Robust Feature Selection on Incomplete Data

Wei Zheng[1], **Xiaofeng Zhu**[1*], **Yonghua Zhu**[2], **Shichao Zhang**[1]
[1] Guangxi Key Lab of Multi-source Information Mining & Security,
Guangxi Normal University, Guilin, 541004, China
[2] Guangxi University, Nanning, 530004, China
zwgxnu@163.com, seanzhuxf@gmail.com, yhzhu66@qq.com, zhangsc@gxnu.edu.cn

Abstract

Feature selection is an indispensable preprocessing procedure for high-dimensional data analysis, but previous feature selection methods usually ignore sample diversity (*i.e.,* every sample has individual contribution for the model construction) and have limited ability to deal with incomplete data sets where a part of training samples have unobserved data. To address these issues, in this paper, we firstly propose a robust feature selection framework to relieve the influence of outliers, and then introduce an indicator matrix to avoid unobserved data to take participation in numerical computation of feature selection so that both our proposed feature selection framework and exiting feature selection frameworks are available to conduct feature selection on incomplete data sets. We further propose a new optimization algorithm to optimize the resulting objective function as well as prove our algorithm to converge fast. Experimental results on both real and artificial incomplete data sets demonstrated that our proposed method outperformed the feature selection methods under comparison in terms of clustering performance.

1 Introduction

Feature selection is one of popular dimensionality reduction techniques for dealing with high-dimensional data due to its interpretability. Based on the ways of the model construction, previous feature selection methods are widely partitioned into three categories [Chandrashekar and Sahin, 2014], *i.e.,* filter model [He *et al.*, 2005], wrapper model [Unler *et al.*, 2011], and embedded model [Peng and Fan, 2017; Zhu *et al.*, 2015; 2017b]. Filter model selects important features from all the features without a training model, and thus is simpler and more efficient than both wrapper model and embedded model. Embedded model integrates the process of feature selection with the training model into a unified framework, *i.e.,* automatically selecting the features in the training process, and has been shown more effective than both filter model and wrapper model.

Given a feature matrix, each of whose row and column, respectively, represent a sample and a feature, feature selection can be regarded as removing redundancy or noise from the vertical direction of the feature matrix, *i.e.,* removing unimportant features, with the assumption that different features have different contribution for data analysis, *i.e.,* feature diversity. Actually, different samples contain different influence for data analysis as well, *i.e.,* sample diversity [Huber, 2011; Nie *et al.*, 2010]. However, a few feature selection methods were focused on considering to remove the influence of the redundancy or noise from the horizontal direction of the feature matrix, *i.e.,* removing or relieving the influence of outliers [Zhu *et al.*, 2014a; Huang *et al.*, 2016].

In practical applications, incomplete data sets can often be found, where some elements of a part of the samples are unobserved/incomplete/missed. In particular, the ratio of incomplete samples may approach to 90% in industrial data [Little and Rubin, 2014]. The popular strategies for dealing with incomplete data sets include discard strategy and imputation strategy [Zhang *et al.*, 2017]. The discard strategy discards incomplete samples (*i.e.,* the samples containing unobserved elements) and uses the observed samples (*i.e.,* all the elements of the samples are observed) for data analysis. The discard strategy ignores the information in incomplete samples (*i.e.,* observed information in incomplete samples) as well as lacks enough samples so that outputting unreliable models [Van Hulse and Khoshgoftaar, 2014; Zhang *et al.*, 2017]. The imputation strategy firstly imputes unobserved elements with guessed values, and then uses all the samples in the data set for data analysis. Compared to discard strategy, the imputation strategy utilizes the observed information in incomplete samples, but the imputed values may be noise or un-useful since unobserved information is never known [Zhang *et al.*, 2006; Van Hulse and Khoshgoftaar, 2014; Doquire and Verleysen, 2012]. The study of data analysis on incomplete data has widely been focused on all kinds of real applications, but few study was focused on conducting feature selection on incomplete data sets.

To address the above issues, in this paper, we propose a new method to conduct robust unsupervised feature selection on incomplete data sets. Specifically, our proposed method

*Xiaofeng Zhu and Wei Zheng had equally contribution on this work, Corresponding author: Xiaofeng Zhu.

