

Feature Selection Based on Sparse Imputation

Jin Xu, Yafeng Yin, and Hong Man

Department of Electrical
and Computer Engineering
Stevens Institute of Technology
Hoboken, New Jersey, USA, 07030
Email: {jxu4, yyin1, hman}@stevens.edu

Haibo He

Department of Electrical, Computer
and Biomedical Engineering
University of Rhode Island
Kingston, RI, USA, 02881
Email: he@ele.uri.edu

Abstract—Feature selection, which aims to obtain valuable feature subsets, has been an active topic for years. How to design an evaluating metric is the key for feature selection. In this paper, we address this problem using imputation quality to search for the meaningful features and propose feature selection via sparse imputation (FSSI) method. The key idea is utilizing sparse representation criterion to test individual feature. The feature based classification is used to evaluate the proposed method. Comparative studies are conducted with classic feature selection methods (such as Fisher score and Laplacian score). Experimental results on benchmark data sets demonstrate the effectiveness of FSSI method.

I. INTRODUCTION

Feature selection is an important technique in machine learning and data mining. How to select the useful features is a key factor in many applications, such as pattern recognition [1] and computer vision [2]. There are many benefits of feature selection: alleviating the curse of dimensionality, enhancing learning process, improving data visualization and optimizing prediction performance.

In feature selection, there are three models: filter models [3], wrapper model [4] and embedded model. Filter model and wrapper model are popular in recent research. Filter model utilized different metrics to evaluate the individual feature, and remove some features before the prediction process. In wrapper method, the prediction results (or change of the results) of a model are used to measure the value of a feature. The computation costs limit the application of wrapper model on large data sets.

How to design a meaningful evaluation metrics is the key for a good filter model. There are various metrics to build filters. Normally the metrics are kinds of relationship between the features and labels. Two popular filter metrics are mutual information [3] and correlation [5]. In [3], max-relevance and min-redundancy feature subset are realized based the mutual information of the training data. Correlation [5] based feature selection method, which is simple and fast to execute, is successfully applied on continuous class problems. There are also other effective filter metrics in recent researches. Class separability [6] is applied in a high dimensional kernel model and feature selection is carried on to maximize the separability. In [7], error probability is considered as discriminating power, and it has been utilized to design feature selection.

Regarding the labeled training data and unlabeled training data. Feature selection can also be grouped as supervised feature selection and unsupervised feature selection. Supervised feature selection evaluates the relationship between the feature values and the label values. Fisher score [8] ranks the discriminate ability of individual feature according the labels, which is a simple and effective feature selection method. Unsupervised feature selection measures the feature similarity or local information. Laplacian score [9] evaluates the geometrical properties in the feature sets, which is a efficient unsupervised feature selection method. In this paper, Fisher score and Laplacian score are used in the comparison experiment.

Sparse representation (coding) [10], which rebuilds a signal by combination as less as possible atoms from a dictionary, is a new method to acquire and represent signals. In details, a data (signal) $\mathbf{y} \in \mathbb{R}^m$ is sparse representation $\mathbf{y} = \mathbf{D}\mathbf{x}$ by a dictionary $\mathbf{D} \in \mathbb{R}^{m \times d}$, where the correspondent coefficient $\mathbf{x} \in \mathbb{R}^d$ is sparse (the major items are zeros). There are many successful applications with sparse representation, such as blind source separation [11], image denoising [12], sparse representation based classification [13] and sparse imputation [14]. Sparse imputation, which is introduced in [14] with application on speech recognition, is a new technique to use sparse representation to recover missing data.

In this paper, sparse imputation is used to evaluate individual feature of training data, and the imputation quality of each feature is recorded to build a new filter model. This filter model is an unsupervised feature selection method, as the class labels don't contribute to sparse imputation. We use proposed feature selection method to applied on UCI data sets, and compared the classification performance with Fisher score method (supervised filter model) and Laplacian score method (unsupervised filter model). Comprehensive comparisons indicate the effectiveness of our method.

The main contributions of this paper are summarized as follows:

- A new filter model Feature Selection via Sparse Imputation (FSSI) is presented. In particular, the imputation quality for individual feature is utilized as evaluation metrics in feature selection.
- The proposed method is applied to UCI [15] data sets (binary-category and multiple-category). The classifica-

tion results are obtained with classic classifiers (support vector machine, k nearest neighbors and multi-layer feed-forward networks).

- The proposed unsupervised feature selection filter model is compared with other methods, Fisher score method (supervised filter model) and Laplacian score method (unsupervised filter model). The comparison results on UCI data sets demonstrate the capability and efficiency of our method.

The rest of paper is organized as follows: Section 2 focuses on related work, in particular, Fisher score method, Laplacian score method and sparse imputation are introduced. Section 3 presents our proposed feature selection based sparse imputation method. Section 4 presents the detail experiments with UCI data sets. Section 5 gives the conclusion of the paper and discusses some future works.

