeat	**35.75**	54.32	41.12	42.73
	MNIST	**3.87**	9.83	7.26	6.89
Large-dimensional sets	COIL20	61.17	**59.82**	69.68	64.83
	lung	**38.12**	55.71	41.59	56.18
	PCMAC	**25.81**	30.93	27.71	29.32
	BASEHOCK	**26.39**	31.99	33.63	31.24

As Table 5 shows, although the four methods show very close results on most datasets, the information loss of the proposed method on some datasets is still higher than that of PCA (landsat, Satellite, MultiFeat, MNIST, lung, PCMAC, and BASEHOCK), MRMR (breast, musk, and COIL20), or GSFS (breast, wine, waveform3, musk, and lung). It is difficult to avoid greater information loss on all data sets for the proposed method, and these results show that there is still promotion space for the proposed method to retain more information in dimensionality reduction.

3.6 Summary of the performance comparisons

Table 6 illustrates a summary of the above performance comparisons, i.e., how many datasets the proposed model obtains better (or worse, or no significant difference) results in the comparison to the competing method. Here, '+' indicates that the proposed fast hybrid method is superior to the competitive method for the considered measure. Similarly, '–' indicates that the proposed method is inferior to the competing method. '≈' indicates that there is no significant difference between the result of the proposed method and that of the competing method. The statistical analysis method used in this paper is the rank-sum test, and statistically significant differences are indicated as $p < 0.05$.

As far as runtime is concerned, Table 6 clearly illustrates that the proposed fast hybrid method performs better than the other competing methods on all almost all of the datasets except only one. These results highlight the enormous efficiency advantage of the proposed method. When considering the classification performance, it illustrates that the proposed method performs better than or show no significant difference with MRMR and GSFS in most of the cases. The results of the comparison with PCA are a little unsatisfactory when using kNN and SVM as classifiers, while the results using BAG and RF still show superiority. When it comes to the comparisons of information loss, we find that the performance

of the proposed method is inferior to PCA. However, it is still obvious that the proposed method performs better than MRMR and GSFS on more than half of the datasets.

Table 6 Summary of the performance comparisons on time, accuracy, and information loss.

Methods	Time		Accuracy (kNN / SVM / BAG / RF)			Information loss	
	+	–	+	≈	–	+	–
Proposed VS PCA	19	1	0 / 0 / 7 / 7	12 / 11 / 10 / 10	8 / 9 / 3 / 3	2	18
Proposed VS MRMR	20	0	7 / 13 / 6 / 9	10 / 4 / 9 / 11	3 / 3 / 5 / 0	14	6
Proposed VS GSFS	20	0	10 / 11 / 9 / 7	8 / 8 / 9 / 9	2 / 1 / 2 / 4	11	9

4. Conclusion and discussion

A fast hybrid dimensionality reduction method based on feature selection and grouped feature extraction for classification has been presented in this paper. This method is designed to achieve multi-purpose in dimensionality reduction: to remove irrelevant features, to remove redundancy information among features while to cover the latent structure of the datasets. Firstly, multi-strategy fusion based feature selection is used to filter out irrelevant features, in which two classical feature selection methods (FiS and IG) based on different evaluation criteria are combined. To further remove redundant information, we choose hierarchical clustering based on the maximum information compression index for feature grouping, ensuring that compressible features containing redundant information are grouped into the same group. Then, PCA is carried out independently in each group to reduce the dimensionality. At the same time, this can also guarantee to cover the latent data structure of each feature group as far as possible. The first principal component is extracted as the representative feature of each group and finally, these principal component features are combined to form new feature sets. We evaluate and compare the methods based on runtime, classification accuracy, and information loss.

The results show that the proposed method is the fastest and has competitive classification performance. In terms of the information loss of the methods, the results show that the proposed method needs to be further improved. However, what is interesting is that the information loss of the proposed method does not result in significant classification performance decrease, which may indicate, on the other hand, that the lost information is more likely to be redundant information irrelevant or weakly-dependent for classification, or even noise information.

Benefit from the PCA based feature extraction operation we design, the computational complexity of the proposed method is relatively lower than PCA, and so it can save a lot of runtime. We use Figure 5 to provide descriptive overviews of PCA and the proposed method for conceptual understanding. Taking a data set with all features shown in the figure as an example, the PCA operation is performed directly on the original dimensional feature set. Assuming that our method clusters these features into groups, with a certain number of features in each group. PCA operations will be performed on these feature groups respectively in the next step, and these corresponding first principal components are used as the representative feature of each group.

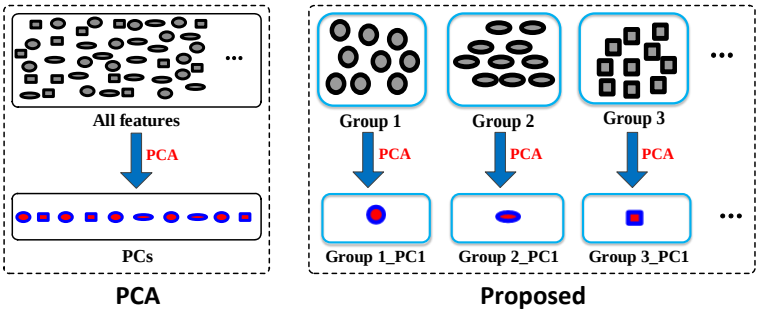

Fig.5 Graphical overview of PCA and the proposed method.