$\mathbf{X}$	the feature matrix of the training data
$\mathbf{x}$	a vector of $\mathbf{X}$
$\mathbf{x}^i$	the *i*-th row of $\mathbf{X}$
$\mathbf{x}_j$	the *j*-th column of $\mathbf{X}$
$x_{i,j}$	the element of the *i*-th row and the *j*-th column of $\mathbf{X}$
$\|\|\mathbf{X}\|\|_F$	the Frobenius norm of $\mathbf{X}$, *i.e.,* $\|\|\mathbf{X}\|\|_F = \sqrt{\sum_{i,j} x_{i,j}^2}$
$\|\|\mathbf{X}\|\|_{2,1}$	the $\ell_{2,1}$-norm of $\mathbf{X}$, *i.e.,* $\|\|\mathbf{X}\|\|_{2,1} = \sum_i \sqrt{\sum_j x_{i,j}^2}$
$\mathbf{X}^T$	the transpose of $\mathbf{X}$
$tr(\mathbf{X})$	the trace of $\mathbf{X}$
$\mathbf{X}^{-1}$	the inverse of $\mathbf{X}$

Table 1: The notations used in this paper.

conducts data analysis by 1) using all the observed information (including the observed information in incomplete samples, whose unobserved elements are not considered to be imputed), and 2) relieving or even removing the influence of the redundancy and noise from both the horizontal and vertical directions of the data set. To do this, firstly, we propose a robust feature selection framework to firstly select a subset of samples (Most Important (MI) samples for short) from the data set to construct an initial feature selection model, and then other MI samples in the left samples were automatically selected to take participation in the revision of this initial feature selection model with the former MI samples. The robustness and generalization ability of the initial feature selection model is gradually improved with the increasing MI samples until all the samples are used up or the model achieves stability. As a result, outliers will be selected later than MI samples or never be selected to take participation in the construction of the feature selection model. Secondly, we introduce an indicator matrix to avoid unobserved values involved into the numerical computation of feature selection models so that both our proposed robust feature selection framework and existing feature selection frameworks can be directly applied on incomplete data sets. Furthermore, we propose a new optimization algorithm to optimize the resulting objective function and also prove the convergence of our proposed optimization algorithm.

To our knowledge, our proposed method is the first work to integrate feature selection, dealing with unobserved data, and robust statistics [Huber, 2011; Huang *et al.*, 2016] in a unified framework since previous works were only designed to focus on a part of these three tasks [Nie *et al.*, 2010; Doquire and Verleysen, 2012; Huang *et al.*, 2016]. Moreover, it is very easily to extend our proposed framework to the tasks such as conducting unsupervised feature selection using other embedded modles, supervised feature selection [Song *et al.*, 2007; Zhu *et al.*, 2018] and semi-supervised feature selection [Zhao and Liu, 2007], on incomplete data sets.

2 Approach

We summarize all the used notations in this paper in Table 1.

2.1 Robust Feature Selection

By considering the efficiency and effectiveness, filter model and embedded model are two common feature selection models. In particular, sparse based embedded model (sparse feature selection for short) has widely been used in real applications since these methods select features by pushing the weight coefficients of unimportant features to small values or even zeros and the weight coefficients of important features to large values. Specifically, given a feature matrix $\mathbf{X} = [\mathbf{x}^1; ...; \mathbf{x}^n] = [\mathbf{x}_1, ..., \mathbf{x}_d] \in \mathbb{R}^{n\times d}$, traditional unsupervised sparse feature selection methods [Bengio *et al.*, 2009; Huang *et al.*, 2011; Zhu *et al.*, 2014b; 2015] can be formulated as:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{X}\mathbf{W}||_F^2 + \lambda||\mathbf{W}||_{2,1} \tag{1}$$

where λ is a non-negative tuning parameter and $\mathbf{W} \in \mathbb{R}^{d\times d}$ is the weight coefficient matrix.

Eq. (1) equivalently considers every sample, so the resulting model is easily influenced by outliers [Nie *et al.*, 2010; Huber, 2011]. To address this issue, Nie *et al.* proposed to replace Frobenius norm in Eq. (1) by an $\ell_{2,1}$-norm loss function [Nie *et al.*, 2010; Zhu *et al.*, 2016; 2017a] to have:

$$\min_{\mathbf{W}} ||\mathbf{X} - \mathbf{X}\mathbf{W}||_{2,1} + \lambda||\mathbf{W}||_{2,1} \tag{2}$$

In Eq. (2), the residual of $\mathbf{X} - \mathbf{X}\mathbf{W}$ (*i.e.,* $||\mathbf{X} - \mathbf{X}\mathbf{W}||_{2,1}$) does not have a squared operator, and thus outliers have less influence compared to Frobrenious norm [Nie *et al.*, 2010]. Compared to Eq. (1), Eq. (2) considers sample diversity, but it lacks flexibility to avoid the influence of outliers. Moreover, both Eq. (1) and Eq. (2) cannot be used for dealing with incomplete data sets.