II. RELATED WORK

We consider a data set $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\} \in \mathbb{R}^{n \times m}$, n is the total number and m is the dimension for each data. The labels of the data set are $C = \{c_1, c_2, \dots, c_n\}$. The original feature sets are $\mathbf{F} = \{F_1, F_2, \dots, F_m\}$, and the purpose of feature selection is to identify the feature subsets $\mathbf{F}_s = \{F_1, F_2, \dots, F_s\}$ with $s < m$. Based on above setting, we introduce Fisher score method, Laplacian score method and sparse imputation in the following parts.

A. Fisher score

Fisher score [16] is one of the most popular supervised feature selection method. The key idea of Fisher score is to search for a feature subset, the distances for different classes in the feature subset are as large as possible, while the distances of the same class in the feature subset are as small as possible. Fisher score method has been improved recently. In [17], a clustering method specialized for Fisher score is developed, which is able to detect important dimensions. Generalized Fisher score is proposed in [18], which can optimize the lower bound of traditional Fisher score. The criterion of Fisher score is described in the following part.

Consider a data set $\{\mathbf{y}_i, c_i\}$, $c_i \in \{1, 2, \dots, k\}$ and $i = 1, 2, \dots, n$, where k is the number of the classes. Let n_j denote the number of data in class j and $j = 1, 2, \dots, k$, so $n_1 + n_2 + \dots + n_k = n$. For a feature set F_z in \mathbf{Y} , let μ and σ^2 denote mean and variance, and μ_j and σ_j^2 are the mean and variance for certain categorical data. According to [16], $\sum_{j=1}^k n_j \sigma_j^2$ is the within-class variance, and $\sum_{j=1}^k n_j (\mu_j - \mu)^2$ is the between-class variance. Fisher score S_z for feature F_z is calculated as:

$$S_z = \frac{\sum_{j=1}^k n_j (\mu_j - \mu)^2}{\sum_{j=1}^k n_j \sigma_j^2} \quad (1)$$

The score for individual feature is recorded based on above equation and it can contribute to feature selection.

B. Laplacian score

Laplacian score is proposed based on Laplacian Eigenmaps [19] and Locality Preserving Projection [20], the key idea is to evaluate the features through the features' locality preserving properties. It is a classical unsupervised filter model for feature selection. For the training data \mathbf{Y} with feature set \mathbf{F} , let L_z is the Laplacian score for the z th feature F_z , the Laplacian score is computed as:

1. It first establishes a nearest neighbor graph G with diverse data nodes (\mathbf{y}_α and \mathbf{y}_β , $\alpha, \beta = 1, \dots, n$) in the data set, define $G_{\alpha, \beta} = e^{-\frac{\|\mathbf{y}_\alpha - \mathbf{y}_\beta\|^2}{t}}$, where t is a predefined constant.

2. For each feature F_z , the feature values are $F_z = \mathbf{f}_z$, then L_z can be calculated as

$$L_z = \frac{\tilde{\mathbf{f}}_z^T L \mathbf{f}_z}{\tilde{\mathbf{f}}_z^T Q \mathbf{f}_z} \quad (2)$$

where $Q = \text{diag}(G\mathbf{1})$, $\mathbf{1} = [1, \dots, 1]^T$, $L = Q - G$, and $\tilde{\mathbf{f}}_z$ is a normalization through:

$$\tilde{\mathbf{f}}_z = \mathbf{f}_z - \frac{\mathbf{f}_z^T Q \mathbf{f}_z}{\mathbf{1}^T Q \mathbf{1}} \quad (3)$$

Similar as Fisher score, Laplacian score based on (2) are recorded for feature selection.

C. Sparse imputation

Assume a data set $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_r, \dots, \mathbf{y}_n\} \in \mathbb{R}^{n \times m}$ and a suitable dictionary $\mathbf{D} \in \mathbb{R}^{d \times m}$ for the sparse coding, where $d > m$ ensures the sparsity (number of the nonzero items) of \mathbf{x} . The original constrain is l_0 norm:

$$\min \|\mathbf{x}\|_0, s.t. \quad \mathbf{D}\mathbf{x} = \mathbf{y} \quad (4)$$

However, the solution of above equation is NP hard. With the research of Restricted Isometry Property (RIP) [21], the l_0 norm could be equivalent to l_1 norm. And the sparse coding could be expressed as:

$$\min \|\mathbf{x}\|_1, s.t. \quad \mathbf{D}\mathbf{x} = \mathbf{y} \quad (5)$$

The l_1 norm problem is transferred to a convex optimization problem, and l_1 -regularized least squares method [22] is proposed to calculate the sparse coefficient \mathbf{x} :

$$\hat{\mathbf{x}} = \arg \min \{\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1\} \quad (6)$$

There are also other efficient methods to do the sparse coding, such as matching pursuit [23] and basis pursuit [24]. The dictionary for the sparse coding is important, a lot of recent researches have contributed establishing appropriate dictionary, such as online dictionary learning [25] and Laplacian score dictionary [26]. As the focus of this work is on the sparse representation for imputation, we just use all the training data to build the dictionary, which is similar as sparse representation based classification [13].

