Neurocomputing 000 (2018) 1–11

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Review of classical dimensionality reduction and sample selection methods for large-scale data processing

Xinzheng Xu[a,b,*], Tianming Liang[a], Jiong Zhu[a], Dong Zheng[a], Tongfeng Sun[a]

[a] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China
[b] Guangxi High School Key Laboratory of Complex System and Computational Intelligence, Guangxi Nanning, 530006, China

ARTICLE INFO

Article history:
Received 24 October 2017
Revised 19 February 2018
Accepted 26 February 2018
Available online xxx

Communicated by Chennai Guest Editor

Keywords:
Large-scale data processing
Sample selection
Dimensional reduction
Machine learning methods

ABSTRACT

In the era of big data, all types of data with increasing samples and high-dimensional attributes are demonstrating their important roles in various fields, such as data mining, pattern recognition and machine learning, etc. Meanwhile, machine learning algorithms are being effectively applied in large-scale data processing. This paper mainly reviews the classical dimensionality reduction and sample selection methods based on machine learning algorithms for large-scale data processing. Firstly, the paper provides a brief overview to the classical sample selection and dimensionality reduction methods. Then, it pays attention to the applications of those methods and their combinations with the classical machine learning methods, such as clustering, random forest, fuzzy set, and heuristic algorithms, particularly deep leaning methods. Furthermore, the paper primarily introduces the application frameworks that combine sample selection and dimensionality reduction in the context of two aspects: sequential and simultaneous, which almost all get the ideal results in the processing of the large-scale training data contrasting to the original models. Lastly, we further conclude that sample selection and dimensionality reduction methods are essential and effective for the modern large-scale data processing. In the future work, the machine learning algorithms, especially the deep learning methods, will play a more important role in the processing of large-scale data.

1. Introduction

Data are increasing in scale and becoming more important. Big data now appears in every industry in modern society. With the continuous development of Artificial Intelligence (AI), big data has become particularly important. From IBM's Deep Blue to Google's Alpha Go—a time span of barely 20 years—AI has developed from neural networks to deep learning algorithms. Big data and improvements in computing power have promoted the development of AI. The computing ability of computers is increasing exponentially, and the accumulation of data is exploding simultaneously. Machine learning and unsupervised learning algorithms must rely on a large number of training samples to ensure powerful performance [1]. However, there are lots of irrelevant, redundant, incomplete and noisy data in training sets as the amount of data becomes increasingly greater [2]. On the other hand, large-scale training data bring problems such as requiring more storage and greater computational complexity, thus influencing the generalization ability and reducing the prediction accuracy. That is quantity and quality of samples influence the performance of the computers and the robustness of the models. In addition to problems associated with samples, another active topic is dimensionality [3]. With recent improvements in storage capacity and analysis technology, researchers have started to focus on the intrinsic properties of samples, namely, the features that are the valuable attributes from all dimensions of the samples. The increasing amount of image data, such as facial images, environmental images and remote sensing images, has been accompanied by an increased focus on computer vision. Moreover, high-dimensional data, especially image data, play an increasingly important role in addressing real-life issues [4]. How to extract and select the most informative or discriminative information is always a very important step in all types of computational fields. Especially in machine learning, even with a model that is very robust and effective, if the data are poor, the model would show lower prediction possibility [5].

For too many problems mentioned above, two frequently used methods, sample selection and dimensionality reduction, are proposed and continuously improved in recent years. Sample selection methods could reduce computational cost and even improve the learning accuracy by discarding the redundant, incomplete, noisy data and other negative samples [6,7]. Traditional sample selec-

* Corresponding author.
E-mail address: xxzheng@cumt.edu.cn (X. Xu).

https://doi.org/10.1016/j.neucom.2018.02.100
0925-2312/

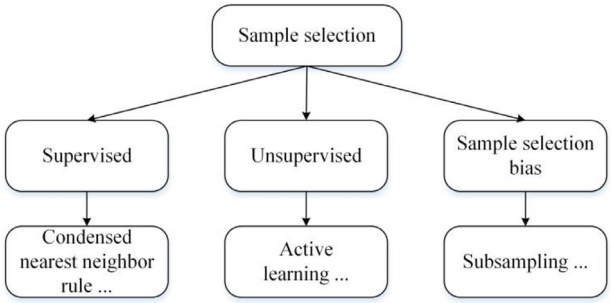

Fig. 1. category of sample selection.

tion methods could be divided into two categories [6]. One kind is data condensing like Condensed Nearest Neighbor rule [8] and Instance-Based Learning (IBL) [9], these methods could remove a part of irrelevant or redundant samples but need a great deal of calculations. The other is active learning which selects a prat of representative unlabeled samples to learn [10,11]. Active learning is an unsupervised method, so the effect of the methods cannot be guaranteed. Features could be seen as a deep partition of the samples, which show the importance in images classification and prediction. Analysis and calculation both are difficult problems for high dimensional data, so the dimensionality reduction technologies arisen at the historic moment. And it is essential in large-scale data models like machine learning nowadays. Different from sample selection methods, dimensionality reduction methods focus on the intrinsic properties of the samples. The effect and calculation speed could be greatly enhanced by selecting the discriminative attributes or discarding a large number of redundant dimensions. However, the features space got by dimensionality reduction means the information is incomplete or even loosed. In addition, this is a challenging problem especially when the data is high in dimensionality. Inadequate features would increase the risk of overfitting and get lower model interpretability [12]. Therefore, the importance of sample selection and dimensionality reduction is self-evident. And almost all of the large-scale data analysis, especially deep learning, use dimensionality reduction and sample selection methods. Machine learning methods have been applied in all kinds of real-world problems like pattern recognition, data mining, predictive analytics, etc. Classical machine learning methods, such as clustering, random forest, fuzzy set, heuristic algorithm, deep leaning and so on, show the perfect effect all along. The applications of efficient machine learning methods for sample selection or dimensionality reduction have been shown to play an increasingly important role in large-scale data processing, and more and more researchers are doing researches on related algorithms and models.

There are already some similar works published before and recently [13–16], but the works either focus on the main theories and the models of samples selection or dimensionality reduction, or focus on reviewing a specific problem like random projection based on dimensionality reduction in [15]. The contributions of this paper are to take a brief overview of the development of sample selection and dimensionality reduction techniques, and focus on the applications of these very useful techniques with machine learning methods. The various machine learning methods used for sample selection and dimensionality reduction are summarized in the paper. And we especially introduce the combination use of sample selection and dimensionality reduction. To the best of our knowledge, this type work is different from the published works, and may be the first time to be done.

This paper will provide researchers with a clear view of two important theories—sample selection and dimensionality reduction—in data processing. And the emphasis of the paper is to review all types of applications with sample selection and dimensionality reduction methods in machine learning tasks as more as possible. This paper is organized as follows. Section 2 introduces sample selection and its applications in all types of machine learning problems. In Section 3, dimensionality reduction techniques are divided into feature extraction and feature selection, and the applications in machine learning methods are further summarized. In particular, the Convolutional Neural Network (CNN) model is emphasized as a combination of feature extraction and feature selection. Comprehensive applications of sample selection and dimensionality reduction are discussed in Section 4. Finally, Section 5 presents the conclusions of the paper and offers discussion.

2. Sample selection in machine learning

Sample selection models in traditional statistical analysis, like random sampling, are easy to understand and most commonly used, and they are far different from the field of machine learning in purpose. In statistical analysis, researchers want to use portions of samples to represent the whole probability distribution, and the aim is to obtain an approximate evaluation scope. However, the purpose of sample selection in machine learning is mainly to remove the redundant and noisy samples. In machine learning, especially deep learning, it seems that more samples are better for the model. Additional high-quality samples can improve the generalization ability and yield better accuracy. Some studies are faced with few samples [17] because the amount of data accumulated is still small or the number of instances is fewer than the number of dimensions [18,19]. However, most researches in machine learning use big data to learn and analyze information such as features as much as possible. Experts want a model to provide only relative answers such as ‘yes’ or ‘no’ . However, analyzing a large-scale input of samples is time-consuming and may cause overfitting of the training model [20]. Sample selection methods or informative subset selection [21] are important for this reason. Scalability is an issue that must be considered [7], especially when very large number is considered. Random selection is used fewer now in machine learning, it usually acts as a slight step before the following process of data processing, or as a random direction selection for some optimization algorithms in machine learning [22,23]. In some large-scale input processing for single data like video sequence and hyperspectral images, random selection could be used in some pixel level selection like [24,25]. In this section, the applications of sample selection are introduced in three aspects: supervised, unsupervised and sample selection bias. An approximate classification of these sample selection methods in machine learning is shown in Fig. 1.