In this paper, we employ the theory of robust statistics [Huber, 2011] such as self-paced learning [Kumar *et al.*, 2010; Lee and Grauman, 2011], half-quadratic minimization [Nikolova and Ng, 2005], and M-estimation (*i.e.,* maximum likelihood type estimation) [Negahban *et al.*, 2009] to replace the $\ell_{2,1}$-norm loss function in Eq. (2) with predefined robust loss functions, aim at achieving flexibly robust feature selection. In robust statistics, functions can be constructed so that they can firstly select MI samples from all the samples to construct an initial model, and the left MI samples are automatically selected from the left samples joining with the former MI samples to revise the robustness and generalization ability of this initial model gradually until all the samples are used up and this model is not further improved [Kumar *et al.*, 2010; Lee and Grauman, 2011]. By replacing the $\ell_{2,1}$-norm loss function in Eq. (2) with a general robust loss function $\phi(\cdot)$, the robust feature selection framework can be formulated as:

$$\min_{\mathbf{W}} \phi(||\mathbf{X} - \mathbf{X}\mathbf{W}||_F) + \lambda||\mathbf{W}||_{2,1} \tag{3}$$

A number of robust loss functions have been designed in the literature [Fan *et al.*, 2017; Nikolova and Chan, 2007; Nikolova and Ng, 2005] such as Cauchy function, ℓ_1-ℓ_2 function, Welsch M-estimator, and Geman–McClure estimator. Every robust loss function $\phi(\cdot)$ has individual characteristics different from the others and can be flexibly selected in real applications, *e.g.,* tuning the parameters of $\phi(\cdot)$ to achieve flexibility. In this way, Eq. (2) can be regarded as a special issue of Eq. (3).

2.2 Robust Feature Selection on Incomplete Data

A few feature selection methods have been designed to deal with incomplete data sets. A possible method of conducting feature selection on incomplete data sets is to construct

feature selection models only using observed samples by discarding incomplete samples. Obviously, the higher the unobserved ratio of the data sets, the lower the robustness and generalization ability of the feature selection model is. To address this issue, an alternative method is to firstly impute unobserved values with imputation methods (such as mean-value method [Little and Rubin, 2014] and k Nearest Neighbors (kNN) imputation method [Zhang *et al.*, 2017]) and then conduct feature selection using all the samples including the imputed incomplete samples. However, the imputation strategy for feature selection is usually unpractical. Specifically, if the imputation models are estimated perfectly, then the constructed feature selection models may be over-fitting. Otherwise, the models will be under-fitting with a bad imputation. Moreover, we have no idea on the ground truth of unobserved data so that it is difficult to evaluate imputation models.

Based on above observations, it may be unnecessary to impute unobserved values for feature selection. To make the use of the observed information in incomplete samples, we propose to use an indicator matrix $\mathbf{D} \in \mathbb{R}^{n\times d}$ (which has the same size as the feature matrix $\mathbf{X}$) to avoid unobserved values involving the numerical computation of feature selection, we thus propose the following objective function to conduct robust feature selection on incomplete data sets:

$$\min_{\mathbf{W}} \phi(||\mathbf{D}\circ(\mathbf{X}-\mathbf{X}\mathbf{W})||_F) + \lambda||\mathbf{W}||_{2,1} \tag{4}$$

where $\mathbf{D} = [\mathbf{d}^1; ...; \mathbf{d}^n]$ is an indicator matrix. Specifically, if the *i*-th row and the *j*-th column element $x_{i,j}$ is unobserved, the value of $d_{i,j}$ is 0, otherwise 1. The symbol $\circ$ is a Hadamard product operator that conducts the element-wise multiplication between two same size matrices.

According to Eq. (4), our proposed method conducts feature selection by using all available information (without imputing unobserved values) as well as reliving the influence of outliers. Moreover, the definition of $\mathbf{D}$ can be applied in any sparse feature selection model for conducting robust unsupervised feature selection using other embedded models, robust supervised feature selection, and robust semi-supervised feature selection, on incomplete data sets.

2.3 Proposed Objective Function

Although a number of robust loss functions have been reported in the literature, the resulting objective functions in these predefined robust functions may be optimized difficultly/ineffieciently or even non-convex, so the half-quadratic technique was designed to address this issue by introducing an auxiliary variable, *i.e.,*

Lemma 1. *Given a fixed scalar z, if a differentiable function $\phi(z)$ satisfies four conditions listed in [Nikolova and Chan, 2007], then the following holds,*

$$\phi(z) = \inf_{\mathbf{z}\in\mathbb{R}}\{\mathbf{v}z^2 + \psi(\mathbf{v})\} \tag{5}$$

where $\mathbf{v}$ is a variable determined by the minimization function of the dual potential function $\psi(\mathbf{v})$ of $\phi(z)$.