Imputation [27] is a statistic method for handling missing data. The sparse coding framework could be transferred to do imputation. In particular, suppose that a data set $\mathbf{Y} \in \mathbb{R}^{n \times m}$

contains $\mathbf{Y}_r \in \mathbb{R}^{n \times q}$ (reliable feature subsets) and $\mathbf{Y}_u \in \mathbb{R}^{n \times (m-q)}$ (unreliable feature subsets):

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_r \\ \mathbf{Y}_u \end{bmatrix}$$

accordingly the dictionary $\mathbf{D} \in \mathbb{R}^{d \times m}$ contains $\mathbf{D}_r \in \mathbb{R}^{d \times q}$ and $\mathbf{D}_u \in \mathbb{R}^{d \times (m-q)}$:

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_r \\ \mathbf{D}_u \end{bmatrix}$$

The sparse coding process is carried on the reliable feature subsets:

$$\mathbf{x} = \arg \min \{ \|\mathbf{y}_r - \mathbf{D}_r \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \} \quad (7)$$

Then use the sparse vector to apply on the unreliable dictionary \mathbf{D}_u to realize imputation, which is $\mathbf{y}_u = \mathbf{D}_u \mathbf{x}$. And sparse imputation can be described as:

$$\hat{\mathbf{y}} = \begin{cases} \hat{\mathbf{y}}_r = \mathbf{y}_r \\ \hat{\mathbf{y}}_u = \mathbf{D}_u \mathbf{x} \end{cases} \quad (8)$$

III. FSSI METHOD

In this section, we introduce the Feature Selection via Sparse Imputation (FSSI) method. The key idea is to use the sparse imputation to recover each feature, then the quality of the representation is the criterion to do the feature selection. The details process is shown in Algorithm 1.

Algorithm 1: Feature Selection based on Sparse Imputation (FSSI)

1: **Input:** a training data set $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots\} \in \mathbb{R}^{n \times m}$, full feature set \mathbf{F} with m features, target feature subsets size s .

2: **for** $z = 1$ to m **do**

3: consider feature F_z in \mathbf{F} to build \mathbf{Y}_r and \mathbf{Y}_u :

$$\mathbf{F}_r = \mathbf{F} - F_z, \mathbf{F}_u = F_z \quad (9)$$

4: based on \mathbf{F}_r and \mathbf{F}_u , obtain \mathbf{Y}_r , \mathbf{Y}_u , \mathbf{D}_r , \mathbf{D}_u

5: **for** $p = 1$ to n **do**

6: obtain sparse vector based \mathbf{D}_r for each data \mathbf{y}'_p in \mathbf{Y}_r

$$\mathbf{x} = \arg \min \{ \|\mathbf{y}'_p - \mathbf{D}_r \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 \} \quad (10)$$

7: conduct sparse imputation: $F(p)^{impu} = \mathbf{D}_u \mathbf{x}$

8: **end for**

9: compute the Euclidean distance between F^{impu} and F_z :

$$score(z) = d(F^{impu}, F_z) = \sqrt{\sum_{p=1}^n (F^{impu}(p) - F_z(p))^2} \quad (11)$$

10: **end for**

11: rank the score to obtain s features

12: Output: \mathbf{F}_s

In this method, we use sparse imputation to evaluate each feature. In details, we first treat a feature set F_z as an unreliable feature sets \mathbf{F}_u . Then we establish reliable feature sets \mathbf{F}_r , which is removed F_z from the whole feature sets \mathbf{F} . The data sets \mathbf{Y}_r and \mathbf{Y}_u are based on feature setting of \mathbf{F}_r and \mathbf{F}_u . In the process of imputation, the dictionary \mathbf{D} are the whole training data [13]. Therefore $\mathbf{Y}_r = \mathbf{D}_r$ and $\mathbf{Y}_u = \mathbf{D}_u$ in our method. Based on sparse coding equation (10), the sparse vector \mathbf{x} for each data \mathbf{y}'_p in \mathbf{Y}_r are calculated (\mathbf{y}'_p is based on the feature set of \mathbf{Y}_r). In step 7, the imputation of unreliable feature for data \mathbf{y}'_p is conducted. Then based on equation (11), sparse imputation quality of the feature set F_k is computed. Finally, step 11 ranks the features based on imputation quality from the worst to the best, and output the feature subsets with target dimensions.

