TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/10 pp629-635
Volume 18, Number 6, December 2013

Efficient Leave-One-Out Strategy for Supervised Feature Selection

Dingcheng Feng, Feng Chen*, and Wenli Xu

Abstract: Feature selection is a key task in statistical pattern recognition. Most feature selection algorithms have been proposed based on specific objective functions which are usually intuitively reasonable but can sometimes be far from the more basic objectives of the feature selection. This paper describes how to select features such that the basic objectives, e.g., classification or clustering accuracies, can be optimized in a more direct way. The analysis requires that the contribution of each feature to the evaluation metrics can be quantitatively described by some score function. Motivated by the conditional independence structure in probabilistic distributions, the analysis uses a leave-one-out feature selection algorithm which provides an approximate solution. The leave-one-out algorithm improves the conventional greedy backward elimination algorithm by preserving more interactions among features in the selection process, so that the various feature selection objectives can be optimized in a unified way. Experiments on six real-world datasets with different feature evaluation metrics have shown that this algorithm outperforms popular feature selection algorithms in most situations.

Key words: leave-one-out; feature selection objectives; evaluation metrics

1 Introduction

Feature selection is a fundamental task for many statistical pattern recognition applications such as image processing, speech recognition, text mining, and bioinformatics[1-6]. Feature selection is usually a discrete process using advanced combinatorial mathematics, especially discrete optimization. Due to its importance in applications and the optimization difficulties, feature selection has attracted much attention with the development of pattern recognition and machine learning methods[1,7,8].

Many new feature selection methods have been recently proposed using the formalism of nonlinear dimensionality reduction, especially the manifold learning techniques, such as the Laplacian eigenmap[9] and locally linear embedding[10]. Laplacian score, which is closely related to the canonical Fisher score[11], evaluates the importance of each feature by its ability to preserve the locality relations among samples[12]. SPEC (SPECtrum decomposition of graph Laplacian) extends the Laplacian score by providing a unified perspective to systematically select features based on the properties of graph Laplacians with a series of heuristic strategies for deriving the SPEC framework and illustrations of the connection between Laplacian score and ReliefF[13,14]. The Trace Ratio (TR) criterion generalizes both the Laplacian score and SPEC by evaluating a single feature to evaluating a feature subset with a "trace-ratio" objective function defined with respect to the feature selection matrix[15]. New objectives for feature selection have recently been proposed. For example, Multi-Cluster Feature Selection (MCFS) aims to select features which best preserve the multi-cluster structure within the data[16],

• Dingcheng Feng, Feng Chen, and Wenli Xu are with the National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: fdc08@mails.tsinghua.edu.cn; {chenfeng, xuwl}@mail.tsinghua.edu.cn.

∗ To whom correspondence should be addressed.

Manuscript received: 2012-10-15; revised: 2013-06-05; accepted: 2013-06-07

while the Similarity Preserving Feature Selection (SPFS) framework attempts to select features which best preserve the sample similarity matrix[6].

Previous feature selection algorithms have been based on specific objective functions, while the "ultimate" feature evaluation metrics (e.g., classification or clustering accuracies) usually differ from these specific objectives. This study optimizes the "ultimate" metrics (which is a more fundamental objective). Specifically, this method selects feature subsets such that the fundamental objectives are optimized. Further, the model requires that the feature selection algorithm returns a quantitative measure of the contribution of each feature to the fundamental objectives by some score function. A leave-one-out feature selection algorithm is used to provide an approximate solution for these purposes. The leave-one-out algorithm is motivated by the conditional independence structure in probabilistic distributions, and can be viewed as an effort to overcome the limitation of the conventional "greedy backward elimination" method. The algorithm is compared with many popular feature selection algorithms for real-world datasets with different evaluation metrics with the results showing that the leave-one-out algorithm outperforms previous methods in most situations.

