Feature Selection with Integrated Relevance and Redundancy Optimization

Linli Xu, Qi Zhou, Aiqing Huang, Wenjun Ouyang, Enhong Chen
School of Computer Science and Technology
University of Science and Technology of China, Hefei, Anhui, China
linlixu@ustc.edu.cn, {zhouqixs, cassie89, oy01}@mail.ustc.edu.cn, cheneh@ustc.edu.cn

***Abstract*—The task of feature selection is to select a subset of the original features according to certain predefined criterion with the goal to remove irrelevant and redundant features, improve the prediction performance and reduce the computational costs of data mining algorithms. In this paper, we integrate feature relevance and redundancy explicitly in the feature selection criterion. Spectral feature analysis is applied here which can fit into both supervised and unsupervised learning problems. Specifically, we formulate the problem into a combinatorial problem to maximize the relevance and minimize the redundancy of the selected subset of features at the same time. The problem can be relaxed and solved with an efficient extended power method with global convergence guaranteed. Extensive experiments demonstrate the advantages of the proposed technique in terms of improving the prediction performance and reducing redundancy in data.**

***Keywords*-spectral feature selection; relevance; redundancy; eigen-optimization; supervised/unsupervised learning**

I. Introduction

The problem of handling high-dimensional data is one of the fundamental challenges in data mining. Given a large number of features, one is often confronted with the problems of overfitting and incomprehensible models. Moreover, irrelevant and redundant features may deteriorate the generalization performance of learning algorithms. To address these issues, feature selection has been considered as an effective method to reduce the dimensionality and remove the irrelevant and redundant features [1], [2].

Specifically, feature selection refers to the process of obtaining an optimal subset from the original feature space, according to some predefined criterion. Given the selected features, traditional data mining models can be applied as normal. By discarding the "bad" features from data and reducing the dimensionality, one can benefit from a reduction in the computational overhead, as well as potentially better predictive performance. Comparing to general dimensionality reduction methods, feature selection is preferable when one wants to preserve the original feature representation.

Numerous approaches have been proposed for the feature selection task, which can generally be categorized into supervised or unsupervised depending on whether the label information is available or not in the data. For supervised feature selection, the relevance of a feature is usually evaluated based on its correlation with the class label. Examples include Pearson correlation coefficients [3], Fisher score [4], ReliefF [5], mutual information [3], [6], [7], [8], [9] and trace ratio [10]. In the unsupervised scenario, it is usually harder to select features in the absence of label information. Among the existing methods, exploiting data variance may be the simplest yet effective way to evaluate feature relevance.

In recent years, researchers start to exploit feature selection methods using the spectrum of the graph induced from the pairwise instance similarities. Relevant features are identified by evaluating the capability of features on producing separability of the instances. These methods generally fall into two classes. The first class of methods select features according to some evaluation criteria which is a function of the graph Laplacian. Specifically, the features are ranked according to some scores computed independently. Representative examples include Laplacian score [11], spectral feature selection [12], trace ratio [10], eigenvalue sensitive feature selection [13], etc. An obvious limitation of the ranking based methods is treating features independently without considering possible correlation between features, which implies the possibility of selecting redundant features with high relevance that can adversely affect the predictive performance. To address this problem, the second class of methods formulate the feature selection problem as regression problems, where sparse weights are introduced for all the features indicating whether they are selected or not, and the objective is to minimize the difference of the spectrum of the graph Laplacian with the span of the selected features. To enforce the sparsity of the feature weights, as well as the cardinality of the selected subset of features, it is necessary to introduce various sparsity inducing norms into the formulation. This type of methods treat features as a group and therefore are able to handle feature redundancy. Methods MRSF [14], MCFS [15] and JELSR [16] belong to this category.

In this paper, we follow a different direction and consider the feature selection problem in terms of two essential factors: relevance and redundancy. That is, features with high relevance will be selected while highly correlated features being discouraged. We design a criterion that explicitly incorporates the relevance as well as the redundancy of the selected subset of features at the same time, and formulate the problem of finding this subset as a combinatorial optimization problem to maximize the criterion. We show that the combinatorial problem can be relaxed and efficiently solved with a linearly constrained eigen-optimization technique that is guaranteed to

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.121

converge globally. As a consequence, we achieve a framework of global feature subset optimization that incorporates the factors of both relevance and redundancy.