In this paper, for a fixed scalar z, we select a concrete robust loss function, *i.e.,* Geman–McClure estimator [Geman and Mcclure, 1987],

$$\phi(z) = \frac{\mu z^2}{\mu + z^2} \tag{6}$$

to replace $\phi(\cdot)$ in Eq. (4) to have:

$$\min_{\mathbf{W}} \frac{\mu||\mathbf{D}\circ(\mathbf{X}-\mathbf{X}\mathbf{W})||_F^2}{\mu + ||\mathbf{D}\circ(\mathbf{X}-\mathbf{X}\mathbf{W})||_F^2} + \lambda||\mathbf{W}||_{2,1} \tag{7}$$

where μ is used to tune the number of MI samples in each iterations [Geman and Mcclure, 1987]. Eq. (7) can be optimized by any gradient methods. By the consideration of efficient and scalable optimization, we apply Lemma 1 to Eq. (7), and then obtain our final objective function:

$$\min_{\mathbf{W},\mathbf{v}} \sum_{i=1}^{n}(v_i||\mathbf{d}^i\circ(\mathbf{x}^i-\mathbf{x}^i\mathbf{W})||_2^2 + \mu(\sqrt{v_i}-1)^2) + \lambda||\mathbf{W}||_{2,1} \tag{8}$$

where $\psi(\mathbf{v}) = \mu(\sqrt{\mathbf{v}}-1)^2$, $\mathbf{v} = [v_1, ..., v_n] \in \mathbb{R}^n$ is an auxiliary variable, μ and λ are two non-negative tuning parameters. According to Lemma 1, Eq. (7) and Eq. (8) are equivalent with respect to the optimization of $\mathbf{W}$, and are more flexible than either Eq. (2) or Eq. (1) since $\mathbf{v}$ controls the selection of MI samples as well as enables to define additional constraints to make the model (*e.g.,* Eq. (8)) more flexible. Specifically, by considering the optimization result of v_i (*i.e.,* $v_i = \frac{\mu^2}{(\mu+\|\mathbf{d}^i\circ(\mathbf{x}^i-\mathbf{x}^i\mathbf{W})\|_2^2)^2}$), if the residual (*i.e.,* $\|\mathbf{d}^i\circ(\mathbf{x}^i-\mathbf{x}^i\mathbf{W})\|_2^2$) is small, then the *i*-th sample $\mathbf{x}^i$ can be regarded a MI sample and thus its weight v_i will be large. By contrast, the weight of an outlier is small due to its large residual. Moreover, in the alternative iterations of optimizing Eq. (8), the value of every v_i will vary with the iteratively updated $\mathbf{W}$. In this way, the initial model constructed in the former iterations will be updated gradually with different weights for each sample until the model achieves stability. Furthermore, the number of MI samples in every iteration can be flexibly controlled by the tuning parameter μ according to the practical demand [Geman and Mcclure, 1987], *i.e.,* $\mu \gg \tilde{f}$ where $\tilde{f}$ is the average of $\{f^i = ||\mathbf{d}^i\circ(\mathbf{x}^i-\mathbf{x}^i\mathbf{W})||_2^2, i = 1, ..., n\}$.

The optimization of Eq. (8) is challenging due to introducing an auxiliary variable $\mathbf{v}$, the convex but no-smooth constraint on the variable of $\mathbf{W}$ (*i.e.,* $||\mathbf{W}||_{2,1}$), and introducing the indicator matrix $\mathbf{D}$. In this paper, we utilize the alternative optimization strategy (*i.e.,* the framework of Iteratively Reweighted Least Squares (IRLS) [Daubechies *et al.*, 2008]) to optimize Eq. (8), *i.e.,* update $\mathbf{v}$ by fixing $\mathbf{W}$ and update $\mathbf{W}$ by fixing $\mathbf{v}$.

2.4 Convergence Analysis

By denoting $\mathbf{v}^{(t)}$ and $\mathbf{W}^{(t)}$, respectively, as the *t*-th iteration result of $\mathbf{v}$ and $\mathbf{W}$, we change Eq. (8) to the following:

$$\begin{aligned} J(\mathbf{W}^{(t)}, \mathbf{v}^{(t)}) = &\sum_{i=1}^{n}(v_i^{(t)}||\mathbf{d}^i\circ(\mathbf{x}^i-\mathbf{x}^i\mathbf{W}^{(t)})||_2^2 \\ &+\mu(\sqrt{v_i^{(t)}}-1)^2) + \lambda||\mathbf{W}^{(t)}||_{2,1} \end{aligned} \tag{9}$$

With the fixed $\mathbf{W}^{(t)}$ and according to half-quadratic theory [Nikolova and Ng, 2005], we have:

$$J(\mathbf{W}^{(t)}, \mathbf{v}^{(t+1)}) \leq J(\mathbf{W}^{(t)}, \mathbf{v}^{(t)}) \tag{10}$$

While fixing $\mathbf{v}^{(t+1)}$, we have the following inequality according to the IRLS framework:

$$J(\mathbf{W}^{(t+1)}, \mathbf{v}^{(t+1)}) \leq J(\mathbf{W}^{(t)}, \mathbf{v}^{(t+1)}) \tag{11}$$

By integrating Eq. (10) with (11), we have

$$J(\mathbf{W}^{(t+1)}, \mathbf{v}^{(t+1)}) \leq J(\mathbf{W}^{(t)}, \mathbf{v}^{(t)}) \tag{12}$$

According to Eq. (12), Eq. (9) is non-increasing at each iteration in optimization. Hence, our proposed method finally achieves converged in the optimized process.