2 Feature Selection Objectives

A common "framework" for most feature selection algorithms can be summarized as the following two closely related procedures: (1) Formulate the feature selection to optimize some objective function which represents some reasonable intuitive idea explicitly. (2) Utilize an optimization strategy which leads to meaningful solutions of the designed objectives. The objective functions should relate to the feature selection motivation. However, the feature selection formulation can differ greatly, which can lead to different optimization objectives. The following illustrates the differences by describing several of the most popular feature selection objectives.

• SPEC[14] (The Laplacian score[12] is used as special case). The objective is to select a feature subset to form the best lower dimensional representation by

$$\min_{\boldsymbol{S}} \mathcal{J}(\boldsymbol{S}) = \mathrm{tr}(\boldsymbol{Y}\mathcal{L}\boldsymbol{Y}^{\mathrm{T}}) \tag{1}$$

where $\mathcal{L} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$ is the normalized Laplacian matrix, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ is the Laplacian matrix, $W_{ij} = \dfrac{1}{n_l}$ if $y_i = y_j = l$ and 0 elsewhere (supervised case) or $W_{ij} = \exp\left(-\dfrac{||\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}||^2}{2\sigma^2}\right)$ (unsupervised case) is the pairwise instance similarity matrix, $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{W}$ defined by a diagonal matrix with $D_{ii} = \sum_{j=1}^{m} W_{ij}$ $(i = 1, \cdots, n)$, $\boldsymbol{Y} = [y_j^{(i)}]$ $(i = 1, \cdots, m,\ j = 1, \cdots, k)$ is the submatrix of the data matrix defined by $\boldsymbol{y}^{(i)} = \boldsymbol{S}\boldsymbol{x}^{(i)}$ (representing data points with respect to selected variables), and $\boldsymbol{S} \in \{0,1\}^{k\times n}$ $(\boldsymbol{S}\boldsymbol{1}_{n\times 1} = \boldsymbol{1}_{k\times 1},\ \left\|\boldsymbol{1}_{k\times 1}^{\mathrm{T}}\boldsymbol{S}\right\|_0 = k)$ is the feature selection matrix.

• TR criterion[15]. The objective is to select a feature subset to simultaneously minimize the within-class affinity and maximize the between-class affinity by

$$\max_{\boldsymbol{S}} \mathcal{J}(\boldsymbol{S}) = \frac{\mathrm{tr}(\boldsymbol{S}\boldsymbol{X}\boldsymbol{L}_{\mathrm{b}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})}{\mathrm{tr}(\boldsymbol{S}\boldsymbol{X}\boldsymbol{L}_{\mathrm{w}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}})} \tag{2}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ is the data matrix with m samples and n features, $\boldsymbol{S} \in \{0,1\}^{k\times n}$ $(\boldsymbol{S}\boldsymbol{1}_{n\times 1} = \boldsymbol{1}_{k\times 1},\ \left\|\boldsymbol{1}_{k\times 1}^{\mathrm{T}}\boldsymbol{S}\right\|_0 = k)$ is a feature selection matrix, $\boldsymbol{L}_{\mathrm{w}} = \boldsymbol{D}_{\mathrm{w}} - \boldsymbol{A}_{\mathrm{w}}$, $\boldsymbol{L}_{\mathrm{b}} = \boldsymbol{D}_{\mathrm{b}} - \boldsymbol{A}_{\mathrm{b}}$, $\boldsymbol{A}_{\mathrm{w}}$ and $\boldsymbol{A}_{\mathrm{b}}$ represent the within-class (or local) and between-class (or global) affinity relationships among samples, $\boldsymbol{D}_{\mathrm{w}}$ is the degree matrix of $\boldsymbol{A}_{\mathrm{w}}$ defined by a diagonal matrix with $(\boldsymbol{D}_{\mathrm{w}})_{ii} = \sum_{j=1}^{m}(\boldsymbol{A}_{\mathrm{w}})_{ij}$ $(i = 1, \cdots, n)$, and $\boldsymbol{D}_{\mathrm{b}}$ is the degree matrix of $\boldsymbol{A}_{\mathrm{b}}$ defined by a diagonal matrix with $(\boldsymbol{D}_{\mathrm{b}})_{ii} = \sum_{j=1}^{m}(\boldsymbol{A}_{\mathrm{b}})_{ij}$ $(i = 1, \cdots, n)$.