Our FSSI method is proposed based on two perspectives:

- FSSI method shares the prosperities of AdaBoost [28]. In the loop of Adaboost, the misclassified training data would be increased the weights in the next iteration. In our method, the feature with the worst imputation quality would be rank first in feature selection. There are successful feature selection method based Adaboost. In [29], AdaBoost is efficiently used to select the global and local appearance features for face recognition. Adaboost has applied on selecting Gabor Feature for image classification [30], and results are competitive with low memory and computation cost.
- Sparse imputation aims to represent a data (or feature) with existing dictionary. When a data (or feature) has idea sparse imputation quality, it may conclude that the existing dictionary contains the information of the data. Therefore the data is not necessary for the learning system considering memory and computation factors. Whereas a data (or feature) has unacceptable sparse imputation based on existing dictionary, the data should be added in the learning system to improve the diversity.

IV. EXPERIMENT

In this section, the empirical studies are conducted on the nine data sets from UCI Repository [15] to show effectiveness of FSSI method. There are two binary data sets and seven multiple categorical data sets. We focus on multiple classification based on two factors: (1) Compared with binary classification, multiple classification is rare in research. (2) The performance of multiple classification needs improvement in a lot of applications. The details information of the experimental data sets are shown in Table I. Data “credit card” and data “ionsphere” are binary data sets. The rest are multiple categorical data sets. “CMC” is the abbreviation of “Contraceptive Method Choice”, and “image seg” is the abbreviation of “Statlog (Image Segmentation)”.

TABLE I
UCI AND FACE EXPERIMENT DATASETS

Name	Features	Training size	Testing size	Class
credit card	14	345	345	2
ionosphere	34	176	175	2
wine	13	89	89	3
CMC	9	737	736	3
breast tissue	9	53	53	6
wine quality	11	2449	2449	6
glass	10	108	108	7
image seg	19	1165	1165	7
libras	90	180	180	15

TABLE II
FEATURE SELECTION ON DATA WINE

Feature	Fisher rank	Lap rank	SI rank
Alcohol	4	11	5
Malic acid	8	5	6
Ash	11	10	10
Alcalinity	3	4	2
Magnesium	2	12	3
Total phenols	10	9	11
Flavanoids	9	6	8
Nonflavanoid	5	7	13
Proanthocyanins	7	1	4
Color intensity	13	2	7
Hue	6	3	12
OD280/OD315	12	8	9
Proline	1	13	1

A. Configuration

In the experiment, each data set is randomly separated into two equal parts, in which one part are training data and the other part are testing data. The training data are used to establish the model for feature selection. Three filter feature selection methods are utilized in the experiment for comparison: Fisher score (supervised feature selection method), Laplacian score (unsupervised feature selection method) and our proposed feature selection via sparse imputation method. We use Fisher, Lap and FSSI as abbreviations to represent these 3 methods in the experiment.

We use feature based classification as the evaluation criterion [31] to judge different feature selection methods. In particular, a feature selection operator Φ is defined:

- The feature selection operator Φ is trained on the training data based on different algorithms, such as Fisher, Lap and FSSI.
- Update the data \mathbf{Y} based on the operator Φ : $\mathbf{Y}' \leftarrow \Phi(\mathbf{Y})$
- Establish a classifier based the training part of \mathbf{Y}' and record the classification performance on the testing part of \mathbf{Y}'

The experiment on each data set is conducted five times and average results are obtained. The target features size is from one to around 80% of whole feature size to show the comprehensive performances. In order to give more details about feature selection process, Table II shows a case study on data wine. All 13 features are listed in the table. Fisher, Lap and sparse imputation (SI) have been trained on the training data to rank the features respectively. In particular, when the target feature size is 2, Fisher method would select “Proline and Magnesium” features for classification, Lap method would choose “Proanthocyanins and Color intensity” features, and SI method would select “Proline and Alcalinity” features.

Three classic classifiers are used in our experiment: k nearest neighbor ($k = 5$ in the experiment), LibSVM [32] and multi-layer feed-forward networks [33]. We abbreviate above classifiers as 5-NN, LibSVM and NeuralNet. “1-v-r” method [34] is utilized when LibSVM and NeuralNet handle multi-class data sets. The sparse coding toolbox is from [22].