3 Experimental Analysis

3.1 Experimental Settings

In our experiments, we separated an incomplete data set into two sets, *i.e.,* Incomplete Set (IS) including all incomplete samples and Observed Set (OS) including all observed samples. Firstly, we denoted the method of conducting k-means clustering with original features of OS as Baseline. We also employed a filter feature selection method and two embedded feature selection methods. Three comparison algorithms are summarized as follows:

- Laplacian Score (LPscore) [He *et al.*, 2005] is a unsupervised filter feature selection model that evaluates the importance of each features based on the Laplacian scores of the data.
- Regularized Self-Representation (RSR) [Zhu *et al.*, 2015] firstly uses a feature-level self-representation to reconstruct each feature by the linear relationship of all features, and then uses an $\ell_{2,1}$-norm regularization to sparsely penalize the regression matrix to conduct feature selection. RSR can be considered as the basic algorithm of our proposed method.
- General Framework for Sparsity Regularized (GSR) [Peng and Fan, 2017] is a general sparsity feature selection model that can achieve the different combination of different loss functions and sparse regularizations in the same framework by tuning parameters. By comparing to other sparsity feature selection methods, GSR is more flexible and could be robust to the outliers.

Moreover, we use our proposed Robust Feature Selection (RFS) framework in Eq.(3) as one of our comparison algorithms.

Secondly, we used OS to impute unobserved values in IS by the imputation method, *i.e.,* mean-value imputation method (Mean) and kNN imputation method (kNN) [Zhang *et al.*, 2017], and then conducted feature selection on OS $\bigcup$ IS by GSR, *i.e.,* GSR_mean and GSR_kNN, followed by conducting k-means clustering on OS with selected features.

Thirdly, we used our proposed Eq. (8) to conduct feature selection on OS $\bigcup$ IS, and then conducted k-means clustering on OS with selected features.

We employed the 10-fold cross-validation scheme to repeat every method on a data set ten times. We reported the average results of ten times, each of which is the average of 20 clustering results. We set the ranges of the parameters of the comparison methods according to the corresponding literature, and set the ranges of the parameter (*i.e.,* λ) of our method in Eq. (8) as $\{10^{-3}, 10^{-2}, ..., 10^{3}\}$. We further set the number of clusters in k-means clustering as the number of real classes of the data sets.

We used two evaluation metrics (*i.e.,* Accuracy (ACC) and Normalized Mutual Information (NMI)) to evaluate the performance of all the methods on four real incomplete data sets. We listed the definitions of these two valuation metrics as follows:

- ACC indicates the percentage of correctly classified samples in the total samples, *i.e.,*

$$ACC = \frac{N_c}{N} \tag{13}$$

  where *N* denotes the number of the samples and N_c is the number of the correctly classified samples.

- NMI uncovers a correlation between the obtained labels and the real labels, *i.e.,*

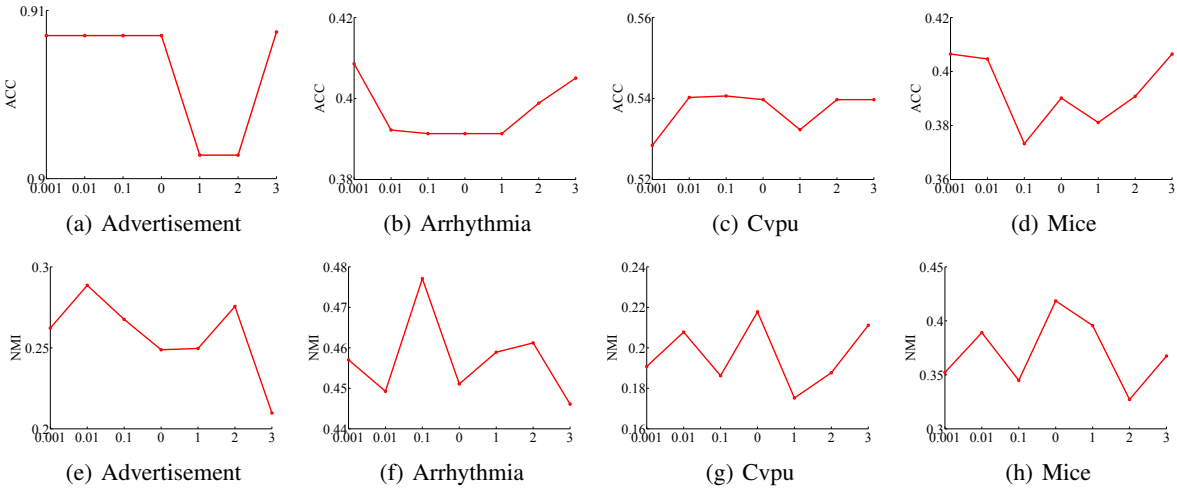
$$NMI = \frac{2I(\mathbf{X}, \mathbf{Y})}{(H(\mathbf{X}) + H(\mathbf{Y}))} \tag{14}$$