2.1. Supervised sample selection

Sample selection has been successfully applied in many fields, such as visual domain adaptation [26], text classification [27], satellite imagery classification [28], classification of RNA [29], etc. Li [30] et al. proposed a novel sample selection method in visual domain adaptation using sparse coding (SSSC). The algorithm first maps the source and target domains into a common subspace to avoid problems associated with crossing domains. Then, the source domain is treated as the dictionary, with which the target domain can be represented via linear combinations using sparse coding. The most irrelevant samples are discarded via $L_{2,1}$ norm regularization. Additionally, the algorithm also trains the classifier using the specific part from the target domain such that the model can simultaneously select samples from multiple domains. The experimental results show the effectiveness on popular data sets such as MNIST, Caltech256 and others. In text classification, Liao [27] et

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

al. proposed a new sample selection method to remove noisy samples using a representativeness score that indicates the importance of samples. The experimental results demonstrated the effectiveness and efficiency of the method applied to a support vector machine (SVM) classifier. To select the boundary samples and improve the classification accuracy for SVM, Xia [31] et al. presented a novel boundary sample selection mechanism named BSS-SVM. Niu [32] applied the same sample selection flow to predict network traffic, referring to their algorithm as FCM-LSSVM. The FCM part refers to the fuzzy-means clustering algorithm used to remove outliers from the original samples. An improved SVM is used to classify and predict, and the Artificial Bee Colony algorithm is also utilized for the optimization of the model. As intended, the speed and accuracy are both improved. Zhai [33] et al. applied probabilistic neural networks (PNNs) and K-L divergence to select support vectors for SVM to reduce both the time and space complexity. Hao [34] et al. proposed an effective framework for handwritten character recognition. The first step is to condense and select boundary samples using an improved weighted condensed nearest neighbor algorithm. Then, the Backpropagation (BP) Neural network is adopted to predict effectively. The model improves the generalization ability and reduces the training time. Chellasamy [28] et al. presented an automatic training sample selection approach for the crop classification of satellite imagery named Ensemble-based Cluster Refinement Approach (ECRA). The ensemble framework clusters the satellite imageries of crops on the basis of texture, spectral and vegetation indices, respectively. Then, the border samples of each cluster are determined. In this case, three separate Multi-Layer Perceptron (MLP) neural networks are used to evaluate the informative samples and update the sample subsets. Finally, Endorsement Theory (ET) is used to finish the classification. The classification accuracy is improved by nearly 10%, and the most important advantage is the automatic classification. Zhang [20] et al. used the maximum entropy and the contribution of samples to classes based on the nearest neighbor rule (KNN). This new algorithm reduces the storage requirements and speeds up the classification process. Chen [35] et al. proposed a sample selection method based on rough sets. Santiago-Ramirez [36] et al. proposed an optimal mechanism that selects the best subset from the training samples for face recognition. Garc [7] et al. presented a new method, which could be applied to any instance selection method without any modification. The method first divides the original dataset into several disjoint subsets. Next, some weak classifiers are selected to make decisions for each sample. The mean of the voting results is used to determine the final subsets. In various experiments, the classification results of the combination of the weak classifiers are better than those of each single classifier alone, and the voting method could achieve a significant reduction in storage while keeping the testing error tolerable.

2.2. Unsupervised sample selection

In many cases, an excessive number of instances can result in a high computational cost of assigning labels; conversely, there can sometimes be too few known labels [37]. Thus, algorithms for unlabeled training samples are of equal importance. Based on the margin sampling (MS) strategy, Guo [38] et al. introduced an active learning approach to select a small number of the most effective training samples in large-scale remote sensing image classification problems. Compared to stratification systems and random sampling, the classification effect is better, and the space and time complexity are reduced. Wang [6] et al. provided a mechanism that selects representative samples based on the maximum ambiguity in a pre-built fuzzy decision tree. The algorithm selects some samples randomly from the original sample set as the initial training set and labels them by experts. Then, the remaining samples are evaluated using the principle maximum classification ambiguity in the newly built fuzzy decision tree. The selected samples are finally labeled via estimation. The model can reduce the storage space effectively and obtain the desired number of samples. Yuan [39] et al. utilized fuzzy clustering to select the initial training samples for active learning. A hybrid selection (HS) model that contains border-based selection (BS) and center-based selection (CS) to select the border and center samples separately is proposed. The performance of the method for active learning is good. In [40], a Gibbs sampler that converges more easily for the structural learning of directed acyclic graphs is constructed. Xu [11] et al. first interpreted active learning from a purely algebraic point of view and combined it with semi-supervised manifold learning, in which a heuristic sample selection method was employed for labeling. According to the Gershgorin circle theorem, an upper band is calculated to label the samples, and the framework performs well at both regression and classification tasks. Yang [37] et al. also focused on sample selection in active learning for multi-class problem.

2.3. Sample selection bias

In economics, sample selection bias [41] is present when samples are not randomly selected. Zadrozny [42] introduced this problem into machine learning and proved that it was true by analyzing the theoretical formulas of some machine learning methods, such as Bayesian classifiers, SVMs, etc. Sample selection bias in the classification of machine learning refers to whether the sample is selected according to the sample or the corresponding label or both. Moreover, Zadrozny presented that global learners, such as soft margin SVM and naive Bayes (NBC), are affected by sample selection bias, whereas local learners, such as hard margin SVM and logistic regression, are not. Later, Wu [43] et al. further studied the correcting sample selection bias problem in image classification. Kernel density estimation (KDE) is used to predict the distribution of the test set according to the training set. The classification results demonstrate the effectiveness of the framework with classifiers like NBC and SVM.

Sample selection bias is also called covariate shift [44], and it is usually used in the data that the distribution is different between training set and test set [45]. Especially in semi-supervised construction of training data, how to select the unlabeled samples according to the few labeled samples is the key to make a better classifier. For example, in medical publications, the data imbalance makes the biomedical texts classification a great challenge to machine learning. When training the SVM classifier, the imbalance of number on every class would decrease the test accuracy. Romero [46] et al. tried to construct a balanced biomedical texts training set through three sample selection bias techniques: undersampling, resampling and subsampling strategies. Oversampling and subsampling could reconstruct the number and distribution of minority class and majority class, respectively. And the author finds that the subsampling with polynomial SVM could get better classification performance on imbalanced biomedical texts. Krautenbacher [47] et al. also used sample selection bias to correct classifiers in stratified data of epidemiological studies.

3. Dimensionality reduction

With datasets becoming increasingly larger, the dimensionality of single data is also increasing; this issue is sometimes referred to as ultrahigh dimensionality [48]. Dimensionality reduction techniques map high-dimensional data to a lower-dimensional space [49]; such techniques are used widely in machine learning, especially in deep learning, as a necessary data pre-processing method. The main purpose of dimensionality reduction is to find the most

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

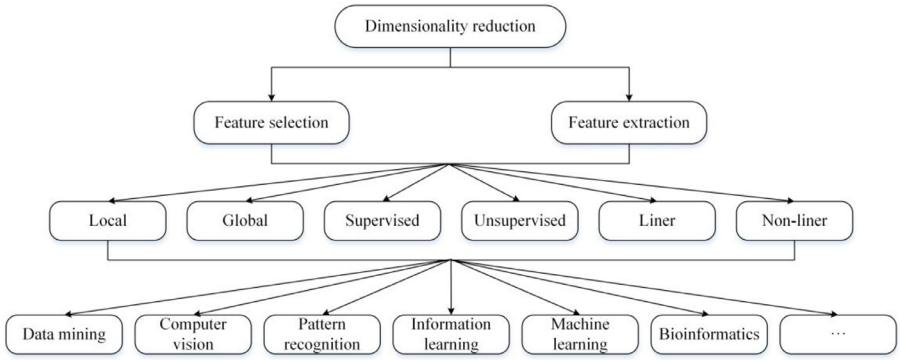

Fig. 2. types and application fields of dimensionality reduction.

useful and informative subspace that not only reduces the computational complexity but also, most importantly, adapts the model to the greatest extent [50]. In some circumstances, the number of sample dimensions is far more than the number of samples [51]. For example, many researchers have found that the number of features is often greater than the number of samples in bioinformatics applications [52]. Thus, to avoid dimension disaster and overfitting, dimensionality reduction is necessary. There are two main aspects in dimensionality reduction: feature extraction and feature selection [53]. Feature selection refers to selecting a portion of the original dimensions that are most important to the task, whereas feature extraction refers to extracting a new and smaller representation set from the original dimension space [54]. Sometimes, feature extraction is difficult to explain and is difficult to adopt in some critical applications [55]. Usually, feature selection methods are easier and used more widely than feature extraction [55]. In feature selection, some reduced dimensions that are most relevant for the target model are selected from the original input space. Feature extraction transforms the original space to a low-dimensional subspace. The structure may be changed relative to the original structure by (non-)linearly combining existing features [56]. Thus, feature selection is more useful in text classification and gene selection, and feature extraction is commonly applied in image classification and pattern recognition. Dimensionality reduction techniques can be further divided according to the following aspects: linear or nonlinear, supervised or unsupervised, local or global [57]. Such methods are extensively applied in classification, clustering, regression, prediction, and so on. The types and application fields of dimensionality reduction are illustrated in Fig. 2.

