Unsupervised Feature Selection via Unified Trace Ratio Formulation and K-means Clustering (TRACK)

De Wang, Feiping Nie, and Heng Huang*

Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
{wangdelp,feipingnie}@gmail.com, heng@uta.edu

Abstract. Feature selection plays a crucial role in scientific research and practical applications. In the real world applications, labeling data is time and labor consuming. Thus, unsupervised feature selection methods are desired for many practical applications. Linear discriminant analysis (LDA) with trace ratio criterion is a supervised dimensionality reduction method that has shown good performance to improve classifications. In this paper, we first propose a unified objective to seamlessly accommodate trace ratio formulation and K-means clustering procedure, such that the trace ratio criterion is extended to unsupervised model. After that, we propose a novel unsupervised feature selection method by integrating unsupervised trace ratio formulation and structured sparsity-inducing norms regularization. The proposed method can harness the discriminant power of trace ratio criterion, thus it tends to select discriminative features. Meanwhile, we also provide two important theorems to guarantee the unsupervised feature selection process. Empirical results on four benchmark data sets show that the proposed method outperforms other sate-of-the-art unsupervised feature selection algorithms in all three clustering evaluation metrics.

1 Introduction

Feature selection is to select relevant and informative features from the high-dimensional feature space. Because it can improve the mode generalization capability, prevent model over-fitting, identify useful features, and hugely reduce the computational time, feature selection has been playing a crucial role in many scientific and practical applications, such as text mining [7], bioinformatics [5,23,3], medical image analysis [22,24], computer vision [4,12], *etc*.

There are three types of feature selection methods: filter method [20,13,19,5], wrapper method [11], and embedded method [26]. The filter methods compute a score to each feature, so the computational cost is relatively low, but the selected features often cannot achieve good classification performance. Wrapper methods treat the classifier as a black box, and use classification results to evaluate potential feature subset, thus the features selected by wrapper methods usually have good performance. However, their computational cost is very high since it need to use the classifier all the way through the

* This work was partially supported by NSF IIS-1117965, IIS-1302675, IIS-1344152, DBI-1356628.

T. Calders et al. (Eds.): ECML PKDD 2014, Part III, LNCS 8726, pp. 306–321, 2014.

process of feature selection. The embedded methods treat classifier as a white box, and incorporate feature selection and classification model into a single optimization problem. Thus, the classification performance is good, and the computational cost is much lower than wrapper method.

From another point of view, feature selection techniques can be categorized into supervised method (using label information) and unsupervised method (without using label information). Supervised feature selection methods determine the importance of a feature by evaluating the feature's correlation with label. The higher correlation indicates a more important feature. Unsupervised feature selection approaches select features with maximum representative and discriminant power. In the real world data mining applications, labeling data is time and labor consuming. Thus, the unsupervised feature selection methods are crucial for practical applications.

Many unsupervised feature selection methods have been proposed. Among them, maximum-variance is the simplest one, which just selects top ranked features with maximum variance. Although selected features are representative for data variance, they are not guaranteed to be discriminant for classification [9]. Laplacian Score [9] selects features that can preserve the local manifold structure of data, and such features are supposed to be discriminative. SPEC [27] selects features that are most consistent with the graph structure of data. MCFS [2] first performs regression using the eigenvector of graph Laplacian, and then selects features with maximum spectral regression coefficients.

In this work, we focus on the unsupervised feature selection model design. Most existing unsupervised feature selection methods are similar to filter methods in supervised learning, and define different score systems to select features. Considering the advantages of embedded feature selection methods in supervised learning, we hope to use the embedded feature selection mechanism in an unsupervised way. In this paper, we address this problem using the unsupervised trace ratio formulation, and rigorously prove that our unsupervised trace ratio formulation is the unified and unique objective of both trace ratio linear discriminant analysis (LDA) and K-means clustering. After that, we propose an unsupervised feature selection method using unsupervised trace ratio formulation and $\ell_{1,2}$-norm regularization. The proposed method can harness the discriminant power of trace ratio formulation, thus it tends to select discriminative features. The optimization algorithm is derived with rigorous convergence analysis. Moreover, we provide important theoretical analysis to guarantee the unsupervised feature selection process. Empirical results on four benchmark data sets show that the proposed method outperforms other sate-of-the-art unsupervised feature selection methods on all three standard evaluation metrics.

2 Notations and Definitions

In this paper, matrices are written as uppercase letters and vectors are written as bold lowercase letters. Given a matrix $W = \{w_{ij}\}$, its i-th row, j-th column are denoted as $\mathbf{w}^i$, $\mathbf{w}_j$, respectively. The $\ell_{1,2}$-norm of matrix W is defined as $||W||_{1,2} = \sum_{i=1}^{d} ||\mathbf{w}^i||_2$. $Tr(W)$ means the trace operation for matrix W.