  where $I(\mathbf{X}, \mathbf{Y})$ denotes the mutual information between the samples and the labels and $H(\cdot)$ is the corresponding entropy.

3.2 Real Incomplete Data Sets

We downloaded four real incomplete data sets from UCI website, *i.e.,* Advertisement, Arrhythmia, Cvpu, and Mice, with unobserved ratios of samples, are 28.60%, 84.96%, 94.99% and 48.89%, respectively. We reported clustering performance of all the feature selection methods with different ratios of selected features, *i.e.,* $\{10\%, 20\%, ..., 90\%\}$ of all the features in Figure 3 and 4.

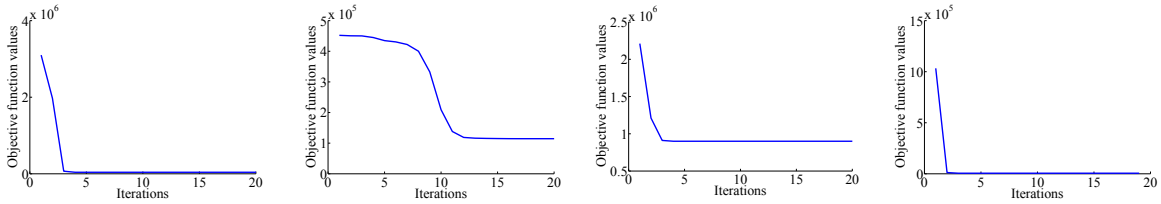
From Figure 3 and 4, firstly, our method achieved the best clustering performance, followed by GSR_kNN, GSR_Mean, RFS, GSR, RSR, LapScore and Baseline. For example, our proposed method on average improved by 2.15% and 18.70%, compared to the best comparison method (*i.e.,* GSR_kNN) and the worst comparison method (*i.e.,* Baseline), in term of ACC results. The reason may be that our method uses all the available information and relieves the influence of outliers. It is noteworthy that imputation methods (*i.e.,* GSR_Mean and GSR_kNN) use all the available information but not taking the influence of outliers into account, while feature selection methods (*i.e.,* LapScore, RSR, GSR, and RFS) only use the information in observed samples and not use the observed information in incomplete samples. Secondly, our method achieved the maximal improvement, compared to these four feature selection methods on data set Cvpu because this data set only has 5.01% observed samples. This indicates that reliable feature selection models need enough training samples.



(a) Advertisement (b) Arrhythmia (c) Cvpu (d) Mice

(e) Advertisement (f) Arrhythmia (g) Cvpu (h) Mice

Figure 1: The ACC (upper row) and NMI (bottom row) results of our method with varied parameter' setting (*i.e.,* λ), where four real incomplete data sets were kept 50% features.



(a) Advertisement (b) Arrhythmia (c) Cvpu (d) Mice

Figure 2: The convergence of our proposed method, where four real incomplete data sets were kept 50% features.

3.3 Parameters' Sensitivity and Convergence

We listed both the ACC and the NMI variations of our proposed method at different values of the parameter λ in Figure 1. Obviously, our proposed method is sensitive to parameters' setting a little, and may achieve stability and good clustering in some ranges. For example, the ACC results of our method varied from 40.65% to 37.33%, but achieved the best results while $\lambda \in [10^{-3}, 10^{-2}]$ on the data set Mice. This enables our method to achieve reliable clustering performance via easy parameter tuning. Figure 2 reported the variations of the objective function value of Eq. (8) at different iterations, and clearly showed that our proposed method (*i.e.,* Eq. (8)) can be efficiently optimized, *i.e.,* convergence within about 15 iterations.

4 Conclusion

This paper proposed a novel robust feature selection framework to conduct unsupervised feature selection on incomplete data sets. Specifically, we employed robust statistics to consider sample diversity for conducting feature selection, and used an indicator matrix to avoid unobserved values taking participation in the process of feature selection as well as making the use of all the available information. Experimental results showed that our method achieved the best clustering results, compared to the methods under comparison. We will extend our framework to conduct supervised/semi-supervised feature selection on incomplete data sets in our future work.