In this section, we first briefly introduce the basic theory of dimensionality reduction. Subsequently, a more detailed summary of feature extraction and feature selection, two main aspects of dimensionality reduction, and their broad applications are given.

3.1. Feature extraction

As mentioned above, feature extraction methods extract information from the original sample space and make a transformation such that the features are changed [58]. They are typically used in image data. Feature extraction methods can be divided into linear and nonlinear techniques [59]. Representative linear methods include Principal Component Analysis (PCA) [60] and linear discriminant analysis (LDA) [61]. PCA is one of the most classical linear dimensionality reduction methods. The main idea is to find the optimal subspace that represents the data distribution, namely, a mapping matrix consisting of the first n feature vectors corresponding to the largest feature values from the covariance matrix. There are also many nonlinear techniques, such as Kernel PCA [62], Multidimensional scaling (MDS) [63] and Isometric Feature Mapping (Isomap) [64], that perform well on complex nonlinear data [65]. Many studies are conducted to explore simple and efficient feature extraction methods in machine learning, such as [66,67].

For the construction of a framework or a complete framework, the design of the feature extraction method is more important than training a complex classifier [68]. Here, we will introduce a very successful feature extraction model used in image processing. Deep learning is developing quickly, and gets so many state-of-art effects today. CNN [69] is a kind of feed-forward deep neural network with a convolutional structure that performs very well. It is comprised of two parts: an automatic feature extractor and a trainable classifier. A primitive CNN structure, such as that used in LeNet-5, is shown in Fig. 3.

The model inputs image directly, global and local features are extracted via linear convolution layers. After convolution layers, the following usually is non-linear polling layer. The polling layer could reduce the resolution of the extracted features, and there are two methods of polling layer that are usually used in the network. For average polling, it can be seen as a further feature extraction process to reduce computation. For the way of max polling, it can be seen as a feature selection to get the most important point of local features. The complete training process could be observed as a wonderful combination of feature extraction and feature selection, as shown in Fig. 4.

It can obtain very good results because the model considers the intrinsic characteristics of the image. In reference [70], the author proposes a novel CNN–SVM framework in which the CNN model is used to extract and select the discriminative features, and the SVM is used as the super classifier. The fusion model obtains a very high recognition rate on the MNIST handwriting database.

3.2. Feature selection

The aim of feature selection is to select the most important part of the feature subsets under a specific evaluation criterion and retain the original construct and information [18]. It is an integrated optimization problem with high computational expense. Thus, there are works that study how to select local features that represent the sample space rather than global features [51]. Traditional feature selection methods usually compute every feature's score in one area independently, and subsequently, the top n features are selected according to the scores. This type of score is used to evaluate the ability of distinguishing different clusters for a certain feature. Obviously, the methods perform well at binary classification but are not good for multi-classification problems. For big data and high dimensionality, feature selection methods encounter the following challenges: efficiency, universality, implementation ease, and nonlinearity [71]. On the other hand, many increasing unlabeled data with high dimensionality need to be processed in machine learning, which is making unsupervised feature selection an increasingly challenging and important problem [72].

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

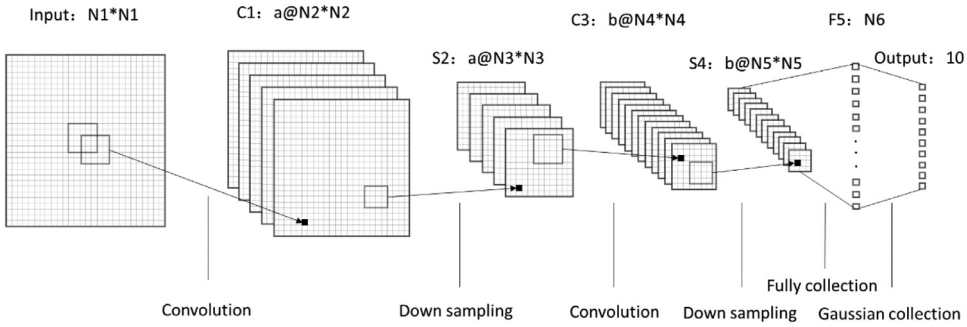

Fig. 3. the CNN architecture.

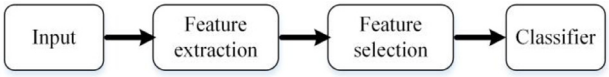

Fig. 4. data flow in CNN.

According to the search models of feature selection, feature selection methods include wrappers, filters and hybrids (ensemble) [73,74]. Wrapper methods are black-box systems that use the current prediction information. Wrapper methods perform well at finding the optimal feature subset to obtain better results than filter methods [75]. However, wrapper methods evaluate the current optimal subset heuristically. When the number of dimensions is high, the computational time and complexity are also high.For example, clustering is the representative method of wrapper methods. However, it has a high computational cost. Wrapper methods also include sequential selection algorithms and heuristic search algorithms [76]. The widely used heuristic approaches are mainly evolutionary algorithms, including Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Genetic Algorithm (GAs) and others [77]. Recent studies have used the weighted A* for the optimization of unsupervised feature selection [78,79]. In contrast with wrapper methods, filter methods use some other indirect measures. Examples include ranking methods and space search methods [73]. These types of methods use general characteristics, such as distance, that are used primarily to select a large part of a feature subset, sometimes even all features. Filter methods are common and easy to extend; examples include the Maximum Variance, Laplacian Score, and Fisher Score [80]. Hybrid methods refer to optimized combinations of filter and wrapper methods. Hybrid methods combine the advantages of filter and wrapper methods and have received much attention as novel feature selection methods. In hybrid methods, also referred to as ensemble methods, the filter method can be considered a pre-processing step; subsequently, the wrapper method is applied to finish the task [76]. Based on the above discussion, the specific methods of feature extraction and feature selection can be broadly summarized as shown in Fig. 5.

3.3. Applications of dimensionality reduction

Dimensionality reduction methods have been used in many real-world applications, including feature selection or feature extraction alone and combinations of feature selection and feature extraction. All types of traditional and improved dimensionality reduction methods combined with advanced machine learning methods yield the desirable effects.

Loderer [81] et al. proposed a training strategy that combines PCA, Local Binary Patterns (LBP) and a clustering algorithm like k-means to select noticeable features automatically for Face Recognition. The classification results for an SVM classifier reveal a balance between storage and accuracy. In image classification, Pighetti [82] et al. considered that SVM is very effective and used widely, so a new framework that selects the fewest effective samples for SVM in fine-grained classification is presented. The Multi-Objective Genetic Algorithm (MOGA) is used to select and optimize effective samples for SVM, where Locality Sensitive Hashing (LSH) is utilized to correct the parameters in the procession of MOGA. The experiments demonstrate that the method achieves state-of-the-art results. Moreover, Peng [74] et al. directly adapted sparsity optimization using the $l_{2,p}$-norm $(0 < p \leq 1)$ for feature selection. The performance on a linear SVM classifier for multi-class classification is improved. To develop tractable algorithms and exploit the theoretical criteria of feature selection, a novel method that combines the mutual information and class label is proposed in [83]. By controlling the criteria of the tractable feature selection method, the computational complexity can be reduced by up to two orders of magnitude. Furthermore, both the classification accuracy and speed are superior to those of the state-of-art baselines. Omara [84] et al. presented a specific geometric feature extraction method for

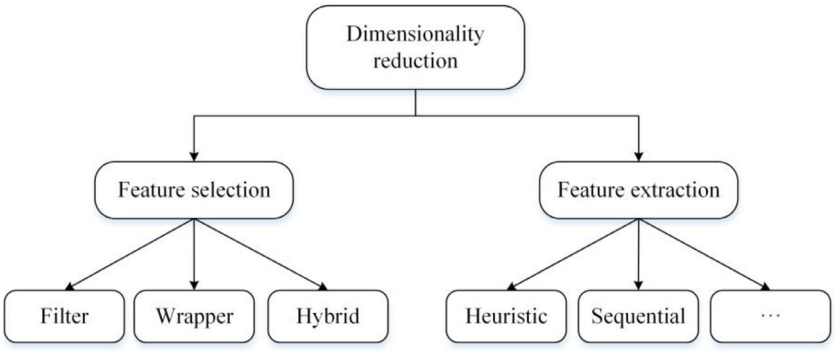

Fig. 5. the specific methods of dimensionality reduction.