Specifically, in our framework, we evaluate feature relevance following the spectral feature selection (SPEC) principle which is applicable to both supervised and unsupervised learning tasks. Therefore, our algorithm naturally fits into both supervised and unsupervised scenarios and selects a subset of features that maximizes the relevance while minimizing the redundancy at the same time.

The rest of the paper is organized as follows. We first present our framework of feature selection with integrated relevance/redundancy optimization (FSIR^2) in Section II. After that some related work is discussed in Section III. Next we conduct extensive experiments with feature selection in both classification and clustering tasks to demonstrate the effectiveness of the proposed method in Section IV. The paper is then concluded in Section V.

II. Feature Selection with Integrated Relevance and Redundancy Optimization

To integrate relevance and redundancy at the same time in a mathematical formulation, we need to first define the relevance and redundancy criteria.

Let $X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N) = (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_M)^\top$ be a data set of N instances, and $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_M$ denote the M feature vectors. For supervised learning, class labels $Y = (y_1, y_2, ..., y_N)$ are also given.

A. Spectral Feature Relevance

To measure the relevance of features, one can exploit the graph spectrum induced from the pairwise instance similarities as proposed in [12]. Specifically, similarities of instances can be calculated with an RBF kernel function:

$$S_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}, \tag{1}$$

while in the supervised case, label information can be used and the similarity measure can be defined as:

$$S_{ij} = \begin{cases} \frac{1}{N_l}, & y_i = y_j = l \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

where N_l is the number of data points in class l.

Given the similarity measure, one can construct a graph $\mathbb{G}$ with vertices corresponding to data instances, and the weight between the i-th vertex and the j-th vertex W_{ij} is defined by the similarity between the i-th and j-th instances S_{ij}. That is, the adjacency matrix W associated with the graph $\mathbb{G}$ is equal to the similarity matrix S.

Spectral feature relevance is defined based on the assumption that a feature which separates data better is more relevant to the target, where the target is usually reflected by the structure of $\mathbb{G}$. Therefore, one can evaluate features according to the graph structure, or by analyzing the spectrum of the graph.

Several feature relevance score measures are defined in [12]. In our framework for simplicity we will adopt the following function to evaluate the relevance of the i-th feature:

$$\mathrm{Rel}_i = \hat{\mathbf{f}}_i^{\top} \mathcal{L} \hat{\mathbf{f}}_i = \frac{\mathbf{f}_i^\top L \mathbf{f}_i}{\mathbf{f}_i^\top D \mathbf{f}_i}. \tag{3}$$

In the equation, D is the degree matrix: $D = \mathrm{diag}(W\mathbf{e})$ where $\mathbf{e}$ is a vector of all 1's, L is the Laplacian matrix: $L = D - W$, while $\mathcal{L}$ is the normalized Laplacian matrix: $\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$; and $\hat{\mathbf{f}}_i$ is the normalized weighted feature vector: $\hat{\mathbf{f}}_i = \frac{D^{\frac{1}{2}}\mathbf{f}_i}{\|D^{\frac{1}{2}}\mathbf{f}_i\|}$.

It can be derived that if the eigen-system of $\mathcal{L}$ is $(\lambda_j, \mathbf{v}_j), j = 1, ..., N$, and $\alpha_j = \cos\theta_j$ where θ_j is the angle between $\mathbf{f}_i$ and $\mathbf{v}_j$, the feature relevance (3) can be rewritten as

$$\mathrm{Rel}_i = \sum_{j=1}^{N} \alpha_j^2 \lambda_j, \quad \text{where} \sum_{j=1}^{N} \alpha_j^2 = 1. \tag{4}$$

One should notice that for the i-th feature a small Rel_i value indicates good separability, since $\hat{\mathbf{f}}_i$ aligns closely with the nontrivial eigenvectors corresponding to small eigenvalues.