Acknowledgements

This work was supported in part by the Nation Natural Science Foundation of China (Grants No: 61573270 and 61672177), Innovation Project of Guangxi Graduate Education (Grant No: YCSW2018093), the Guangxi High Institutions Program of Introducing 100 High-Level Overseas Talents, the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing, and the Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (18-A-01-01).

References

[Bengio *et al.*, 2009] Samy Bengio, Fernando Pereira, Yoram Singer, and Dennis Strelow. Group sparse coding. In *NIPS*, pages 82–89, 2009.

[Chandrashekar and Sahin, 2014] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[Daubechies *et al.*, 2008] Ingrid Daubechies, Ronald Devore, Massimo Fornasier, and Sinan Gunturk. Iteratively

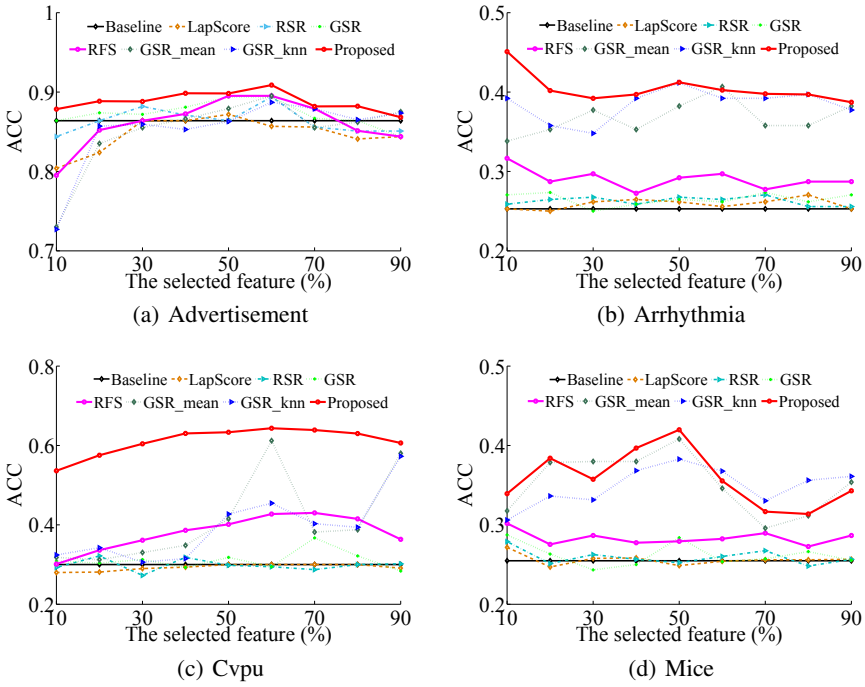


Figure 3: The ACC results of all the methods on different data sets.

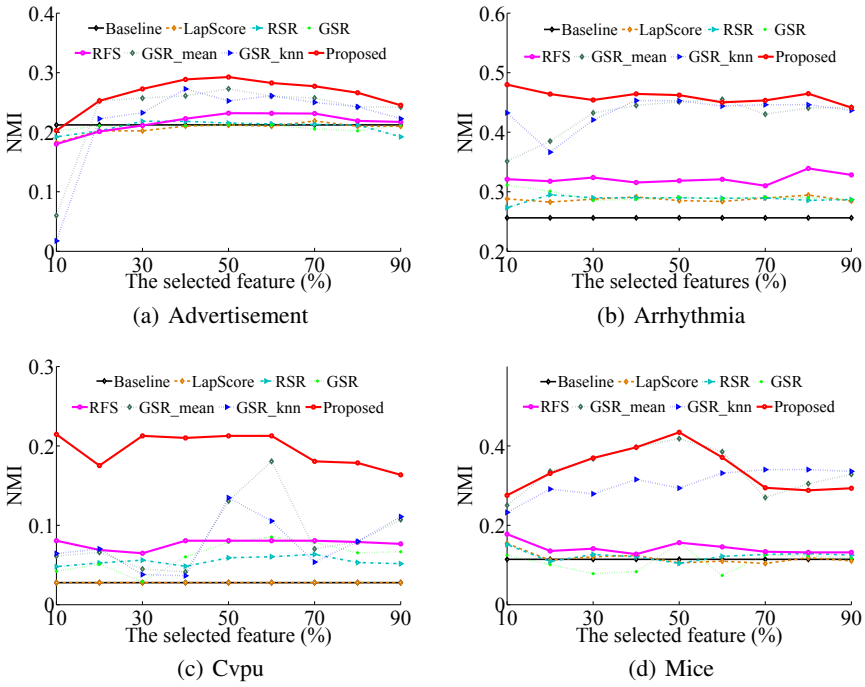


Figure 4: The NMI results of all the methods on different data sets.

reweighted least squares minimization for sparse recovery. *Communications on Pure & Applied Mathematics*, 63(1):1–38, 2008.