• SPFS[6]. The objective is to select a feature subset which best preserves the similarity among samples by

$$\min_{\boldsymbol{S}} \mathcal{J}(\boldsymbol{S}) = \left\|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{S}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{X} - \boldsymbol{W}\right\| \tag{3}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ is the data matrix with m samples and n features, $\boldsymbol{S} \in \{0,1\}^{k\times n}$ $(\boldsymbol{S}\boldsymbol{1}_{n\times 1} = \boldsymbol{1}_{k\times 1},\ \left\|\boldsymbol{1}_{k\times 1}^{\mathrm{T}}\boldsymbol{S}\right\|_0 = k)$ is a feature selection matrix, and $\boldsymbol{W}$ is a predefined similarity matrix which can be constructed either by using label information (supervised case) or using distance matrices (unsupervised case).

• MCFS[16]. The objective is to select a feature subset which best preserves the multi-cluster structure among samples by

$$\max_{\boldsymbol{S}} \mathcal{J}(\boldsymbol{S}) = \sum_{j\in F(\boldsymbol{S})} \mathrm{MCFS}(j) \tag{4}$$

where $F(\boldsymbol{S})$ is the selected feature subset, $\mathrm{MCFS}(j)$ is the score for feature j defined by $\mathrm{MCFS}(j) = \max_k \left|a_{k,j}\right|$, $\boldsymbol{a}_k \in \mathbb{R}^{n\times 1}$ is the optimal solution to the L1-regularized regression problem

$$\min_{\boldsymbol{a}_k} \|\boldsymbol{y}^{(k)} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{a}_k\| + \|\boldsymbol{a}_k\|_1 \tag{5}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ is the data matrix with m samples and n features, $\boldsymbol{y}^{(k)} \in \mathbb{R}^{m\times 1}$ is the solution to the generalized eigenvalue decomposition problem $\boldsymbol{L}\boldsymbol{y} = \lambda \boldsymbol{D}\boldsymbol{y}$ ($\boldsymbol{y}^{(k)}$ corresponds to the k-th smallest eigenvalue), $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ is the Laplacian matrix, $\boldsymbol{W}$ is the weight matrix defined on a graph (can have multiple choices), and $\boldsymbol{D}$ is the degree matrix of $\boldsymbol{W}$ defined by a diagonal matrix with $D_{ii} = \sum_{j=1}^{m} W_{ij}$ ($i = 1, \cdots, n$).

SPEC, TR, SPFS, and MCFS are all appropriate feature selection formulations, which select feature subsets based on reasonable intuition. However, the direct connections among these different motivations are difficult to identify, and users have difficulty choosing which algorithm to apply in specific applications. Most feature selection evaluation criteria are the classification or clustering accuracies, which are the "ultimate" objective for the feature selection. The next section describes a more direct perspective to optimizing the more fundamental objectives of the feature selection.

3 Leave-One-Out Feature Selection Framework

Three popular evaluation metrics for feature selection are used here to evaluate a dataset $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(m)}] \in \mathbb{R}^{n\times m}$ with the corresponding labels $\boldsymbol{y} = [y^{(1)}, \cdots, y^{(m)}] \in \mathbb{R}^{1\times m}$ where $y^{(i)} \in \{1, \cdots, C\}$ is the label of sample point $\boldsymbol{x}^{(i)}$. Specifically, the selected feature subset $F(\boldsymbol{S}) \in \{1, \cdots, n\}$ is evaluated using the following evaluation metric functions on $\boldsymbol{X}(F, :) = [\boldsymbol{x}_F^{(1)}, \cdots, \boldsymbol{x}_F^{(m)}] \in \mathbb{R}^{k\times m}$ (denoted by $\boldsymbol{X}_F$) and $\boldsymbol{y}$, which can be computed and $\boldsymbol{x}_F^{(i)} = \boldsymbol{x}^{(i)}(F)$ denotes the elements indexed by F in $\boldsymbol{x}^{(i)}$.