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

ear recognition. Chen [85] et al. considered clustering and group-sparsity regularization for feature selection and then proposed a clustering-based multi-task joint feature selection framework for semantic attribute prediction in multi-task learning. Chen [55] et al. integrated feature selection and modeling, discarding indifferent and bad features, with simultaneous construction of the fuzzy rule. Therefore, the model is not a separate system and can obtain a small set of features. A new feature extraction method, Orthogonal Least Squares Regression (OLSR), is presented in [86]. The model constructs the LSR based on the orthogonal constraint such that the discriminative information can be obtained. The recognition accuracy for UCI, face data and footprint data demonstrates that the features extracted from the model are representative. Computer vision is currently a very active research field. Gao [87] et al. presented a novel large-scale image data extraction method, Centered Convolutional Restricted Boltzmann Machines (CCRBM), for scene recognition. The model adds the centered factors to improve the Convolutional Restricted Boltzmann Machines (CRBM) such that the stabilities of the model are enhanced. The visible unit and hidden unit influence each other and optimize continuously. Using a greedy layer-wise training method, the model's generative ability improves. The extensive experiments on large-scale image data show the method's effectiveness at scene recognition. In RNN, the inner states transform is very complex. Løkse [88] et al. applied the dimensionality reduction technique in each state network to obtain better performance. Before the inner state values are transmitted to the next layer, the regularization constraint PCA and kPCA is adapted on these inner states. The improved model has enhanced generalization capability and obtains better performance in experiments. Li [57] et al. considered the wide applications of LDA in dimensionality reduction and its disadvantages, such as the applications being restricted to Gaussian distributions and the number of features extracted being few. The maximum nonparametric margin projection (MNMP) model for feature extraction is proposed. Moreover, the within-class and between-class scatters are used to obtain the discriminative subspace. The experiment demonstrates the efficiency of the method. In multi-label learning, the data often have redundant and noisy features of high dimensionality because there is more than one label associated with one instance [89]. Jian [90] et al. presented a multi-label informed feature selection framework to improve the effectiveness and efficiency of the classification. The high-dimensional multi-labeled feature space is decomposed into a low-dimensional space according to the idea of Latent Semantic Indexing. Then, further features selection is adapted, and the Armijo rule is employed for optimization. Feature selection selects some top features form the mapping data from the original high-dimensional space, in order to remove noisy features as more as possible, Xu [91] improved LDA to optimize the framework of multi-label feature extraction. The new framework is encouraged from two existed multi-label LDA methods. The features got by LDA will be further weighed by considering both feature and label information. Besides, the weights got through maximizing Hilbert–Schmidt independence criterion make the multi-label feature extraction more sufficiently and effectively. Angelo [92] et al. evidenced that the effective feature extraction method applied in a soft computing method was the key for aerospace structure defects in experiments. Tao [17] et al. utilized the PSO to optimize the parameters in the stage of feature extraction on small sample sets. Next, according to the decision level, different feature extraction methods are built to improve the classification accuracy. In [93], the researchers proved that traditional methods that improved the inter-class discrimination maximally in dimensionality reduction would result in overfitting, especially for high-dimensional data. Consequently, a new method that solves the problem of overfitting, named Soft Discriminant Map (SDM), is proposed. An experimental comparison with PCA and LDA demonstrates that this method yields superior performance. Wei [94] et al. mentioned that the visual area V4 in the neural mechanism was responsible for the shape recognition for vision. Therefore, they constructed a V4 neural network for vision tasks based on shape feature extraction. The low-level layers obtain the orientations and edges features. Then, the features around salient points are encoded into a RBM layer to generate representations of the shapes. These shape representations and their corresponding distributions are finally used for object recognition. In the field of clinical medicine, there are many studies regarding dimensionality reduction [18,54]. Mi [76] et al. proposed a robust wrapper algorithm for the prediction of tumor treatment outcomes using a small dataset and obtained a promising accuracy. The study presented in [54] demonstrates that when the sample number is sufficiently large, feature extraction methods perform better than feature selection methods. Embedded feature selection methods are good choices for small data sets. In the challenging face recognition problem, Wang [19] et al. presented an ensemble learning method with random sampling of subspace feature vectors and optimal parameters to obtain higher accuracy. The high dimensionality of the samples is first reduced via PCA. Then, the subspace of the feature vectors is obtained using the LDA method. Through the combination algorithms and multiple LDA classifiers, the accuracy of the recognition gets better. Thanh [95] et al. proposed a dynamic online sparse coding features selection mechanism for the field of robotics based on reinforcement learning in the real-world environment's high-dimensional spaces. In the brain-computer interface (BCI) field, Luo [96] et al. presented a feature selection approach named dynamic frequency feature selection (DFFS) to select the most useful features. The random forest (RF) algorithm is then used for classification. In [97], feature extraction based on block projection is applied on multi-unit recordings, namely spike sorting, to extract information of extracellular action potentials.

4. Combinations of sample selection and dimensionality reduction

Until now, this paper has reviewed the theories of sample selection and dimensionality reduction in detail, respectively. We have also introduced a large number of these methods combined with machine learning separately. For all types of special questions, such as classification and prediction, either sample selection or feature reduction alone is used in most cases. However, recently, many researchers have combined sample selection with dimensionality reduction to obtain a better optimal solution. In this section, we collect various methods and applications that fuse sample selection and dimensionality reduction in either sequential or simultaneous manners.

4.1. Sequential combinations

In text classification, Pang [98] et al. proposed a new model that combines sample selection and feature selection based on [27]. Noise samples are selected according to the representativeness score. Then, a feature weight adjustment method is used according to the distribution of samples to further select discriminative features for the classifier. The experimental results on SVM demonstrate that the model could discard many irrelevant samples effectively and obtain a better classification performance. Thung [99] et al. employed feature selection and sample selection sequentially on multi-modality incomplete data [100] by data grouping and multi-task learning. First, the incomplete data matrix is grouped into some overlapping submatrices. Then, the feature selection process is performed using multi-task sparse regression to remove redundant and noisy features. Next, the framework selects representative

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

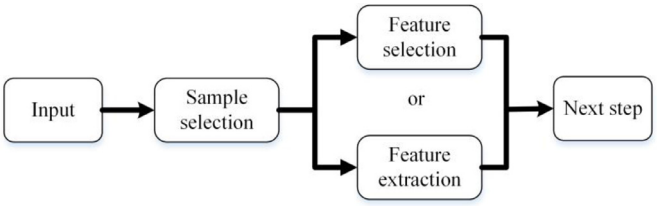

Fig. 6. sequential combination framework1.

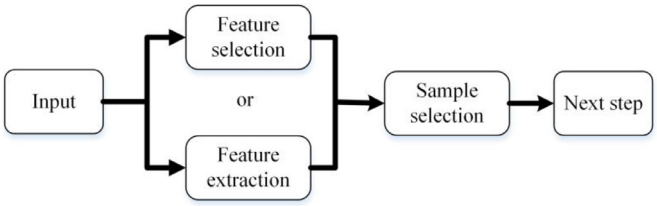

Fig. 7. sequential combination framework2.

samples using a multi-task learning algorithm. The framework exhibits improvements in terms of both accuracy and speed of classification. Xia [101] et al. considered two main problems in domain adaptation, labeling and instance adaptation, to present a comprehensive framework named feature ensemble plus sample selection (SS-FE). The model trains the Naïve Bayes (NB) classifier first by feature extraction, and then a sample selection method based on PCA (PCA-SS) is adopted. The effects of feature extraction and PCA-SS are both good, but SS-FE exhibits better performance in an experimental comparison. Xuan [29] et al. presented an effective sample selection method on pre-miRNAs (miSampleSelect), which is based on two-stage clustering and feature selection. The training set is first clustered according to the stem similarity of pre-miRNAs. Then, 27 discriminative features are selected from the whole ensemble of 48 features for each sample. The third step implements another clustering algorithm, which is based on the distribution of samples, and the most effective samples are selected according the density. The miSampleSelect method can resolve the problem that caused by the imbalance of the training samples by the SVM classifier. Neagoe [102] et al. applied ACO to the recognition of space imagery. ACO Band Selection (ACO-BS) is first used to reduce the dimensionality by creating bands of multispectral images. Then, ACO Training Label Purification (ACO-TLP) is used to select the training sample subset that is the most informative. The model can reduce as many as samples and is effective in experiments. In addition to the applications, there are also theoretical analyses of sample selection and dimensionality reduction. To obtain an upper bound on the samples number for PCA learning, Hanneke [103] solved a long-standing open problem by completely eliminating the logarithmic factor based on the work of Hans Simon. All fusion methods mentioned above are sequential combinations of dimensionality reduction and sample selection. The frameworks mentioned above are illustrated in Figs. 6 and 7.