Given data matrix $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$, d is the number of features and n is the number of data samples. In the classic Linear Discriminant Analysis (LDA), the total scatter matrix S_t, within-class scatter matrix S_w, and between-class scatter matrix S_b are defined as following:

$$\begin{aligned} S_t &= \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T, \\ S_w &= \sum_{k=1}^{c} \sum_{\mathbf{x}_i \in l_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T, \\ S_b &= \sum_{k=1}^{c} n_k (\mathbf{m}_k - \bar{\mathbf{x}})(\mathbf{m}_k - \bar{\mathbf{x}})^T, \end{aligned}$$

where $\mathbf{x}_i \in \Re^{d \times 1}$ is the i-th data sample, c is the number of clusters, n_k is the number of data points belong to class l_k, $\mathbf{m}_k$ is the center of the k-th cluster, *i.e.* $\mathbf{m}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in l_k} \mathbf{x}_i$, $\bar{\mathbf{x}}$ is the center of all data, *i.e.* $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$. It is well known that $S_t = S_b + S_w$.

Suppose $X \in \Re^{d \times n}$ is the data matrix after centralization, *i.e.* $\bar{x} = 0$, the formulations of total scatter matrix S_t and between-class scatter matrix S_b can be thus reduced to:

$$S_t = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T = XX^T, \;\; S_b = \sum_{k=1}^{c} n_k \mathbf{m}_k \mathbf{m}_k^T . \tag{1}$$

Denote $G \in \Re^{n \times c}$ as the class indicator matrix, where $G_{ij} = 1$ if x_i belongs to the j-th class and $G_{ij} = 0$ otherwise. We define a cluster centroid matrix M to include the centroid vector of each cluster as $M = [\mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_c]$. Using the class indicator matrix G, we can represent the cluster centroid matrix M as:

$$M = XG(G^T G)^{-1} . \tag{2}$$

Using matrices G and M, we can re-write the scatter matrices into more compact manner as:

$$S_b = MG^T GM^T \tag{3}$$

$$S_w = (X - MG^T)(X - MG^T)^T \tag{4}$$

3 Trace Ratio Linear Discriminant Analysis Review

In recent research, Linear Discriminant Analysis (LDA) with trace ratio criterion has shown better performance than the traditional LDA with ratio trace criterion [18,10]. Thus, the trace ratio LDA has attracted more and more attention and has been well studied. The problem of trace ratio LDA is as follows:

$$\max_{W^T W = I} \frac{Tr(W^T S_b W)}{Tr(W^T S_w W)} , \tag{5}$$

where $W \in \Re^{d \times m}$ is the projection matrix, which is constrained to be orthonormal, and m is the reduced dimension.

Using the optimal solution W of the problem (5), the data points are projected to a lower dimensional subspace such that the Euclidean distances of data pairs within the same class are minimized while the Euclidean distances of data pairs between different classes are maximized. That is to say, the data points are easy to be classified after the dimensionality reduction with W.

Because of $S_t = S_b + S_w$, problem (5) is equivalent to the following problem:

$$\max_{W^T W=I} \frac{Tr(W^T S_b W)}{Tr(W^T S_t W)}. \tag{6}$$

4 Discriminative Unsupervised Feature Selection

Because the LDA can enhance the classification tasks, several recent research works have used this criterion for supervised feature selection and shown promising results [15,21]. However, the LDA strategy cannot be applied to unsupervised feature selection, because the unsupervised learning models don't provide the data labels which are required to compute the within-class and between-class scatters. In previous work [6], the authors utilized the clustering results to calculate S_b and S_w and iteratively do LDA and K-means clustering, such that the LDA criterion can be applied to improve clustering results. However, the authors only presented a heuristic algorithm and didn't have a unified objective for two different processes, *i.e.* the LDA and K-means clustering minimize different objectives. Thus, the optimality and convergence of their algorithm are NOT guaranteed.

In this work, we are interested in designing a powerful unsupervised feature selection method. To address the above problems, we will derive a new formulation and rigorously prove it unifies both trace ratio LDA and K-means clustering, such that the trace ratio LDA criterion can be applied to unsupervised model seamlessly. Combining with the structured sparsity-inducing norms, we will propose a novel unsupervised feature selection method.