B. Integrated Relevance/Redundancy Criterion

Next we incorporate redundancy into our framework. To measure the pairwise redundancy between the i-th and j-th features, we simply use the squared cosine distance between the normalized feature vectors $\hat{\mathbf{f}}_i$ and $\hat{\mathbf{f}}_j$. Given that, along with the feature relevance measure discussed above, we can construct an $M \times M$ matrix R to record the relevance and redundancy values for all features:

$$R_{ij} = \begin{cases} \mathrm{Rel}_i, & \text{if } i = j \\ \cos^2(\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j), & \text{otherwise.} \end{cases}$$

Further to make all the entries of the R matrix on the same scale, we observe that Rel_i as shown in (4) corresponds to a weighted sum of $\lambda_1, ...\lambda_N$, while all the weights sum up to 1, which implies a rough estimate of $Average(\mathrm{Rel}_i)$ over all features being $\frac{\sum_i \lambda_i}{N}$. On the other hand, the value of $\cos^2(\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j)$ is upper bounded by 1. Therefore, we can normalize the R matrix by

$$\tilde{R}_{ij} = \begin{cases} \mathrm{Rel}_i, & \text{if } i = j \\ \frac{\sum_i \lambda_i}{N} \cos^2(\hat{\mathbf{f}}_i, \hat{\mathbf{f}}_j), & \text{otherwise.} \end{cases} \tag{5}$$

The constructed $\tilde{R}$ matrix integrates both the relevance and the pairwise redundancy of features. In particular, a small value on the diagonal $\tilde{R}_{ii}$ indicates a highly relevant feature while a small entry off the diagonal $\tilde{R}_{ij}$ implies a pair of diverse features. Intuitively for a good subset of features, all entries in $\tilde{R}$ should be small, implying a small value of $\sum_{ij} \tilde{R}_{ij}$.

Therefore, the problem of selecting a fixed-size subset of features with both high relevance and low redundancy can be formulated as

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^\top \tilde{R} \mathbf{w} \\ \text{s.t.} \quad & \textstyle\sum_i w_i = d \\ & w_i \in \{0, 1\}, \forall i. \end{aligned} \tag{6}$$

where d is the number of features to be selected, which needs to be specified in advance. The binary variable w_i indicates whether the i-th feature will be selected; if so, its corresponding diagonal and off-diagonal elements will be added to the objective.

C. Optimization

Solving the quadratic integer programming problem (6) is NP-hard. However, we will show that if we relax the problem to the real value domain, an approximate solution can be found efficiently.

First using the fact that $\mathbf{z} = 2\mathbf{w} - 1 \in \{-1, +1\}^{M\times 1}$, one can reformulate the problem (6) as

$$\begin{aligned} \min_{\mathbf{z}} \quad & \frac{1}{2}\mathbf{z}^\top \tilde{R}\mathbf{z} + \mathbf{e}^\top \tilde{R}\mathbf{z} \\ \text{s.t.} \quad & \mathbf{z}^\top \mathbf{e} = 2d - M, \\ & \mathbf{z} \in \{-1, +1\}^{M\times 1} \end{aligned} \tag{7}$$

where $\mathbf{e}$ is a vector of all 1's.

Due to the discrete constraints on $\mathbf{z}$, the problem above is non-convex and NP-hard. So we replace the discrete constraints with a relaxed norm constraint $\|\mathbf{z}\| = \sqrt{M}$, and rewrite (7) as

$$\begin{aligned} \min_{\mathbf{z}} \quad & \frac{1}{2}\mathbf{z}^\top \tilde{R}\mathbf{z} + \mathbf{e}^\top \tilde{R}\mathbf{z} \\ \text{s.t.} \quad & \mathbf{z}^\top \mathbf{e} = 2d - M, \\ & \|\mathbf{z}\| = \sqrt{M} \end{aligned} \tag{8}$$

which is still non-convex.

To solve the feature subset optimization problem (8), we design an iterative procedure called *extended power method* to solve the general problem

$$\max_{\mathbf{z}} \frac{1}{2}\mathbf{z}^\top A\mathbf{z} + \mathbf{b}^\top \mathbf{z} \quad \text{s.t.} \quad \|\mathbf{z}\| = r, B\mathbf{z} = \mathbf{c} \tag{9}$$

where A is a semidefinite positive matrix.