Figs. 6 and 7 show two different sequences of the use of dimensionality reduction and sample selection. They also represent two kinds of processing way for the original data. For the first way, the model first considers the distribution of the data, then takes the appropriate method to get features for the following task. For Fig. 7, the model usually extracts all samples' features, then the features of every sample would replace the original sample to the next sample selection step. For different data and different tasks, researchers would select suited way as shown before.

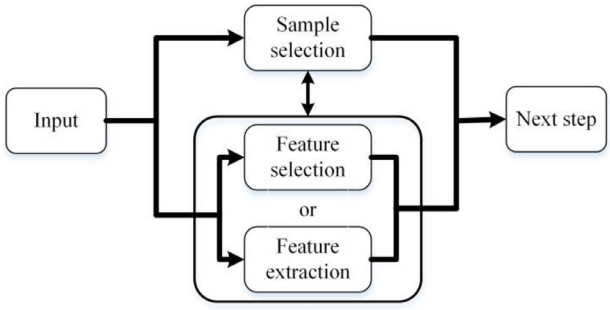

Fig. 8. new simultaneous framework.

4.2. Simultaneous combinations

There are also studies that find that the optimized subset obtained from the two independent processes of sample selection and dimensionality reduction may be not the optimal result because of the overlap of the two sub-problems [104,1]. A new framework that combines sample selection and dimensionality reduction simultaneously is shown in Fig. 8.

Ng [1] et al. thought that the most important aspect of developing a combination of sample selection and feature selection is to find a same measure to evaluate the importance of each such that there would be no overlap. They improved a new methodology, RBFNN-SM (Radial Basis Function Neural Networks-sensitivity measure), which can be used to select useful samples only or perform sample selection and feature selection together (RBFMV-SM-SS). The sensitivity measure (SM) is used widely in neural networks because it can evaluate the features and samples in the same manner. On the other hand, RBF is easy to understand and use. Thus, the ensemble architecture has obvious advantages. In addition, experimental results demonstrate that the model can select a few features and samples to obtain a high accuracy. Mohsenzadeh [105] et al. presented a joint structure called relevance sample feature machine (RSFM) to select the relevance samples and features simultaneously. The model is an extension of the basic relevance vector machine (RVM), which can generate sparse subsets from the original training samples, namely the relevance samples. RSFM can obtain a higher classification accuracy than the classical RVM because it can obtain the sparse relevance feature vectors and relevance samples simultaneously. Increased sparsity can decrease the complexity and avoid overfitting of the model such that the model is more efficient. The improved algorithm has better generalization ability and removes noisy and irrelevant features effectively, but it has the drawback that the model can only obtain local convergence. Therefore, Adeli [104] et al. proposed a joint feature-sample selection (JFSS) structure that uses a sparse linear regression model to remove redundant samples and irrelevant features simultaneously. JFSS adopts the single optimization problem instead of the marginal likelihood maximization to change the problem of the simple convex optimization problem. Thus, the formulation is easier to solve than RSFM. Additionally, considering that there is still some random noise after selecting the best features and samples, the whole model has a de-noising process. The model is applied for PD diagnosis, and the results demonstrate its reliability and good accuracy. Based on the disadvantages of RSFM, Mohsenzadeh et al. also proposed an improved model called Incremental Relevance Sample-Feature Machine (IRSFM) [75]. The new variant changes the marginal likelihood maximization approach that optimizes the initialized construction of the model, adding kernel functions step by step and only computing the relevant parameters by the current features and samples. As a result, IRSFM can be applied to large-scale data sets more easily and is more

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

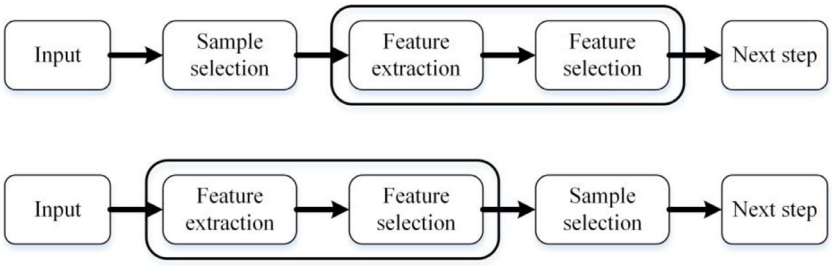

Fig. 9. new combination mechanism 1.

effective than RSFM. For surveillance videos, the image classification task is often based on a large number of video clips. A serious problem is the large space and time complexity caused by the large-scale samples and dense features. Zhang [4] et al. presented a joint feature selection and sample selection method to simultaneously solve these two problems. A compact hashing function is used to create binary codes based on the rules that the samples obtain large interclass and small intra-class distances. In this way, the feature space is reduced significantly, and redundant samples are removed in a low-rank manner. Moreover, the author also adapted a kernel method to address nonlinear and complex images. Experiments on three standard face recognition datasets demonstrate that the proposed method is effective and efficient.

5. Discussion and conclusion

The fast development of machine learning makes a big progress of the social such as Face recognition, disease diagnosis, speech recognition, image classification and any other real problems. Sample selection and dimensionality reduction techniques are extremely important in large-scale data analysis, especially in machine learning [7,106]. This paper briefly reviews the problems that are caused by the accumulation of large-scale data. 'Large-scale' means that the quantity and dimensionality of the samples obtained are becoming increasingly greater. As a result, more irrelevant, redundant, incomplete and noisy data with ultrahigh attributes are generated. Accordingly, problems such as overfitting, high computational complexity and low computational speed are becoming more serious, especially in the active field of image data processing. Thus, the concepts and current situations of sample selection and dimensionality reduction are outlined. The contribution of this paper is to summarize the different types of applications of sample selection and dimensionality reduction combined with various methods of machine learning, such as clustering, classification, rough sets, fuzzy sets, soft computing, heuristic algorithms, etc., that have been developed in recent years. Among these methods, we focus on CNN in particular. The development of CNN has enabled great progress for many active research fields such as image recognition and classification. Its complete training mechanism, which combines feature extraction and feature selection, has yielded state-of-the-art results. From this paper, it is clear that these two techniques have been applied widely, but as yet there is no universal method for sample selection and dimensionality reduction that can be applied to all problems. Every special problem typically adopts one unique method to make progress relative to previous work. According to the existing literatures summarized in this paper, we found that sample selection and dimensionality reduction methods either use only one of these techniques or use all sequentially or simultaneously. Facing the challenging problems caused by large-scale data, it is possible that a general mechanism that fuses sample selection methods with dimensionality reduction methods, such as those illustrated Figs. 9 and 10, would obtain better performance.

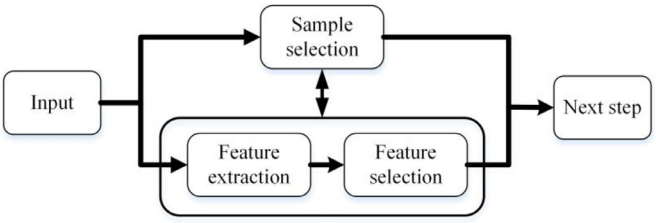

Fig. 10. new combination mechanism 2.

We have a idea that whether the structure that combines two dimensionality reduction methods, feature extraction and feature selection, like CNN could be used in data processing with sample selection at the same time. Figs. 9 and 10 show the sequential and simultaneous mechanisms, respectively.

From the conclusions of this paper, we can see that the data for traditional machine learning is different from CNN. In CNN, the model usually processes the image data directly, that is to say the network would extract the useful features by itself, and the samples are almost as more as possible. But in traditional machine learning methods, the application fields are wilder and the data processing methods like sample selection and dimensionality reduction are often used. The applications of machine learning methods in this paper contain vary kinds of fields, such as medicine, biology, agriculture, aerospace, traffic and so on. And the most purposes are to classify and predict. In recent years, the multi-task and multi-label classification get popular. As referred before, sample selection and dimensionality reduction play the essential role for the successful applications. In every study field, the way of sample selection and dimensionality reduction are far different because of different data form, different tasks and different simulation environments. That is why we collect so many methods with all kinds of forms in this paper.