Nevertheless, such information loss of the proposed method mentioned ablove may also be closely related to the feature extraction operation we design. For every group obtained through feature clustering, we just only retain the first principal component and abandon all other components in the

process of grouped PCA. Thus it may inevitably cause more information loss. So in the future works, we will try to design other processing strategies to remove redundant information, such as assigning priority weights to different feature groups so as to determine how many principal components of the corresponding feature groups should be retained, which may help us to further retain more effective information.

Another problem to be further studied and solved is that the performance of the proposed method depends on the multi-strategy feature selection parameters C to some extent. For different data sets, the number of irrelevant features and noise features is different, so the parameter should also be adapted to different data sets. For a definite data set, different parameter setting inevitably results in various performances. Too small parameters may cause incomplete removing of redundancy, whereas too large parameters will lead to the filtering of the effective features. Therefore, we hope to solve this parameter dependency problem and determine the optimal parameter by designing more effective approachs next.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician, 46*, 175-185.

Bennett, R. (1969). The intrinsic dimensionality of signal collections. *IEEE Transactions on Information Theory, 15*, 517-525.

Bharti, K. K., & Singh, P. K. (2012). A two-stage unsupervised dimension reduction method for text clustering. In J. C. Bansal, P. Singh, K. Deep, M. Pant & A. Nagar (Eds.), *2012 7th International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)* (pp. 529-542). India: Springer, India.

Bharti, K. K., & Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science, 5*, 156-169.

Bharti, K. K., & Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. *Expert Systems with Applications, 42*, 3105-3114.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-12.

Camastra, F. (2003). Data dimensionality estimation methods: A survey. *Pattern Recognition, 36*, 2945-2954.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*, 1-27.

Choi, S., Shin, J. H., Lee, J., Sheridan, P., & Lu, W. D. (2017). Experimental demonstration of feature extraction and dimensionality reduction using memristor networks. *Nano Letters, 17*, 3113-3118.

Das, A. K., Sengupta, S., & Bhattacharyya, S. (2018). A group incremental feature selection for classification using rough set theory based genetic algorithm. *Applied Soft Computing, 65*, 400-411.

Dehghan, Z., & Mansoori, E. G. (2018). A new feature subset selection using bottom-up clustering. *Pattern Analysis and Applications, 21*, 57-66.

Dessì, N., & Pes, B. (2015). Similarity of feature selection methods: An empirical study across data intensive classification tasks. *Expert Systems with Applications, 42*, 4632-4642.

Devijver, P. A., & Kittler, J. V. (1982). *Pattern recognition: A statistical approach*. London: Prentice/hall International.

Dubey, V. K., Saxena, A. K., & Shrivas, M. M. (2017). A cluster-filter feature selection approach. In *2016 International Conference on ICT in Business Industry & Government (ICTBIG)* (pp. 1-5). Indore, India: IEEE.

Ferreira, A. J., & Figueiredo, M. A. T. (2012). Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters, 33*, 1794-1804.

GenaroDaza-Santacoloma, JuliánD.Arias-Londoño, Godino-Llorente, J., NicolásSáenz-Lechón, VíctorOsma-Ruiz, & GermánCastellanos-Domínguez. (2009). Dynamic feature extraction: An application to voice pathology detection. *Intelligent Automation and Soft Computing, 15*, 667-682.

Golay, J., & Kanevski, M. (2017). Unsupervised feature selection based on the Morisita estimator of intrinsic dimension. *Knowledge-Based Systems, 135*, 125-134.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology, 24*, 417-441.

Kolhe, S., & Deshkar, P. (2017). Dimension reduction methodology using group feature selection. In *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 789-791). Bangalore, India: IEEE.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills and London: Sage Publications.

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering, 3*, 1787-1797.

Levina, E., & Bickel, P. J. (2004). Maximum likelihood estimation of intrinsic dimension. In *2004 17th International Conference on Neural Information Processing Systems* (pp. 777-784). Vancouver, British Columbia, Canada: MIT Press Cambridge.

Maeda, E. (2014). Dimensionality reduction. In K. Ikeuchi (Ed.), *Computer Vision: A Reference Guide*. Boston, MA: Springer US.

Malina, W., . (1981). On an extended fisher criterion for feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 3*, 611-614.

Mitra, P., Murthy, C. A., & Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 301-312.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*, 1226-1238.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81-106.

Rokach, L., & Maimon, O. (2005). Clustering methods. *Data Mining and Knowledge Discovery Handbook, 3*, 321-352.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*, 2323-2326.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications, 33*, 1-5.

Shang, Z., & Li, M. (2017). Feature selection based on grouped sorting. In T. Y. Lawry J, Chai C. (Ed.), *2017 9th International Symposium on Computational Intelligence and Design* (pp. 451-454). Hangzhou, China: IEEE.

Song, Q., Ni, J., & Wang, G. (2013). A fast clustering-based feature subset selection algorithm for high dimensional data. *IEEE Transactions on Knowledge and Data Engineering, 25*, 1-14.

Subasi, A., & Gursoy, M. I. (2010). EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications, 37*, 8659-8666.

Sun, W., Gang, Y., Bo, D., Zhang, L., & Zhang, L. (2017). A sparse and low-rank near-isometric linear embedding method for feature extraction in hyperspectral imagery classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*, 4032-4046.

Tenenbaum, J. B., Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*, 2319-2323.

Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. *Methods, 111*, 21-31.

Zhang, D., Zou, L., Zhou, X., & He, F. (2018). Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *IEEE Access, 6*, 28936-28944.