Algorithm 1 Extended power method to solve problem (9)

1: $\mathbf{n}_0 = B^\top (BB^\top)^{-1}\mathbf{c}$
2: $\gamma = \sqrt{r^2 - \|\mathbf{n}_0\|^2}$
3: $P = I - B^\top (BB^\top)^{-1}B$
4: $\mathbf{z}_0 = \mathbf{n}_0$, $k = 0$
5: **repeat**
6: $\quad \mathbf{u}_{k+1} = \gamma \frac{P(A\mathbf{z}_k + \mathbf{b})}{\|P(A\mathbf{z}_k + \mathbf{b})\|}$
7: $\quad \mathbf{z}_{k+1} = \mathbf{u}_{k+1} + \mathbf{n}_0$
8: $\quad k = k + 1$
9: **until** $\mathbf{z}$ converges

Output: $\mathbf{z}$

Algorithm 1 works by updating $\mathbf{z}$ along the gradient direction $(A\mathbf{z}+\mathbf{b})$ of the current estimate in each iteration, followed by a projection step into the null space of B (line 6) where P is the projection matrix, and a "pulling" step with $\mathbf{n}_0$ which is the vector from the origin to its projection onto the hyperplane $B\mathbf{z} = \mathbf{c}$ to enforce the constraints. Notice that Algorithm 1 is similar to the projected power method [17] and it solves a more general problem with a linear term $\mathbf{b}^\top \mathbf{z}$ in the objective. When $\mathbf{b} = \mathbf{0}$ Algorithm 1 is equivalent to the projected power method. Global convergence can be derived similarly as in [17].

Proposition 1: Algorithm 1 is guaranteed to converge to the global solution of the optimization problem (9).

The proof can be sketched as follows. In each iteration, for the new solution $\mathbf{z}_{k+1}$ and the current estimate $\mathbf{z}_k$, the inequality holds if the convergence has not been reached yet:

$$\mathbf{z}_{k+1}^\top (\frac{1}{2}A\mathbf{z}_{k+1} + \mathbf{b}) > \mathbf{z}_{k+1}^\top (\frac{1}{2}A\mathbf{z}_k + \mathbf{b}) > \mathbf{z}_k^\top (\frac{1}{2}A\mathbf{z}_k + \mathbf{b})$$

which implies monotonically increasing objective values during updates until convergence. The fixed point at the convergence satisfies $P\mathbf{z} = \mathbf{z} - \mathbf{n}_0 = \gamma \frac{P(A\mathbf{z}+\mathbf{b})}{\|P(A\mathbf{z}+\mathbf{b})\|}$, which corresponds to the conditions of a critical point of the optimization problem (9):

$$(PA - \lambda I)P\mathbf{z} = -P(A\mathbf{n}_0 + \mathbf{b}), \quad \|P\mathbf{z}\| = \gamma \tag{10}$$

for $\lambda = \frac{\|P(A\mathbf{z}+\mathbf{b})\|}{\gamma}$. And the global optimum of (9) is reached at the largest feasible λ for (10).

The overall procedure for feature selection with integrated and redundancy optimization (FSIR^2) is summarized in Algorithm 2. Notice we need to switch the minimization problem (9) to a maximization problem.

Algorithm 2 FSIR^2

Input : Data set X, class labels Y if available, d.

1: Build the similarity matrix S based on X (and Y);
2: Calculate relevance and redundancy for all the features, construct the matrix $\tilde{R}$ according to (5);
3: $A = \eta I - \tilde{R}$ where η is an arbitrary constant such that $A \succeq 0$ without affecting the solutions
4: $\mathbf{b} = -\tilde{R}\mathbf{e}$, $B = \mathbf{e}^\top$, $\mathbf{c} = 2d - M$, $r = \sqrt{M}$
5: Solve the problem (9) according to Algorithm 1

Output: $\mathbf{z}$

The computation of Algorithm 1 mainly lies in the matrix-vector multiplication step, which costs $O(M^2)$ in each iteration. It provides an efficient approach for solving problem (9) considering that the complexity of quadratic programming or eigen-computation is $O(M^3)$ in general.

Once $\mathbf{z}$ is returned, one can easily apply various rounding schemes to it and select features accordingly. This can also be regarded as partitioning the features into a "selected" subset and a "discarded" subset according to their z values with a cardinality constraint, which corresponds to selecting d features with higher values in $\mathbf{z}$.