The sample selection could remove the redundant data and make the training set balanced, so it can train a more robust classifier. It is worth noting that with the data accumulation more and more, unsupervised sample selection methods and sample selection bias method get important gradually. As for feature selection and feature extraction, more and more studies keep the point on that how to get the features considering the relationship with the current task. At last, the paper makes attentions on the combination of these two important techniques in machine learning.

In the nowadays large-scale data processing, sample selection and dimensionality reduction are both the necessary steps. Because the different granularities of data representation, the most selected order is sample selection first, and then is dimensionality reduction. Sample selection often focuses on the number and the distribution of the whole original data set. On the contrary, dimensionality reduction would process the samples into deeper representation as features or dimensions. In order to understand the data essentially, there are some researches finding the way to

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

adopt the dimensionality reduction and sample selection at the same time. In this way, it can not only reduce the time further but also select better data and features for the task. Through the conclusions above, we have known that the dimensionality reduction includes feature selection and feature extraction, but almost all applications only use one of them. Feature selection and feature extraction are two different ways for getting data features, but they can help each other. For example, we consider that whether the features got by feature extraction are all effective? And can we further select the extracted features through feature selection methods? And next, how to combine sample selection, feature extraction and feature selection may be the new direction for the large-scale data processing.

Acknowledgments

This work was supported by the National Natural Science Foundation [Nos. 61672522, 41704115], and Guangxi High School Key Laboratory of Complex System and Computational Intelligence [No. 2017CSCI01].

References

[1] W.W.Y. Ng, D.S. Yeung, I. Cloete, Input sample selection for RBF neural network classification problems using sensitivity measure, systems, man and cybernetics, 2003, in: Proceedings of the IEEE International Conference, 3, 2003, pp. 2593–2598, doi:10.1109/ICSMC.2003.1244274.
[2] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1–2) (1997) 245–271 (97)00063-5, doi:10.1016/S0004-3702.
[3] Y. Zhai, Y.S. Ong, I. Tsang, Making trillion correlations feasible in feature grouping and selection, IEEE Trans. Pattern Anal. Mach. Intell. 38 (12) (2016) 2472–2486, doi:10.1109/TPAMI.2016.2533384.
[4] M. Zhang, R. He, D. Cao, Z. Sun, T. Tan, Simultaneous feature and sample reduction for image-set classification, 30th AAAI Conference on Artificial Intelligence, AAAI press, 2016, pp. 1401–1407.
[5] J.V. Hulse, Data Quality in Data Mining and Machine Learning, Florida Atlantic University, 2007.
[6] X.Z. Wang, L.C. Dong, J.H. Yan, Maximum ambiguity-based sample selection in fuzzy decision tree induction, IEEE Trans. Knowl. Data Eng. 24 (8) (1997) 1491–1505, doi:10.1109/TKDE.2011.67.
[7] C. Garc A-Osorio, D. Haro-Garc, A. Aida, A-Pedrajas.N. Garc, Democratic instance selection: a linear complexity instance selection algorithm based on classifier ensemble concepts, Artif. Intell. 174 (5–6) (2010) 410–441, doi:10.1016/j.artint.2010.01.001.
[8] P.E. Hart, The condensed nearest neighbour rule, IEEE Trans. Inf. Theory 14 (3) (1968) 515–516, doi:10.1109/TIT.1968.1054155.
[9] D.W. Aha, D. Kibler, M.K. Albert, Instance-based learning algorithms, Mach. Learn. 6 (1) (1991) 37–66, doi:10.1007/BF00153759.
[10] D.J.C. Mackay, Information-based objective functions for active data selection, Neural Comput. 4 (4) (1992) 590–604, doi:10.1162/neco.1992.4.4.590.
[11] H. Xu, H. Zha, R.C. Li, M.A. Davenport, Active manifold learning via gershgorin circle guided sample selection, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI Press, 2015, pp. 3108–3114.
[12] A. Moayedikia, K.L. Ong, Y.L. Boo, W.G. Yeoh, R. Jense, Feature selection for high dimensional imbalanced class data using harmony search, Eng. Appl. Artif. Intell. 57 (2017) 38–49, doi:10.1016/j.engappai.2016.10.008.
[13] C.O.S. Sorzano, J. Vargas, A.P. Montano, A survey of dimensionality reduction techniques, arXiv:1403.2877, (2014)1-35.
[14] F. Wang, J. Sun, Survey on Distance Metric Learning and Dimensionality Reduction in Data Mining, Kluwer Academic Publishers, 2015.
[15] H. Xie, J. Li, H. Xue, A Survey of Dimensionality Reduction Techniques Based on Random Projection, 2017.
[16] M. Arellano, S. Bonhomme, Sample Selection in Quantile Regression: A Survey, Working Papers, 2017.
[17] C.G. Tao, L.L. Zhao, X.H. Su, P.J. Ma, PSO-based feature extraction for high dimension small sample, in: Proceedings of the IEEE Fifth International Conference on Advanced Computational Intelligence, 8267, 2012, pp. 229–233, doi:10.1109/ICACI.2012.6463157.
[18] J. Krawczuk, T. Łukaszuk, The feature selection bias problem in relation to high-dimensional gene data, Artif. Intell. Med. 66 (2016) 63–71, doi:10.1016/j.artmed.2015.11.001.
[19] X. Wang, X. Tang, Random sampling for subspace face recognition, Int. J. Comput. Vis. 70 (1) (2006) 91–104, doi:10.1007/s11263-006-8098-z.
[20] N. Zhang, T. Xiao, A sample selection algorithm based on maximum entropy and contribution, in: Proceedings of the International Conference on Machine Learning and Cybernetics, ICMLC 2010, 1, Qingdao, China, 2010, pp. 397–402. July 11-14, 2010, Proceedings. doi:10.1109/ICMLC.2010.5581031.
[21] E. Elhamifar, G. Sapiro, S.S. Sastry, Dissimilarity-based sparse subset selection, IEEE Trans. Pattern Anal. Mach. Intell. 38 (11) (2016) 2182–2197, doi:10.1109/TPAMI.2015.2511748.
[22] Y.H. Zhou, L. Huang, X.I. Mao-Long, Quantum-behaved particle swarm optimization algorithm with random selection of optimal individual, J. Comput. Appl. 4 (6) (2009) 189–193.
[23] W. Sun, A.P. Lin, H.S. Yu, Q.K. Liang, G.H. Wu, All-dimension neighborhood based particle swarm optimization with randomly selected neighbors, Inf. Sci. Int. J. 405 (C) (2017) 141–156, doi:10.1016/j.ins.2017.04.007.
[24] W. Guicquero, P. Vandergheynst, T. Laforest, A. Dupret, On adaptive pixel random selection for compressive sensing, IEEE Signal Inf. Process. Netw. 234 (2015) 701–708, doi:10.1109/GlobalSIP.2014.7032211.
[25] B. Du, L. Zhang, Random-selection-based anomaly detector for hyperspectral imagery, IEEE Trans. Geosci. Remote Sens. 49 (5) (2011) 1578–1589, doi:10.1109/TGRS.2010.2081677.
[26] B.Q. Gong, K. Grauman, F. Sha, Connecting the dots with landmarks: discriminatively learning domain-invariant features for unsupervised domain adaptation, JMLR W&CP 28 (1) (2013) 222–230.
[27] Y.X. Liao, X.Z. Pan, A new method of training sample selection in text classification, Int. Workshop Educ. Technol. Comput. Sci. 1 (2010) 211–214, doi:10.1109/ETCS.2010.621.
[28] M. Chellasamy, P.A.T. Ferre, M. Humlekrog Greve, Automatic training sample selection for a multi-evidence based crop classification approach, Int. Arch. Photogramm. Remote Sens. & S XL-7 (7) (2014) 63–69, doi:10.5194/isprsarchives-XL-7-63-2014.
[29] P. Xuan, M.Z. Guo, L.L. Shi, J. Wang, X.Y. Liu, W.B. Li, Y.P. Han, Two-stage clustering based effective sample selection for classification of pre-miRNAs, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, 10, 2010, pp. 549–552, doi:10.1109/BIBM.2010.5706626.
[30] X. Li, M. Fang, J.J. Zhang, J.Q. Wu, Sample selection for visual domain adaptation via sparse coding, Signal Process. Image Commun. 44 (2016) 92–100, doi:10.1016/j.image.2016.03.009.
[31] J.T. Xia, M.Y. He, Y.Y. Wang, Y. Feng, A fast training algorithm for support vector machine via boundary sample selection, in: Proceedings of the International Conference on Neural Networks and Signal Processing, 1, 2004, pp. 20–22, doi:10.1109/ICNNSP.2003.1279203.
[32] X.T. Niu, Fcm-lssvm based on training sample selection, Metall. Mining Ind. 7 (9) (2015) 751–757.
[33] J.H. Zhai, C. Li, T. Li, Sample selection based on k-l divergence for effectively training SVM, IEEE Syst. Man Cybern. 8215 (2013) 4837–4842, doi:10.1109/SMC.2013.823.
[34] H.W. Hao, R.R. Jiang, Training sample selection method for neural networks based on nearest neighbor rule, Acta Autom. Sin. 33 (33) (2007) 1247–1251, doi:10.1360/aas-007-1247.
[35] D.G. Chen, X. Zhang, E.C.C. Tsang, Y.P. Yang, Sample selection with rough set, in: Proceedings of the International Conference on Machine Learning and Cybernetics, 1, 2010, pp. 291–295, doi:10.1109/ICMLC.2010.5581051.
[36] E. Santiago-Ramirez, J.A. Gonzalez-Fraga, E. Gutierrez, O. Alvarez-Xochihua, Optimization-based methodology for training set selection to synthesize composite correlation filters for face recognition, Signal Process. Image Commun. 43 (2016) 54–67, doi:10.1016/j.image.2016.02.002.
[37] Y. Yang, Z. Ma, F. Nie, X. Chang, A.G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, Int. J. Comput. Vis. 113 (2) (2015) 113–127, doi:10.1007/s11263-014-0781-x.
[38] Y. Guo, L. Ma, F. Zhu, F.J. Liu, Selecting training samples from large-scale remote-sensing samples using an active learning algorithm, Comput. Intell. Intell. Syst. 575 (2015) 40–51, doi:10.1007/978-981-10-0356-1_5.
[39] W.W. Yuan, Y.K. Han, D.H. Guan, S.Y. Lee, Y.K. Lee, Initial training data selection for active learning, in: Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, 2011, doi:10.1145/1968613.1968619.
[40] R.J. Goudie, S. Mukherjee, A Gibbs sampler for learning DAGs, J. Mach. Learn. Res. 17 (30) (2016) 1–39.
[41] J.J. Heckman, Sample selection bias as a specification error, Econometrica 47 (1979) 153–161, doi:10.2307/1912352.
[42] B. Zadrozny, Learning and evaluating classifiers under sample selection bias, in: Proceedings of the Twenty-first International Conference on Machine Learning, 114, 2004, doi:10.1145/1015330.1015425.
[43] D. Wu, D.Z. Lin, L. Yao, W.J. Zhang, Correcting sample selection bias for image classification, in: Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering, 2008, pp. 1214–1220, doi:10.1109/ISKE.2008.4731115.
[44] B. Schölkopf, J. Platt, T. Hofmann, Mixture regression for covariate shift, in: Proceedings of the International Conference on Neural Information Processing Systems, MIT Press, 2006, pp. 1337–1344.
[45] A.T. Smith, C. Elkan, Making generative classifiers robust to selection bias, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2007) 657–666, doi:10.1145/1281192.1281263.
[46] R. Romero, E.L. Iglesias, L. Borrajo, Building biomedical text classifiers under sample selection bias, in: Proceedings of the International Symposium on Distributed Computing and Artificial Intelligence, 91, Berlin Heidelberg, Springer, 2011, pp. 11–18.
[47] N. Krautenbacher, F.J. Theis, C. Fuchs, Correcting classifiers for sample selection bias in two-phase case-control studies, Comput. Math. Methods Med. (2017), doi:10.1155/2017/7847531.