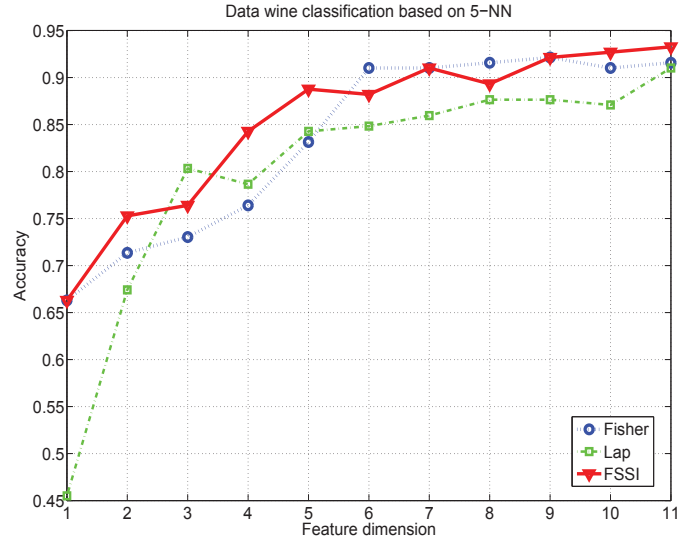


Fig. 1. An example of feature selection based classification on data wine. The feature selection projector is trained with Fisher, Lap and FSSI respectively. Then updated training data and testing data are applied on 5-NN classifier to obtain the classification results on different dimensions. Data wine has 13 features, feature selection has targeted to 11 features, which is 84% of the whole feature sets

Fig. 1 shows an example of classification performance on data wine with different selected features. The outputs of FSSI are more accurate than that of Lap, we may claim that the feature selection method FSSI is more appropriate than Lap for data wine.

B. Comparison among FSSI, Fisher and Lap

In this section, the comparison results among FSSI, Fisher and Lap are shown. For brevity and clarity, we only show the plots of six data sets (two binary data sets and four multiple categorical data sets).

Fig. 2 shows the comparison results for data credit. When the selected feature size is larger than 3, the advantage of FSSI could be observed with all three classifiers. It is also interesting to notice that the results of Fisher and Lap are

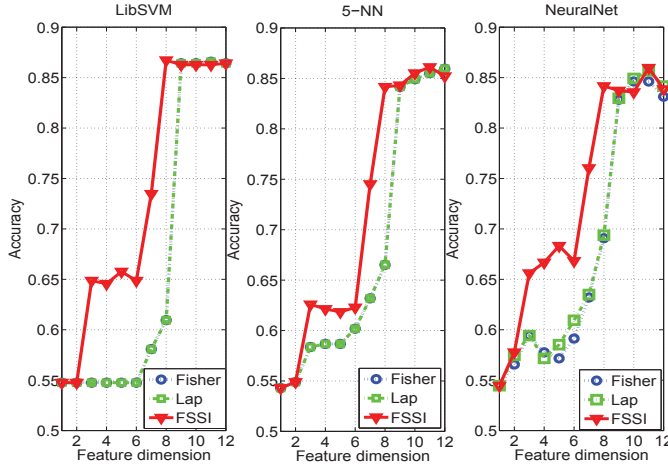


Fig. 2. Comparison of feature selection based classification accuracies for data credit card. Feature selection methods Fisher, Lap and FSSI are used. (Left) Average accuracy of data credit card with LibSVM. (Center) Average accuracy of data credit card with 5-NN. (Right) Average accuracy of data credit card with NeuralNet.

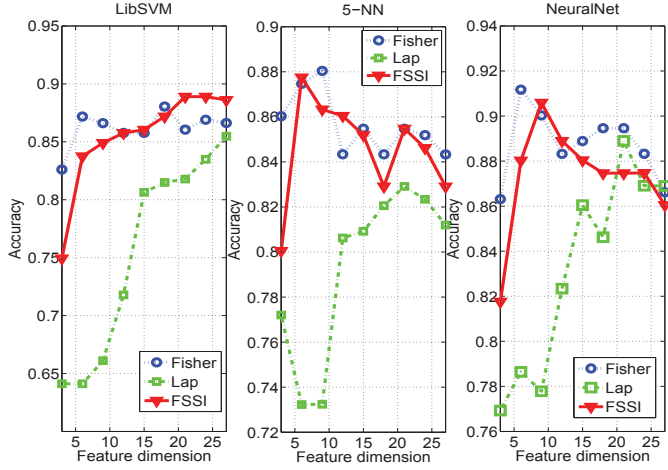


Fig. 3. Comparison of feature selection based classification accuracies for data ionosphere. Feature selection methods Fisher, Lap and FSSI are used. (Left) Average accuracy of data ionosphere with LibSVM. (Center) Average accuracy of data ionosphere with 5-NN. (Right) Average accuracy of data ionosphere with NeuralNet.

almost same in the left two subfigures. However, the results of another binary data ionosphere from Fig. 3 is unremarkable. Through the curves of FSSI are higher than that of Lap, the curves of FSSI are in the same level of Fisher.