III. Related Work

Relevance and redundancy are two important factors for the feature selection problem. To address them, various methods have been proposed, which include the minimum Redundancy

Maximum Relevance (mRMR) principle [9], quadratic programming feature selection (QPFS) [18], feature selection with redundancy-constrained class separability [19], etc. However, these methods are all designed for feature selection in supervised learning. Among them, mRMR and QPFS resemble in that they both use mutual information as the similarity measure, while mRMR selects features in a greedy way, and QPFS formulates the problem as a quadratic program. In the experimental investigation, we will compare to mRMR and QPFS in the supervised scenarios.

On the other hand, in this paper we mainly focus on the algorithms that are applicable to both supervised learning and unsupervised learning, and the proposed FSIR^2 model is formulated as a constrained eigen-optimization problem. In our experiments below, we will compare the proposed algorithm to the representative approaches including the ranking based Laplacian score, spectral feature selection (SPEC), eigenvalue sensitive feature selection (EVSC), as well as the regression based multi-cluster feature selection (MCFS) and minimum redundancy spectral feature selection (MRSF).

IV. Experiments

We now empirically evaluate the performance of the algorithm FSIR^2 in both supervised and unsupervised learning.

A. Data sets

Data sets used in our experiments are briefly described in Table I. The first 4 data sets with relatively fewer numbers of features are taken from UCI ML repository [20]. In addition we also include 6 high-dimensional data sets which have been widely used to evaluate spectral feature selection methods. They are 2 image data sets: PIE10P, PIX10P; and 4 Microarray data sets: GLI-85, CLL-SUB-111, SMK-CAN-187 and TOX-171.

TABLE I
SUMMARY OF THE DATA SETS

Data set	Size	Features	Classes
Austra	690	14	2
Clean	476	166	2
Heart	270	13	2
Vote	435	16	2
CLL-SUB-111	111	11340	3
GLI-85	85	22283	2
PIE10P	210	2420	10
PIX10P	100	10000	10
SMK-CAN-187	187	19993	2
TOX-171	171	5748	4

B. Evaluation of selected features

In the experiments, 5 representative spectral feature selection algorithms are chosen as baselines for both unsupervised and supervised investigation: Laplacian Score, SPEC, EVSC, MCFS and MRSF. For Laplacian Score, SPEC, MRSF and FSIR^2, different similarity measures are applied in unsupervised and supervised cases. In unsupervised senarios, similarity measure is calculated using the RBF kernel function defined by (1), and we set $\sigma^2 = \frac{\sum_{i,j}(\mathbf{x_i}-\mathbf{x_j})^2}{N^2}$ in the experiments; while in supervised learning, S is calculated according to (2). Furthermore, in supervised tasks, we also compare to 2 mutual information based redundancy minimization approaches, mRMR and QPFS, which have been developed exclusively for supervised learning.

In supervised setting, algorithms are evaluated with (i) classification accuracy and (ii) squared $\cos\mathrm{ine}$ redundancy rate (RED). Assume F is the set of the d selected features, and X_F only contains features in F, the redundancy rate is defined as:

$$\mathrm{RED}(F) = \frac{1}{d(d-1)} \sum_{\mathbf{f}_i,\mathbf{f}_j \in F, i \neq j} \cos^2(\mathbf{f}_i, \mathbf{f}_j) \ ,$$

The measurement assesses the average similarity among all feature pairs and takes values in $[0, 1]$; a large value indicates potential redundancy in F.

For unsupervised cases, two evaluation measurements are used: (i) clustering accuracy (AC) and (ii) normalized mutual information (NMI).

1) Study of Unsupervised Cases: In this experiment, we investigate the performance of various feature selection algorithms in clustering. We perform k-means clustering (the number of clusters is obtained from the ground truth) by using the selected features and compare the results of different algorithms. On the high-dimensional data sets, 5% of features are selected; while on the UCI data sets, since the numbers of features are generally quite small, we simply set d to the number of features selected by Weka [21][1]. The results are averaged over 100 repeats. Tables II and III present the clustering accuracy and normalized mutual information achieved by different algorithms on the benchmark data sets, where the last column records the results on the data sets without using any feature selection methods. The last row of the tables records the average clustering performance over 10 data sets for each method. In the tables, the bold values indicate the best performance that is statistically significant with 95% confidence. We can observe that the performance of the proposed FSIR^2 method is superior to the rest of the algorithms on most of the data sets in terms of both clustering accuracy and NMI. An interesting thing to observe is that our method uses the same spectral feature relevance as SPEC, and produces better performance on a majority of the data sets, which demonstrates the advantage of integrating feature redundancy into the framework.