• Classification accuracy[6, 12, 14-16]:

$$\max \mathcal{J}_{\text{classify}} = \frac{\sum_{i=1}^{m_{\text{test}}} \delta(y^{(i)}, \hat{f}(\boldsymbol{x}_F^{(i)}))}{m_{\text{test}}} \tag{6}$$

where $\hat{f}$ is a classifier trained with m_{train} data points randomly sampled from $(\boldsymbol{X}_F, \boldsymbol{Y})$, m_{test} is the number of remaining samples (test samples), and $\delta(x, y)$ is the delta function which equals 1 if $x = y$ and 0 otherwise.

• Clustering accuracy[12, 16]:

$$\max \mathcal{J}_{\text{cluster}} = \frac{\sum_{i=1}^{m} \delta(y^{(i)}, \text{map}(c^{(i)}))}{m} \tag{7}$$

where $y^{(i)}$ is the provided label, $c^{(i)}$ is the cluster label of data point $\boldsymbol{x}^{(i)}$, and $\text{map}(c^{(i)})$ maps each cluster label $c^{(i)}$ to the equivalent label in data.

• Normalized mutual information[12, 16]:

$$\max \mathcal{J}_{mi} = \frac{\text{MI}(C, C')}{\max(H(C), H(C'))} \tag{8}$$

where

$$\text{MI}(C, C') = \sum_{c_i \in C, c_j' \in C'} p(c_i, c_j') \log_2 \frac{p(c_i, c_j')}{p(c_i)\, p(c_j')}$$

and $p(c_i)$ is the probability of a sample proportion in cluster c_i, and $H(C)$ is the entropy of cluster C.

The feature selection objectives (Eqs. (6)-(8)) are more fundamental than the case in Eqs. (1)-(4). Previous studies have proposed some "intuitive" feature selection criteria (e.g., Eqs. (1)-(4)), and evaluate the selected feature subset by the "ultimate" evaluation criteria (e.g., Eqs. (6)-(8)). Though these two classes of objectives are closely connected as illustrated by the effectiveness of previous feature selection algorithms, good solutions of the intuitive criteria cannot guarantee good solutions of the ultimate in generic situations.

This study aims to provide a unified framework to find good feature subsets with respect to arbitrary evaluation objectives. Specifically, the model is based on

$$\max_F \mathcal{J}(F) \tag{9}$$

where $\mathcal{J}$ can be arbitrary feature selection objectives. We are interested in cases when $\mathcal{J}$ is Eqs. (6)-(8) which are more fundamental and direct feature selection criteria. Moreover, the model will give individual representation of the importance of each feature, i.e., give each feature a "score" which describes its contribution to the objective. Though the "fundamental objective" and "individual representation" are desirable, Eq. (9) cannot be solved in the general case.

Thus, the structure of conditional independence relations[17] is used to provide reasonable approximations for Eq. (9). The Markov blanket of a random variable y with respect to $\{y, \boldsymbol{x}\}$ is a variable subset $\text{MBL}(y) \subset \boldsymbol{x} = \{x_1, \cdots, x_n\}$ when y is independent of all the variables in $\boldsymbol{x}$. The smallest Markov blanket is called the Markov boundary of y, denoted by $\text{MB}(y)$, which is usually unique with one sufficient condition that $p(y, \boldsymbol{x})$ is strictly positive[17]. One property of a Markov blanket is that the intersection of two Markov blankets is still a Markov blanket. Note that the feature subset selection can also be viewed as identifying the Markov boundary of class label y from the feature variables $\boldsymbol{x}$ such that $p(y|\boldsymbol{x}) = p(y|\text{MB}(y))$. Given $\text{MB}(y)$, where y is