When the experiments are carried on the multiple categorical data sets (Fig. 4 to Fig. 7), the effectiveness of FSSI could be noted. In Fig. 4, the results of FSSI are encouraged. In the subfigure of “LibSVM”, the outputs of FSSI dominate the competitors from the dimension of 5. In the subfigures of “5-NN” and “NeuralNet”, the results of FSSI surpass other methods from the dimension of 2. The advantages of FSSI also could be observed in Fig. 5 for data breast tissue, in which the curves of FSSI always appear in the top of subfigures. Fig. 6 shows the experimental results for data glass, Fisher and FSSI methods have improved results compared to Lap method.

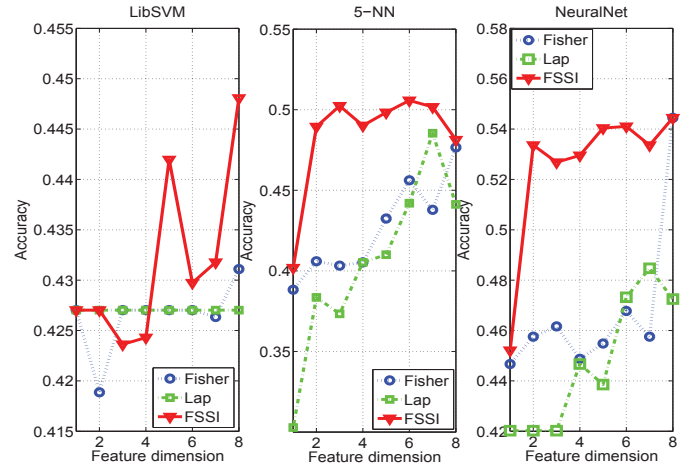


Fig. 4. Comparison of feature selection based classification accuracies for data CMC. Feature selection methods Fisher, Lap and FSSI are used. (Left) Average accuracy of data CMC with LibSVM. (Center) Average accuracy of data CMC with 5-NN. (Right) Average accuracy of data CMC with NeuralNet.

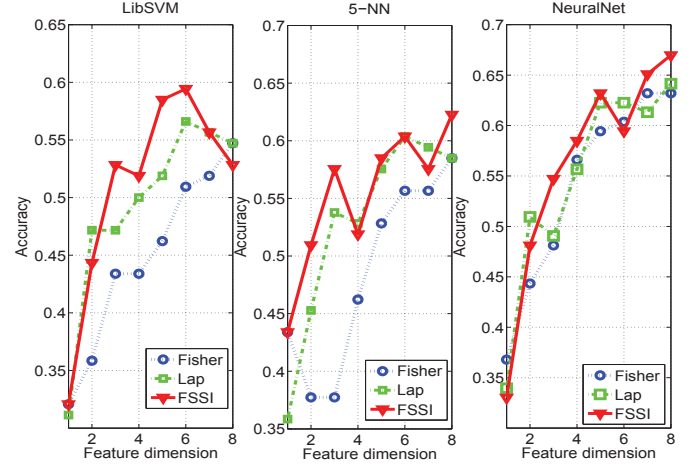


Fig. 5. Comparison of feature selection based classification accuracies for data breast tissue. Feature selection methods Fisher, Lap and FSSI are used. (Left) Average accuracy of data breast tissue with LibSVM. (Center) Average accuracy of data breast tissue with 5-NN. (Right) Average accuracy of data breast tissue with NeuralNet.

And in Fig. 7 for data Libras, Lap and Fisher methods have enhanced outputs compared to Fisher method.

For intensive comparison of different feature selection methods, the statistic analysis is applied with experiment results. For each data set and each method, the classification accuracies, from feature size one to around 50% of whole feature sizes, are averaged and standard deviations are calculated [35] in the Table III. The highest accuracy is highlighted. We can observe that FSSI can win 5, 6 and 6 times in LibSVM, 5-NN and NeuralNet separately. The standard deviation of FSSI is slightly larger than competitors, which may indicate that the features selected by FSSI are contributable.

V. CONCLUSIONS

In this paper, we have explored the use of sparse imputation for feature selection. The proposed FSSI method utilizes the