We further investigate the influence of the number of selected features on the clustering performance. Figure 1 illustrates the curves of the clustering accuracy and normalized mutual information versus the number of selected features for each algorithm. Due to the space limit, we only plot the results on Vote and PIE10P. The results show a clear advantage of FSIR^2. Especially on PIE10P when the number of features is large, the performance of FSIR^2 demonstrates stable and significant superiority.

[1] weka.attributeSelection.AttributeSelection

TABLE II
STUDY OF UNSUPERVISED CASES: CLUSTERING ACCURACY (THE HIGHER THE BETTER). THE "ALL" COLUMN CORRESPONDS TO LEARNING WITH ALL THE FEATURES.

Data set	$FSIR^2$	SPEC	LScore	EVSC	MCFS	MRSF	All
Austra	0.56	**0.60**	**0.60**	0.56	0.56	0.56	0.56
Clean	**0.50**	**0.50**	0.50	**0.50**	**0.51**	0.50	0.49
Heart	**0.60**	**0.60**	**0.60**	0.53	**0.60**	**0.60**	0.60
Vote	**0.88**	0.82	0.82	0.56	0.75	0.75	0.88
CLL-SUB-111	0.37	0.54	0.54	**0.55**	0.53	**0.55**	0.55
GLI-85	**0.56**	0.53	0.51	0.51	0.54	0.49	0.55
PIE10P	**0.37**	0.18	0.19	0.24	0.24	0.22	0.23
PIX10P	**0.65**	0.59	0.60	0.57	**0.62**	**0.65**	0.67
SMK-CAN-187	**0.58**	0.52	0.52	0.53	0.50	0.55	0.52
TOX-171	**0.41**	0.37	0.38	0.37	0.38	0.38	0.39
AVG	0.55	0.52	0.52	0.49	0.51	0.52	0.54

TABLE III
STUDY OF UNSUPERVISED CASES: NORMALIZED MUTUAL INFORMATION (THE HIGHER THE BETTER). THE "ALL" COLUMN CORRESPONDS TO LEARNING WITH ALL THE FEATURES.

Data set	$FSIR^2$	SPEC	LScore	EVSC	MCFS	MRSF	All
Austra	0.01	**0.02**	**0.02**	0.01	0.01	0.01	0.01
Clean	0.00	**0.03**	**0.03**	0.00	0.00	0.00	0.00
Heart	**0.02**	**0.02**	**0.02**	0.01	**0.02**	**0.02**	0.02
Vote	**0.49**	0.33	0.34	0.05	0.29	0.24	0.49
CLL-SUB-111	0.11	**0.24**	**0.23**	0.19	0.16	0.19	0.19
GLI-85	**0.08**	0.00	0.00	**0.07**	0.04	0.06	0.11
PIE10P	**0.44**	0.12	0.12	0.22	0.32	0.26	0.26
PIX10P	**0.86**	0.83	0.83	0.78	**0.85**	0.84	0.88
SMK-CAN-187	**0.02**	0.00	0.00	0.01	0.00	0.01	0.00
TOX-171	0.10	0.11	0.11	**0.12**	0.10	**0.12**	0.14
AVG	0.25	0.19	0.20	0.17	0.21	0.20	0.21

TABLE IV
STUDY OF SUPERVISED CASES: CLASSIFICATION ACCURACY (THE HIGHER THE BETTER). THE "ALL" COLUMN CORRESPONDS TO LEARNING WITH ALL THE FEATURES.