[Doquire and Verleysen, 2012] Gauthier Doquire and Michel Verleysen. Feature selection with missing data using mutual information estimators. *Neurocomputing*,

90:3–11, 2012.

[Fan *et al.*, 2017] Yanbo Fan, Ran He, Jian Liang, and Bao-Gang Hu. Self-paced learning: an implicit regularization perspective. In *AAAI*, pages 1877–1883, 2017.

[Geman and Mcclure, 1987] S. Geman and D. E. Mcclure. Statistical methods for tomographic image reconstruction. *Proc Session of the Ici Bulletin of the Ici*, lii-4, 1987.

[He *et al.*, 2005] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. In *NIPS*, pages 507–514, 2005.

[Huang *et al.*, 2011] Junzhou Huang, Tong Zhang, and Dimitris Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research*, 12(Nov):3371–3412, 2011.

[Huang *et al.*, 2016] Dong Huang, Ricardo Cabral, and Fernando De la Torre. Robust regression. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):363–375, 2016.

[Huber, 2011] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. 2011.

[Kumar *et al.*, 2010] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *NIPS*, pages 1189–1197, 2010.

[Lee and Grauman, 2011] Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, pages 1721–1728, 2011.

[Little and Rubin, 2014] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[Negahban *et al.*, 2009] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS*, pages 1348–1356, 2009.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[Nikolova and Chan, 2007] M Nikolova and R. H. Chan. The equivalence of half-quadratic minimization and the gradient linearization iteration. *IEEE Transactions on Image Processing*, 16(6):1623–7, 2007.

[Nikolova and Ng, 2005] Mila Nikolova and Michael K Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 27(3):937–966, 2005.

[Peng and Fan, 2017] Hanyang Peng and Yong Fan. A general framework for sparsity regularized feature selection via iteratively reweighted least square minimization. In *AAAI*, pages 2471–2477, 2017.

[Song *et al.*, 2007] Le Song, Alex Smola, Arthur Gretton, Karsten M Borgwardt, and Justin Bedo. Supervised feature selection via dependence estimation. In *ICML*, pages 823–830, 2007.

[Unler *et al.*, 2011] Alper Unler, Alper Murat, and Ratna Babu Chinnam. mr 2 pso: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Information Sciences*, 181(20):4625–4641, 2011.

[Van Hulse and Khoshgoftaar, 2014] Jason Van Hulse and Taghi M Khoshgoftaar. Incomplete-case nearest neighbor imputation in software measurement data. *Information Sciences*, 259(3):596–610, 2014.

[Zhang *et al.*, 2006] Shichao Zhang, Yongsong Qin, Xiaofeng Zhu, Jilian Zhang, and Chengqi Zhang. Optimized parameters for missing data imputation. In *PRICAI*, pages 1010–1016, 2006.

[Zhang *et al.*, 2017] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. Efficient knn classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems,DOI: 10.1109/TNNLS.2017.2673241*, 2017.

[Zhao and Liu, 2007] Zheng Zhao and Huan Liu. Semi-supervised feature selection via spectral analysis. In *SDM*, pages 641–646, 2007.

[Zhu *et al.*, 2014a] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. Multi-modality canonical feature selection for alzheimer's disease diagnosis. In *MICCAI*, pages 162–169, 2014.

[Zhu *et al.*, 2014b] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *CVPR*, pages 1542–1549, 2014.

[Zhu *et al.*, 2015] Pengfei Zhu, Wangmeng Zuo, Lei Zhang, Qinghua Hu, and Simon C. K. Shiu. Unsupervised feature selection by regularized self-representation. *Pattern Recognition*, 48(2):438–446, 2015.

[Zhu *et al.*, 2016] Yingying Zhu, Xiaofeng Zhu, Minjeong Kim, Dinggang Shen, and Guorong Wu. Early diagnosis of alzheimer's disease by joint feature selection and classification on temporally structured support vector machine. In *MICCAI*, pages 264–272, 2016.

[Zhu *et al.*, 2017a] Pengfei Zhu, Wencheng Zhu, Qinghua Hu, Changqing Zhang, and Wangmeng Zuo. Subspace clustering guided unsupervised feature selection. *Pattern Recognition*, 66:364–374, 2017.

[Zhu *et al.*, 2017b] Yingying Zhu, Xiaofeng Zhu, Minjeong Kim, Daniel Kaufer, and Guorong Wu. A novel dynamic hyper-graph inference framework for computer assisted diagnosis of neuro-diseases. In *IPMI*, pages 158–169. Springer, 2017.

[Zhu *et al.*, 2018] Pengfei Zhu, Qian Xu, Qinghua Hu, Changqing Zhang, and Hong Zhao. Multi-label feature selection with missing labels. *Pattern Recognition*, 74:488–502, 2018.