independent of all $\boldsymbol{x}\backslash\mathrm{MB}(y)$, then $\mathrm{MB}(y)$ contains the information for the conditional distribution of y and, therefore, is a good feature subset. According to the property of Markov blankets, y is independent of all $\boldsymbol{x}\backslash A$ provided that A is a variable subset satisfying $\mathrm{MB}(y) \subset A \subset \boldsymbol{x}$. However, when $\mathrm{MB}(y) \not\subset A$, i.e., $\exists x_i$ such that $x_i \in \mathrm{MB}(y)$ but $x_i \notin A$, y is not independent of $\boldsymbol{x}\backslash A$ given A and, therefore, A cannot be viewed as a good feature subset.

The conditional independence suggests the intuition that removing an important feature may have a bigger impact on the conditional distribution of $p(y|\boldsymbol{x})$ than adding a redundant feature. Note that $\boldsymbol{x}$ is also a (trivial) Markov blanket of y. Therefore, the importance of feature x_i can be reflected in $D(p(y|\boldsymbol{x}), p(y|\boldsymbol{x}_{-i}))$ where $\boldsymbol{x}_{-i} = \boldsymbol{x}\backslash\{x_i\}$ and D is some distance metric of two distributions. Since general supervised learning is based on detecting the properties of the full conditional distribution $p(y|\boldsymbol{x})$, the individual features can reasonably be evaluated by their contributions to $p(y|\boldsymbol{x})$. Specifically, the score for feature x_i can be given as $w(x_i) = \mathcal{J}(\boldsymbol{x}) - \mathcal{J}(\boldsymbol{x}_{-i})$, where $\mathcal{J}$ is an arbitrary evaluation criterion of a feature subset such as in Eqs. (6)-(8).

Algorithm 1 summarizes this feature selection idea which can use the previous evaluation metrics $\mathcal{J}$ in Eqs. (6)-(8). The algorithm returns the quantitative importance for each feature to the metric. If a larger (smaller) value of $\mathcal{J}$ corresponds to a good feature subset, then a larger (smaller) weight, $w(x_i)$, will correspond to a better (worse) feature. Algorithm 1 runs the first step of the "greedy backward elimination", but stops at this step and then performs similar computations for the other features. Therefore, Algorithm 1 avoids the drawback of the "greedy backward elimination" which usually converges to a local optimal feature subset, since the effect of the interactions among features on the evaluation metric is neglected in the selection process. Algorithm 1 overcomes this limitation by evaluating the $n-1$ features together, which preserves possible interactions among features.

Algorithm 1 The Leave-one-out Feature Selection Algorithm

Input: Feature matrix $\boldsymbol{X} = [\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(m)}] \in \mathbb{R}^{n\times m}$, label vector $\boldsymbol{y} = [y^{(1)}, \cdots, y^{(m)}] \in \mathbb{R}^{1\times m}$, a feature evaluation metric $\mathcal{J}$ such as in Eqs. (6), (7), and (8).

Output: A weight function for each feature.

1: Compute $\mathcal{J}(\boldsymbol{X}, \boldsymbol{y})$.
2: **for** $i \in \{1, \cdots, n\}$ **do**
3: $F_{-i} = \{1, \cdots, n\} \setminus \{i\}$.
4: Compute $\mathcal{J}(\boldsymbol{X}(F_{-i}, :), \boldsymbol{y})$.
5: Compute $w(i) = \mathcal{J}(\boldsymbol{X}, \boldsymbol{y}) - \mathcal{J}(\boldsymbol{X}(F_{-i}, :), \boldsymbol{y})$.
6: **end for**
7: **return** w.