Data set	$FSIR^2$	SPEC	LScore	MCFS	MRSF	mRMR	QPFS	All
Austra	**0.85**	**0.86**	**0.86**	0.75	0.71	0.73	0.85	0.75
Clean	0.71	0.75	0.75	0.75	**0.76**	0.73	0.70	0.81
Heart	0.81	0.81	0.81	**0.83**	0.81	0.81	0.81	0.83
Vote	**0.96**	**0.96**	**0.96**	**0.95**	**0.95**	**0.96**	**0.96**	0.96
CLL-SUB-111	**0.68**	0.63	0.63	0.65	0.63	0.64	0.63	0.52
GLI-85	**0.88**	**0.88**	**0.88**	**0.90**	0.87	**0.89**	**0.88**	0.90
PIE10P	0.98	0.98	0.98	**0.99**	**0.99**	0.92	0.97	0.99
PIX10P	0.97	0.96	0.96	**0.99**	**0.98**	0.94	0.97	0.98
SMK-CAN-187	**0.70**	**0.70**	0.69	**0.72**	0.69	0.67	**0.71**	0.72
TOX-171	0.84	0.84	0.84	**0.89**	0.83	0.84	0.85	0.88
AVG	0.84	0.82	0.82	0.84	0.82	0.81	0.83	0.83

TABLE V
STUDY OF SUPERVISED CASES: REDUNDANCY RATE (THE LOWER THE BETTER).

Data set	$FSIR^2$	SPEC	LScore	MCFS	MRSF	mRMR	QPFS
Austra	0.48	0.71	0.71	**0.22**	0.30	0.23	0.26
Clean	0.27	0.76	0.76	0.19	0.25	0.19	**0.14**
Heart	0.77	0.82	0.82	0.64	**0.62**	0.66	**0.61**
Vote	**0.21**	0.37	0.37	0.23	0.34	0.30	0.38
CLL-SUB-111	**0.19**	0.90	0.90	0.59	0.60	0.58	0.72
GLI-85	0.61	**0.24**	**0.24**	0.67	0.67	0.71	0.64
PIE10P	**0.33**	0.84	0.84	0.54	0.50	0.66	0.52
PIX10P	0.62	1.00	1.00	**0.61**	0.60	0.78	0.66
SMK-CAN-187	**0.97**	1.00	1.00	0.98	0.97	0.99	0.98
TOX171	**0.56**	0.97	0.97	0.86	0.84	0.86	0.89
AVG	0.50	0.76	0.76	0.55	0.57	0.60	0.58

2) Study of Supervised Cases: In supervised scenarios, for each data set we randomly sample 60% of all the data points as the training data and the rest for test. This process is repeated for 100 times and results are averaged over them. Linear SVM is used for classification with parameters chosen by cross-validation. The classification accuracies are reported in Table IV, where the last column records the results on the data sets without using any feature selection methods. The last row records the average accuracy over 11 data sets. Similarly bold numbers in the table indicate the best performance with statistical significance.

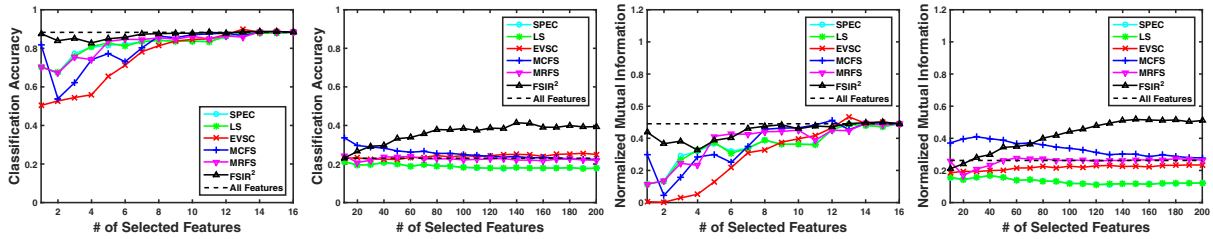
Figure 2(a)-(b) show the classification accuracy versus the number of selected features on Vote and PIE10P respectively for each algorithm. From Figure 2(a)-(b) and Table IV, we can observe that FSIR^2 produces superior classification performance comparing to SPEC, Laplacian Score, MRSF and mRMR, while being comparable to MCFS and QPFS.

To evaluate the effect of reducing redundancy in the selected features of the proposed algorithm, Table V presents the redundancy rates of the feature subsets selected by different algorithms. Figure 2(c)-(d) show the curves of redundancy rates versus the number of selected features on Vote and PIE10P. The results show that FSIR^2 attains low redundancy, which suggests that the redundancy reducing mechanism in our method is effective. In addition, one can notice that among the competing algorithms, the regression based methods including MCFS and MRSF select features with lower average redundancy than the ranking based methods, while being comparable to the two mutual information based redundancy minimization approaches which are developed exclusively for supervised learning, mRMR and QPFS.