[48] M. Tan, I.W. Tsang, L. Wang, Towards ultrahigh dimensional feature selection for big data, J. Mach. Learn. Res. 15 (1) (2014) 1371–1429.
[49] K.Q. Weinberger, L.K. Saul, An introduction to nonlinear dimensionality reduction by maximum variance unfolding, in: Proceedings of the National Conference on Artificial Intelligence, 2, AAAI Press, 2006, pp. 1683–1686.
[50] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (0) (1997) 273–324 (97) 00043-X, doi:10.1016/S0004-3702.
[51] N. Armanfard, J.P. Reilly, M. Komeili, Local feature selection for data classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (6) (2016) 1217–1227, doi:10.1109/TPAMI.2015.2478471.
[52] D. Ramyachitra, M. Sofia, P. Manikandan, Interval-value based particle swarm optimization algorithm for cancer-type specific gene selection and sample classification, Gen. Data 5 (2015) 46–50, doi:10.1016/j.gdata.2015.04.027.
[53] W. Pindah, A. Seman, S. Nordin, M.S.M. Said, Review of dimensionality reduction techniques using clustering algorithm in reconstruction of gene regulatory networks, in: Proceedings of the International Conference on Computer, Communications, and Control Technology, 5, 2015, pp. 1031–1034, doi:10.1109/I4CT.2015.7219560.
[54] S. Pölsterl, S. Conjeti, N. Navab, A. Katouzian, Survival analysis for high-dimensional, heterogeneous medical data: exploring feature extraction as an alternative to feature selection, Artif. Intell. Med. 72 (2016) 1–11, doi:10.1016/j.artmed.2016.07.004.
[55] Y.C. Chen, N.R. Pal, I.F. Chung, An integrated mechanism for feature selection and fuzzy rule extraction for classification, IEEE Tran. Fuzzy Syst. 20 (4) (2012) 683–698, doi:10.1109/TFUZZ.2011.2181852.
[56] N. Chumerin, M.M. Van Hulle, Comparison of Two Feature Extraction Methods Based on Maximization of Mutual Information, Machine Learning for Signal Processing, Proceedings of the 2006, IEEE Signal Processing Society Workshop on 2006; 343-348, doi:10.1109/MLSP.2006.275572.
[57] B. Li, J. Du, X.P. Zhang, Feature extraction using maximum nonparametric margin projection, Neurocomputing 188 (2016) 225–232, doi:10.1016/j.neucom.2014.11.105.
[58] C. Lee, D.A. Landgrebe, Feature extraction based on decision boundaries, IEEE Trans. Pattern Anal. Mach. Intell. 15 (4) (1993) 388–400, doi:10.1016/j.patcog.2017.12.010.
[59] D. Araujo, A.D. Neto, A. Martins, J. Melo, Comparative study on dimension reduction techniques for cluster analysis of microarray data, Hist. Eur. Idea 11 (1) (2011) 1835–1842, doi:10.1109/IJCNN.2011.6033447.
[60] I.T. Jolliffe, Principal Component Analysis, 87, Springer, Berlin, 1986, pp. 41–64, doi:10.1007/b98835.
[61] A.M. Martínez, A.C. Kak, Pca versus lda, Pattern Analysis & Machine Intelligence IEEE Transactions 2001; 23(3-4):228-233, doi:0.1109/34.908974.
[62] V.N. Vapnik, The Nature of Statistical Learning Theory, Neural Networks IEEE Transactions 1995; 10(5):988 – 999, doi:10.1007/978-1-4757-3264-1.
[63] F.W. Young, R.M. Hamer, Multidimensional scaling: history, theory, and applications, Hillsdale, New Jersey, 1988.
[64] J.B. Tenenbaum, V.D. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323, doi:10.1126/science.290.5500.2319.
[65] Jihan, K., Rafic, Y., Dimensionality Reduction on Hyperspectral Images: A Comparative Review Based on Artificial Datas, LECT NOTES COMPUT SC, 2011 4th International Congress on 2011; 4:1875-1883, doi:10.1109/CISP.2011.6100531.
[66] L. Zhang, Q. Zhang, B. Du, D. Tao, J. You, Robust manifold matrix factorization for joint clustering and feature extraction, 31th AAAI Conference on Artificial Intelligence, AAAI press, 2017, pp. 1662–1668.
[67] J. Wangni, N. Chen, Nonlinear feature extraction with max-margin data shifting, 30th AAAI Conference on Artificial Intelligence, AAAI press, 2016, pp. 2208–2214.
[68] J. Li, J. Zhao, K. Lu, Joint feature selection and structure preservation for domain adaptation, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI press, 2016, pp. 1697–1703.
[69] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324, doi:10.1109/5.726791.
[70] X.X. Niu, C.Y. Suen, A novel hybrid cnn–svm classifier for recognizing handwritten digits, Pattern Recognit. 45 (4) (2012) 1318–1325, doi:10.1016/j.patcog.2011.09.021.
[71] A. Barbu, Y. She, L. Ding, G. Gramajo, Feature selection with annealing for big data learning, Eprint Arxiv 39 (2) (2014) 272–286.
[72] F. Nie, W. Zhu, X. Li, Unsupervised feature selection with structured graph optimization, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, 2016, pp. 1302–1308.
[73] H. Mhamdi, F. Mhamdi, Feature selection methods on biological knowledge discovery and data mining: a survey, Int. Workshop Database (2014) 46–50, doi:10.1109/DEXA.2014.26.
[74] H. Peng, Y. Fan, Direct sparsity optimization based feature selection for multiclass classification, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI press, 2016, pp. 1918–1924.
[75] Y. Mohsenzadeh, H. Sheikhzadeh, S. Nazari, Incremental relevance sample-feature machine: a fast marginal likelihood maximization approach for joint feature selection and classification, Pattern Recognit. 60 (2016) 835–848, doi:10.1016/j.patcog.2016.06.028.
[76] H. Mi, C. Petitjean, B. Dubray, P. Vera, R. Su, Robust feature selection to predict tumor treatment outcome, Artif. Intell. Med. 64 (3) (2015) 195–204, doi:10.1016/j.artmed.2015.07.002.
[77] N. Abd-Alsabour, A review on evolutionary feature selection, IEEE (2014), doi:10.1109/EMS.2014.28.
[78] H. Arai, K. Xu, C. Maung, H. Schweitzer, Weighted A* algorithms for unsupervised feature selection with provable bounds on suboptimality, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, 2016, pp. 4194–4195.
[79] H. Arai, C. Maung, K. Xu, H. Schweitzer, Unsupervised feature selection by heuristic search with provable bounds on suboptimality, 30th AAAI Conference on Artificial Intelligence, AAAI press, 2016, pp. 666–672.
[80] C.J.C. Burges, Dimension reduction: a guided tour, Found. Trends® Mach. Learn. 2 (4) (2010) 262–286, doi:10.1561/2200000002.
[81] M. Loderer, J. Pavlovicova, M. Feder, M. Oravec, Data dimension reduction in training strategy for face recognition system, in: Proceedings of the International Conference on Systems, Signals and Image Processing, 2014, pp. 263–266.
[82] R. Pighetti, D. Pallez, F. Precioso, Improving SVM training sample selection using multi-objective evolutionary algorithm and LSH, Comput. Intell. (2015) IEEE Symposium 2015, doi:10.1109/SSCI.2015.197.
[83] L. Lefakis, F. Fleuret, Jointly informative feature selection made tractable by gaussian modeling, J. Mach. Learn. Res. 17 (182) (2016) 1–39.
[84] I. Omara, F. Li, H.Z. Zhang, W.M. Zuo, A novel geometric feature extraction method for ear recognition, Expert Syst. Appl. 65 (2016) 127–135, doi:10.1016/j.eswa.2016.08.035.
[85] L. Chen, B. Li, Clustering-based joint feature selection for semantic attribute prediction, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, AAAI press, 2016, pp. 3338–3344.
[86] H.F. Zhao, Z. Wang, F.P. Nie, Orthogonal least squares regression for feature extraction, Neurocomputing 216 (2016) 200–207, doi:10.1016/j.neucom.2016.07.037.
[87] J.Y. Gao, J.F. Yang, G.H. Wang, M.G. Li, A novel feature extraction method for scene recognition based on centered convolutional restricted boltzmann machines, Neurocomputing 11 (2) (2016) 14–19, doi:10.1016/j.neucom.2016.06.055.
[88] S. Løkse, F.M. Bianchi, R. Jenssen, Training echo state networks with regularization through dimensionality reduction, Cogn. Comput. 9 (3) (2017) 364–378, doi:10.1007/s12559-017-9450-z.
[89] J.H. Liu, Y.J. Lin, Y. Kang, C.X. Wang, Online multi-label group feature selection, Knowl. Based Syst. (2017), doi:10.1016/j.knosys.2017.12.008.
[90] L. Jian, J. Li, K. Shu, H. Liu, Multi-label informed feature selection, in: Proceedings of the International Joint Conference on Artificial Intelligence, AAAI Press, 2016, pp. 1627–1633.
[91] J. Xu, A weighted linear discriminant analysis framework for multi-label feature extraction, Neurocomputing 275 (2017) 107–120, doi:10.1016/j.neucom.2017.05.008.
[92] G. D'Angelo, S. Rampone, Feature extraction and soft computing methods for aerospace structure defect classification, Measurement 85 (2016) 192–209, doi:10.1016/j.measurement.2016.02.027.
[93] R. Liu, D.F. Gillies, Overfitting in linear feature extraction for classification of high-dimensional image data, Pattern Recognit. 53 (C) (2016) 73–86, doi:10.1016/j.patcog.2015.11.015.
[94] H. Wei, Z. Dong, V4 neural network model for shape-based feature extraction and object discrimination, Cognit. Comput. 7 (6) (2015) 753–762, doi:10.1007/s12559-017-9450-z.
[95] T.N. Thanh, Z. Li, T.V. Silander, T.Y. Leong, Online feature selection for model-based reinforcement learning, in: Proceedings of the International Conference on Machine Learning, 2013, pp. 498–506.
[96] J. Luo, Z.R. Feng, J. Zhang, N. Lu, Dynamic frequency feature selection based approach for classification of motor imageries, Comput. Biol. Med. 75 (2016) 45–53, doi:10.1016/j.compbiomed.2016.03.004.
[97] S.C. Wu, A.L. Swindlehurst, Direct feature extraction from multi-electrode recordings for spike sorting, Digit. Signal Process. (2018), doi:10.1016/j.dsp.2018.01.016.
[98] X. Pang, Y. Liao, A text classification model based on training sample selection and feature weight adjustement, in: Proceedings of the International Conference on Advanced Computer Control, 3, 2010, pp. 294–297, doi:10.1109/ICACC.2010.5486615.
[99] K.H. Thung, C.Y. Wee, P.T. Yap, D. Shen, Neurodegenerative disease diagnosis using incomplete multi-modality data via matrix shrinkage and completion, Neuroimage 91 (2) (2014) 386–400, doi:10.1016/j.neuroimage.2014.01.033.
[100] D. Zhang, D. Shen, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease, Neuroimage 59 (2) (2012) 895–907, doi:10.1016/j.neuroimage.2011.09.069.
[101] R. Xia, C. Zong, X. Hu, E. Cambria, Feature ensemble plus sample selection: domain adaptation for sentiment classification, IEEE Intell. Syst. 28 (3) (2013) 10–18, doi:10.1109/MIS.2013.27.
[102] V.E. Neagoe, E.C. Neghina, Feature selection with ant colony optimization and its applications for pattern recognition in space imagery, International Conference on Communications (COMM), IEEE, IEEE, 2016, pp. 101–104.
[103] S. Hanneke, The optimal sample complexity OF PAC learning, J. Mach. Learn. Res. 17 (38) (2016) 1–15.
[104] E. Adeli, F. Shi, L. An, C.Y. Wee, G.R. Wu, T. Wu, D.G. Shen, Joint feature-sample selection and robust diagnosis of parkinson's disease from mri data, Neuroimage 141 (2016) 206–219, doi:10.1016/j.neuroimage.2016.05.054.
[105] Y. Mohsenzadeh, H. Sheikhzadeh, A.M. Reza, N. Bathaee, M.M. Kalayeh, The relevance sample-feature machine: a sparse bayesian learning approach to