4 Experiments

This section shows that the leave-one-out algorithm can outperform many popular feature selection algorithms with real-world datasets (Table 1). Figure 1 compares the different feature selection algorithms in terms of the classification accuracies, clustering accuracies, and normalized mutual information for real-world datasets. "MCFSSup" and "MCFSUns" denote the supervised and unsupervised version of the MSFS algorithm, "infogain" (information gain) and "relief"[13] denote two conventional feature selection algorithms, "fromESupKNN", "fromESupMI", and "fromESupAC" denote the leave-one-out feature selection algorithm where the evaluation metric $\mathcal{J}$ is the classification accuracy in Eq. (6) indicated as "supKNN" (using the k-nearest neighbor classifier with $k = 1$), the clustering accuracy in Eq. (7) indicated as "supAC" (using the best result of the K-means clustering method for 10 times), and the normalized mutual information in Eq. (8) indicated as "supMI". The classification accuracy, Eq. (6), in "fromESupKNN" was computed from test data where the datasets were partitioned as $m_{\mathrm{train}} \approx 0.5\,\mathrm{m}$ and $m_{\mathrm{test}} \approx 0.5\,\mathrm{m}$. The clustering accuracy, Eq. (7), and the normalized mutual information, Eq. (8), in "fromESupMI" and "fromESupAC" were computed using all the data. The unsupervised "redundancy rate"[18] was computed as an evaluation metric and was defined as $\mathcal{J}_{\mathrm{red}} = \dfrac{1}{F(S)} \sum_{i,j\in F(S)} \rho_{i,j}$ where $\rho_{i,j}$ is the correlation between the i-th and the j-th features and $F(S)$ is the selected feature subset. Note

Table 1 Real-world datasets.

Dataset	Dimension	Size	Classes
dermatology	34	366	6
USPS	256	9298	10
MNIST	784	4000	10
mfeat	649	2000	10
COIL20	1024	1440	20
COIL100	1024	7200	100