TABLE III
ACCURACY AVERAGE AND STANDARD DEVIATION IN LOW DIMENSION

Data set	Evaluation	LibSVM			5-NN			NeuralNet		
		Fisher	Lap	FSSI	Fisher	Lap	FSSI	Fisher	Lap	FSSI
credit	Mean	0.552	0.552	0.633	0.583	0.583	0.618	0.582	0.588	0.651
	StDev	0.013	0.013	0.066	0.030	0.030	0.067	0.027	0.029	0.071
ionosphere	Mean	0.856	0.694	0.831	0.863	0.770	0.851	0.889	0.803	0.875
	StDev	0.018	0.070	0.046	0.015	0.038	0.030	0.018	0.038	0.034
wine	Mean	0.760	0.733	0.791	0.789	0.753	0.815	0.795	0.769	0.806
	StDev	0.089	0.207	0.086	0.097	0.146	0.091	0.063	0.174	0.054
CMC	Mean	0.425	0.427	0.429	0.407	0.375	0.476	0.454	0.429	0.516
	StDev	0.004	0	0.008	0.016	0.043	0.042	0.006	0.013	0.036
breast tissue	Mean	0.402	0.455	0.479	0.436	0.491	0.525	0.491	0.504	0.515
	StDev	0.060	0.083	0.102	0.063	0.086	0.061	0.092	0.105	0.117
wine quality	Mean	0.516	0.500	0.524	0.519	0.504	0.528	0.523	0.512	0.533
	StDev	0.042	0.021	0.043	0.057	0.027	0.057	0.047	0.019	0.049
glass	Mean	0.762	0.364	0.738	0.705	0.436	0.709	0.736	0.446	0.743
	StDev	0.037	0.021	0.029	0.007	0.069	0.011	0.015	0.070	0.022
image seg	Mean	0.752	0.494	0.569	0.810	0.573	0.668	0.830	0.597	0.706
	StDev	0.267	0.164	0.211	0.249	0.159	0.216	0.260	0.177	0.230
libras	Mean	0.411	0.530	0.515	0.463	0.594	0.588	0.590	0.784	0.742
	StDev	0.150	0.077	0.111	0.124	0.017	0.068	0.184	0.073	0.109
Mean Wins		3	1	5	2	1	6	2	1	6

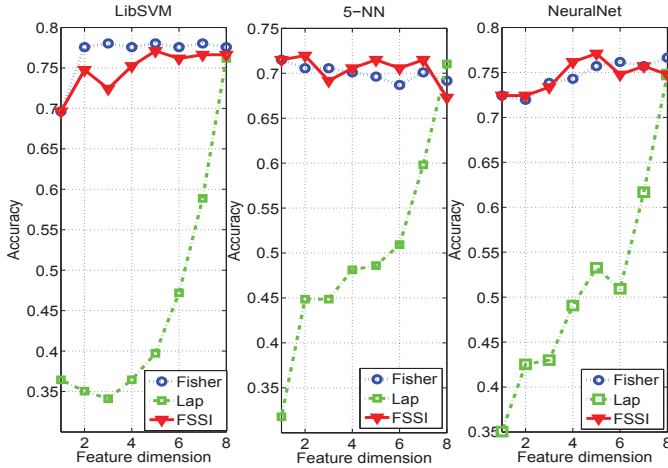


Fig. 6. Comparison of feature selection based classification accuracies for data glass. Feature selection methods Fisher, Lap and FSSI are used. (Left) Average accuracy of data glass with LibSVM. (Center) Average accuracy of data glass with 5-NN. (Right) Average accuracy of data glass with NeuralNet.

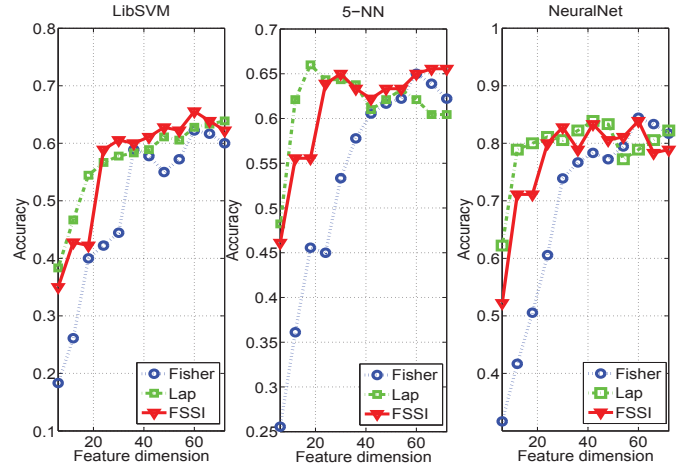


Fig. 7. Comparison of feature selection based classification accuracies for data libras. Feature selection methods Fisher, Lap and FSSI are used. (Left) Average accuracy of data libras with LibSVM. (Center) Average accuracy of data libras with 5-NN. (Right) Average accuracy of data libras with NeuralNet.

sparse imputation quality to rank features. Then we use the feature based classification method to evaluate the feature selection methods. The experiment is conducted in binary and multiple categorical data sets from UCI Machine Learning Repository. The comparison results with Fisher score and Laplacian score demonstrate the effectiveness of FSSI method.

There are several future plans along this topic: (1) The computation cost of sparse imputation should be considered and efficient algorithms are need to speed up the calculation. (2) Robust theoretical analysis is needed to justify the proposed method. (3) High dimension data sets are needed to test capability of FSSI method.

ACKNOWLEDGMENT

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Grant FA8650-11-1-7152 and FA8650-11-1-7148.

REFERENCES

- [1] P. Mitra, C.A. Murthy and S.K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 301-312, 2002.
- [2] J. Mutch and D.G. Lowe, "Multiclass object recognition with sparse, localized features," *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 11-18, 2006.