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100

joint feature-sample selection, IEEE Trans. Cybern. 43 (6) (2013) 2241–2254, doi:10.1109/TCYB.2013.2260736.
[106] S. Xiang, X. Shen, J. Ye, Efficient nonconvex sparse group feature selection via continuous and discrete optimization, Artif. Intell. 224 (2015) 28–50, doi:10.1016/j.artint.2015.02.008.

Xinzheng Xu is currently an associate professor at China University of Mining and Technology, China. He received his B.S. degree from Shandong University of Science and Technology in 2002, and his M.S. degree from Xiamen University in 2005. He received his Ph.D. degree from China University of Mining and Technology in 2012. His research interests include pattern recognition, machine learning, and neural networks et al.

Tianming Liang is currently a master candidate at China University of Mining and Technology, China. He received his bachelor's degree from China University of Mining and Technology in 2015. His research interests include pattern recognition and machine learning.

Jiong Zhu is currently a master candidate at China University of Mining and Technology, China. He received his bachelor's degree from China University of Mining and Technology in 2015. His research interests include neural networks and machine learning.

Dong Zheng is currently a master candidate at China University of Mining and Technology, China. He received his bachelor's degree from China University of Mining and Technology in 2017. His research interests include pattern recognition and machine learning.

Tongfeng Sun is currently an associate professor at China University of Mining and Technology, China. He received his master's degree and Ph.D. degree from China University of Mining and Technology in 2004 and 2012, respectively. His research interests include intelligent information processing, pattern recognition, and machine learning et al.

Please cite this article as: X. Xu et al., Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.02.100