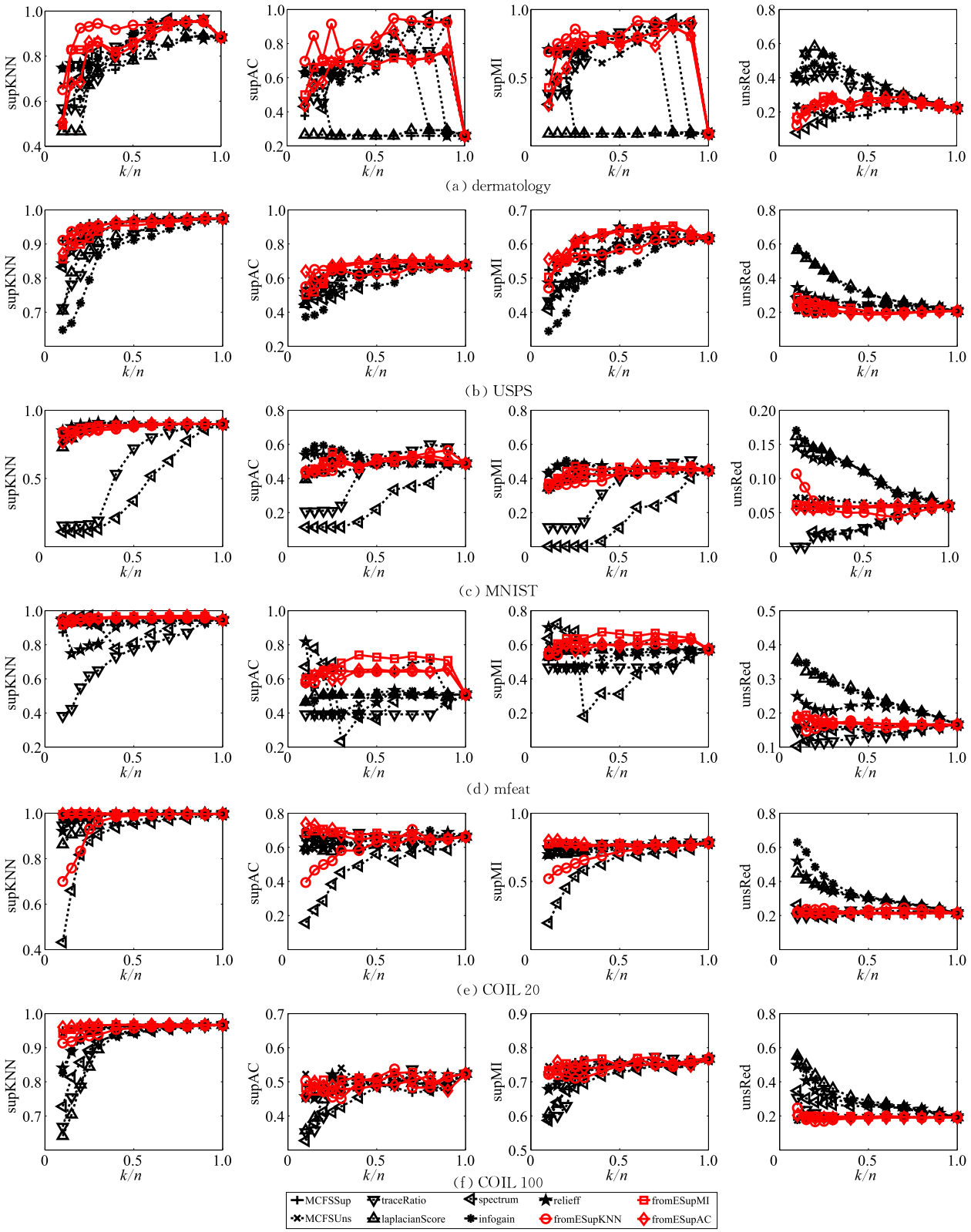


Fig. 1 Classification accuracies (column 1), clustering accuracies (column 2), mutual information (column 3), and redundancy rate (column 4) with different numbers ($\frac{k}{n}$ = 0.1, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1) of selected features in six real-world datasets (row 1-row 6).

that smaller values of $\mathcal{J}_{\text{red}}$ indicate that the redundancy is smaller in $F(S)$ and $F(S)$ is thus a better feature subset.

Figure 1 shows that the leave-one-out feature selection has better performance in most situations, where the classification accuracy, Eq. (6), is computed from the test data with the datasets partitioned as $m_{\text{train}} \approx 0.6\,\text{m}$ and $m_{\text{test}} \approx 0.4\,\text{m}$ and the clustering accuracy, Eq. (7), the normalized mutual information, Eq. (8) and the redundancy rate are computed using all the data. The leave-one-out algorithm selects features which *directly* maximize these classification accuracies, clustering accuracies, and normalized mutual information while preserves the interactions among features in the selection process. Other feature selection algorithms use specific objective functions such as Eqs. (1)-(4) which are reasonable in some situations, but only indirectly solve Eq. (9). The gap between Eq. (9) and the specific feature selection objectives such as in Eqs. (1)-(4) can be quite large, which is why these algorithms are inefficient in some situations.

5 Conclusions

In the past several decades, feature selection has drawn much attention due to its fundamental importance in pattern recognition and the difficulty in optimizing the process. Different from previous algorithms which are built on some specific objective functions, this paper presents a leave-one-out feature selection algorithm that directly optimizes more basic feature selection objectives such as the classification and clustering accuracies. The leave-one-out algorithm borrows the concept from conditional independence structures that removing important features usually causes larger changes in the conditional distribution than removing redundant features, since the label variable is determined by a small set of variables. With this observation, the leave-one-out algorithm defines the score of each feature as the performance change with respect to the absence of the feature from the full feature set. The importance to the objectives is then reflected in the scores, which overcomes the limitation of the greedy backward elimination algorithm that loses interactions among features in the selection process. Experiments on real-world datasets with various kinds of feature evaluation metrics verify the effectiveness of the leave-one-out algorithm.

Acknowledgements

We thank the editors and reviewers for their constructive suggestions and comments. This work was supported by the National Natural Science Foundation of China (Nos. 61071131 and 61271388), the Beijing Natural Science Foundation (No. 4122040), the Research Project of Tsinghua University (No. 2012Z01011), and the Doctoral Fund of the Ministry of Education of China (No. 20120002110036).