- [3] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.
- [4] J.B. Yang and C.J. Ong, "Feature Selection using Probabilistic Prediction of Support Vector Regression," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 954-962, 2011.
- [5] M. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," *Proc. 17th Int'l Conf. Machine Learning*, pp. 359-366, 2000.
- [6] L. Wang, "Feature Selection with Kernel Class Separability," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1534-1546, 2008.
- [7] T. Pavlenko, "On feature selection, curse-of-dimensionality and error probability in discriminant analysis," *Journal of Statistical Planning and Inference*, vol. 115, no. 2, pp. 565-584, 2003.
- [8] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research* 3, pp. 1157-1182, 2003.
- [9] X. He, D. Cai and P. Niyogi, "Laplacian score for feature selection," *Proc. Advances in the Neural Information Processing Systems 18*, Vancouver, Canada, 2005.
- [10] F. Wang and P. Li, "Compressed Nonnegative Sparse Coding," *IEEE 10th International Conference on Data Mining (ICDM)*, pp. 1103-1108, 2010.
- [11] Y. Li, S. Amari, A. Cichocki, D.W.C. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Transactions on Signal Processing*, vol.54, no.2, pp. 423-437, 2006.
- [12] M. Elad, and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, 15(12), pp. 3736-3745, 2006.
- [13] J. Wright, A.Y. Yang, A. Ganesh, S.S Sastry and Y. Ma, "Robust face recognition via sparse representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), pp. 210-227, 2009.
- [14] J. Gemmeke and B. Cranen, "Using sparse representations for missing data imputation in noise robust speech recognition," *Proc. of EUSIPCO 2008*.
- [15] A. Frank and A. Asuncion, "UCI Machine Learning Repository," [<http://archive.ics.uci.edu/ml/>], Irvine, CA: University of California, School of Information and Computer Science. 2010.
- [16] R.O. Duda P.E. Hart and D.G. Stork, "Pattern Classification," Wiley-Interscience Publication, 2001.
- [17] K. Tsuda, M. Kawanabe and K.R. Muller, "Clustering with the Fisher Score", *Advances in Neural Information Processing Systems 15*, MIT Press, 2003.
- [18] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," *International Conference on Uncertainty in Artificial Intelligence*, 2011.
- [19] M. Belkin and P. Niyogi, "Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering," *Advances in Neural Information Processing Systems*, vol. 14, 2001.
- [20] X. He and P. Niyogi, "Locality Preserving Projections," *Advances in Neural Information Processing Systems*, vol. 16, 2003.
- [21] E. Candes and T. Tao, "Near optimal signal recovery from random projections and universal encoding strategies," *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406-5225, 2006.
- [22] S. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, "An interior-point method for largescale l1-regularized least squares", *IEEE Journal of Selected Topics in Signal Processing*, 1(4), pp. 606-617, 2007.
- [23] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries", *IEEE Trans. Signal Process.*, vol. 41, pp.3397-3415, 1993.
- [24] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit", *SIAM Journal on Scientific Computing*, 20(1), pp. 33-61, 1998.
- [25] J. Mairal, F. Bach, J. Ponce and G. Sapiro, "Online Dictionary Learning for Sparse Coding", *International Conference on Machine Learning*, 2009.
- [26] J. Xu and H. Man, "Dictionary Learning Based on Laplacian Score in Sparse Coding," *Lecture Notes in Computer Science*, vol. 6871, pp. 253-264, 2011.
- [27] P.D. Allison, "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research*, vol. 28, pp. 301-309. 2000.
- [28] Y. Freund and R.E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences* 55, pp. 119-139, 1997.
- [29] P. Silapachote, D.R. Karupiah and A. Hanson, "Feature selection using adaboost for face expression recognition," *The Fourth IASTED International Conference on Visualization, Imaging, and Image Processing*, Marbella, Spain, pp. 84-89, 2004.
- [30] L. Shen and L. Bai, "AdaBoost Gabor feature selection for classification," *Image and Vision Computing NewZealand (IVCNZ)*, Akaroa, New Zealand, pp. 77-83, 2004.
- [31] Z. Zhao and H. Liu, "Semi-supervised feature selection via spectral analysis," *Proc. 7th SIAM International Conference on Data Mining*, 2007.
- [32] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>*, 2001.
- [33] I.H. Witten and E. Frank, *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco: Morgan Kaufmann, 2000.
- [34] Y. Lee, Y. Lin and G. Wahba, "Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, vol. 99, pp. 67-81, 2004.
- [35] J. Ren, Z. Qiu, W. Fan, H. Cheng and P.S. Yu, "Forward semi-supervised feature selection," *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining (PAKDD)*, Osaka, Japan, 2008.