From our experimental investigation in both supervised and unsupervised scenarios, it is demonstrated that the proposed FSIR^2 framework can select features containing less redundancy and achieve superior predictive performance.

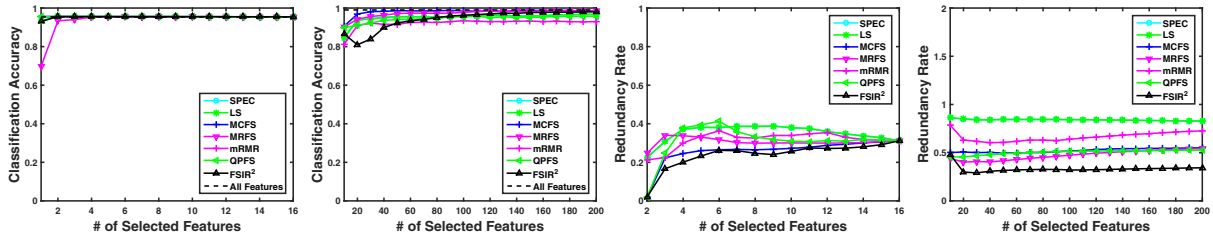
V. CONCLUSION

This paper presents a novel algorithm for feature selection. We design a selection criterion that explicitly integrates feature relevance and redundancy, which are two important factors that affect the quality of selected features. We show that the optimization problem that maximizes the relevance/redundancy criterion can be reformulated and relaxed, after which an efficient extended power method is applied with global convergence guaranteed. Spectral feature analysis is employed here that is applicable to both supervised and unsupervised scenarios. The resulting feature selection procedure (FSIR^2) yields superior predictive performance while reducing redundancy in data.



(a) Accuracy: Vote (b) Accuracy: PIE10P (c) NMI: Vote (d) NMI: PIE10P

Fig. 1. Study of unsupervised cases: (a)-(b) AC vs. # of selected features on "Vote" and "PIE10P" (the higher the better); (c)-(d) NMI vs. # of selected features on "Vote" and "PIE10P" (the higher the better).



(a) Accuracy: Vote (b) Accuracy: PIE10P (c) RED: Vote (d) RED: PIE10P

Fig. 2. Study of supervised cases: (a)-(b) classification accuracy vs. # of selected features on "Vote" and "PIE10P" (the higher the better); (c)-(d) redundancy rate vs. # of selected features on "Vote" and "PIE10P" (the **lower** the better).

ACKNOWLEDGMENT

Research supported by the National Natural Science Foundation of China (No. 61375060) and the National Science Foundation for Distinguished Young Scholars of China (No. 61325010).

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.

[2] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, 1999.

[3] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, pp. 59–66, 1988.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. Wiley-Interscience, 2000.

[5] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of relieff and rrelieff," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[6] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[7] H. Yang and J. Moody, "Feature selection based on joint mutual information," in *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, 1999.

[8] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, 2002.

[9] 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.

[10] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *Proceedings of the 23rd National Conference on Artificial Intelligence*, 2008.

[11] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in Neural Information Processing Systems 18*, 2006.

[12] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[13] Y. Jiang and J. Ren, "Eigenvalue sensitive feature selection," in *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[14] Z. Zhao, L. Wang, and H. Liu, "Efficient spectral feature selection with minimum redundancy," in *Proceedings of the 24th National Conference on Artificial Intelligence*, 2010.

[15] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[16] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2011, pp. 1324–1329.

[17] L. Xu, W. Li, and D. Schuurmans, "Fast normalized cut with linear constraints," in *Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[18] I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C. S. Cruz, "Quadratic programming feature selection," *Journal of Machine Learning Research*, vol. 11, 2010.

[19] L. Zhou, L. Wang, and C. Shen, "Feature selection with redundancy-constrained class separability," *IEEE Transactions on Neural Networks*, vol. 21, no. 5, pp. 853–858, 2010.

[20] A. Frank and A. Asuncion, "Uci machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278