References

[1] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.*, vol. 3, nos. 7-8, pp. 1157-1182, Oct. 2003.

[2] H. Peng, F. Long, and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[3] M. C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[4] J. Li and M. Sun, Non-independent term selection for Chinese text categorization, *Tsinghua Science and Technology*, vol.14, no.1, pp. 113-120, Feb. 2009.

[5] F. Nie, H. Huang, X. Cai, and C. Ding, Efficient and robust feature selection via joint l2, 1-norms minimization, in *Advances in Neural Information Processing Systems 23*, Vancouver, BC, Canada, 2010, pp. 1813-1821.

[6] Z. Zhao, L. Wang, H. Liu, and J. Ye, On similarity preserving feature selection, *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 619-632, Mar. 2013.

[7] C. Hou, F. Nie, D. Yi, and Y. Wu, Feature selection via joint embedding learning and sparse regression, in *Proc. 22nd Int. Joint Conf. on Artificial Intelligence*, Barcelona, Spain, 2011, pp. 1324-1329.

[8] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, Discriminative least squares regression for multiclass classification and feature selection, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738-1754, Nov. 2012.

[9] M. Belkin and M. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Comput.*, vo. 15, no. 6, pp. 1373-1396, Jun 2003.

[10] S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, vol. 290, no. 5500, pp. 2323-2326, Dec. 2000.

[11] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley-Interscience, 2001.

[12] X. He, D. Cai, and P. Niyogi, Laplacian score for feature selection, in *Advances in Neural Information Processing Systems 18*, Vancouver, BC, Canada, 2006, pp. 507-514.

[13] I. Kononenko, Estimating attributes: Analysis and extensions of relief, in *Machine Learning: ECML-94*, Springer, 1994, pp. 171-182.

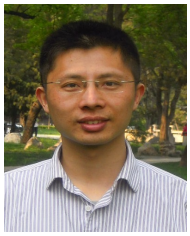
[14] Z. Zhao and H. Liu, Spectral feature selection for supervised and unsupervised learning, in *Proc. 24th Int. Conf. on Machine Learning*, Corvallis, USA, 2007, pp. 1151-1157.

[15] F. Nie, S. Xiang, Y. Jia, C. Zhang, and S. Yan, Trace ratio criterion for feature selection, in *Proc. 23rd AAAI Conf. on Artificial Intelligence*, Chicago, USA, 2008, pp. 671-676.

[16] D. Cai, C. Zhang, and X. He, Unsupervised feature selection for multi-cluster data, in *Proc. 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Washington DC, USA, 2010, pp. 333-342.

[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[18] Z. Zhao, L. Wang, and H. Liu, Efficient spectral feature selection with minimum redundancy, in *Proc. 24th AAAI Conf. on Artificial Intelligence*, Atlanta, USA, 2010, pp. 673-678



Dingcheng Feng received his BS degree in automation from Harbin Institute of Technology, Harbin, China, in 2008. He is currently working toward his PhD degree in control science and engineering in the Department of Automation, Tsinghua University, Beijing, China. His research interests include structure learning, variational inference, causal inference, and feature selection.

Feng Chen received his BS and MS degrees in automation from Saint-Petersburg Poly-technic University, Saint-Petersburg, Russia, in 1994 and 1996, respectively, and his PhD degree in automation from Tsinghua University, Beijing, China, in 2000. He is currently a professor with the National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University. His research interests include computer vision, video processing, variational inference, and structure learning in graphical models.

Wenli Xu received his BS degree in electrical engineering and his MS degree in automatic control engineering from Tsinghua University, Beijing, China, in 1970 and 1980, respectively, and his PhD degree in electrical and computer engineering from the University of Colorado at Boulder, in 1990. He is currently a professor with the National Laboratory for Information Science and Technology, Department of Automation, Tsinghua University. His research interests include automatic control, intelligent robot, and computer vision.