4.1 Unsupervised Dimensionality Reduction Using Trace Ratio Criterion

Trace ratio LDA is a supervised dimensionality reduction method. Plugging Eq. (3) into Eq. (6), the trace ratio LDA objective can be written as:

$$\max_{W^T W=I} \frac{Tr(W^T X G (G^T G)^{-1} G^T X^T W)}{Tr(W^T S_t W)}, \tag{7}$$

where $G \in \Re^{n \times c}$ is the class indicator matrix defined in Section 2.

In unsupervised circumstance where there is no label information, we don't know both projection matrix W and class indicator matrix G. If we apply the trace ratio strategy to unsupervised dimensionality reduction, we need solve the following problem:

$$\max_{W^T W=I, G \in Ind} \frac{Tr(W^T X G (G^T G)^{-1} G^T X^T W)}{Tr(W^T S_t W)}, \tag{8}$$

where Ind is the set of clustering indicator matrices and $G \in Ind$ means G is constrained to be a clustering indicator matrix. This is not LDA anymore. How does problem (8) reduce the data dimensionality to an unsupervised way? Our following theorem will rigorously show that the problem (8) is a unified and unique objective of both trace ratio LDA and K-means clustering.

Solving problem (8) is exactly equivalent to iteratively solving trace ratio LDA and doing K-means clustering. When G is fixed, obviously solving problem (8) is to solve the trace ratio LDA *w.r.t.* W, *i.e.* solving problem (7).

When W is fixed, $Tr(W^T S_t W)$ is irrelevant to G. Thus, we need to solve the following problem:

$$\max_{G \in Ind} Tr(W^T X G (G^T G)^{-1} G^T X^T W) \,. \tag{9}$$

Although the problem (9) only has one variable, it is difficult to solve due to the intractable constraint. Because $Tr(W^T S_t W)$ is a constant now (W is fixed), maximizing between-class distance in problem (9) is equivalent to minimizing within-class distance. Problem (9) is equivalent to the following problem:

$$\min_{G \in Ind, M} Tr(W^T S_w W) \,, \tag{10}$$

where $S_w = (X - MG^T)(X - MG^T)^T$ as shown in Eq. (4). Thus, we need to optimize:

$$\begin{aligned} & \min_{G \in Ind, M} Tr(W^T (X - MG^T)(X - MG^T)^T W) \\ \Longrightarrow & \min_{G \in Ind, M} \left\| W^T X - W^T M G^T \right\|_F^2 \\ \Longrightarrow & \min_{G \in Ind, F} \left\| W^T X - F G^T \right\|_F^2 \,, \end{aligned} \tag{11}$$

where $F = W^T M$.

Problem (11) can be easily solved by alternating optimization, *i.e.*, iteratively optimizing F when G is fixed and optimizing G when F is fixed. Interestingly, this iterative procedure is exactly the procedure of traditional K-means clustering algorithm on the projected data $W^T X$: that is, when G is fixed, the optimal solution of F is the centers of the clusters in the projected subspace; when F is fixed, the optimal solution of G can be computed by assigning the data points to their closest centers. Thus, the objective function in (9) is equivalent to K-means clustering objective.

Therefore, solving problem (8) is equivalent to iteratively solving trace ratio LDA (fix G to solve W) and doing K-means clustering (fix W to solve G).

Therefore, the objective in (8) is a good trace ratio formulation to reduce the dimensionality in an unsupervised way. The K-means clustering indicators can be used as labels to calculate scatter matrices, such that the projection matrix is discriminative to separate different data groups.

Please notice that our method is significantly different from the method in [6], where the traditional ratio trace LDA and K-means clustering algorithms are heuristically combined *without* any optimality and convergence guarantee. Our new Theorem 1 rigorously proves that the trace ratio formulation in (8) is the *unified and unique* objective

when we iteratively solve trace ratio LDA and K-means clustering. Thus, this procedure is guaranteed to converge.

Based on our above derivations, the unsupervised trace ratio formulation in (8) is equivalent to the following problem:

$$\min_{W^TW=I, G\in Ind, F} \frac{\left\|W^TX - FG^T\right\|_F^2}{Tr(W^TS_tW)} \tag{12}$$

4.2 Unsupervised Feature Selection Using Structured Sparse Trace Ratio Formulation

Both supervised and unsupervised trace ratio LDA are dimensionality reduction methods, where the projected feature is a linear combination of all original features. However, in many applications (*e.g.* bioinformatics and document mining), we are more interested in the feature selection model, *i.e.*, selecting a few relevant features. To address this problem, we integrate the structured sparsity-inducing norms with the above unsupervised trace ratio formulation, such that we can select informative features in an unsupervised way.

We hope to learn a row sparse projection matrix W in which only a few rows of W are non-zeros. With this row sparse projection matrix W, only a few important features are involved in the projection. This goal can be achieved by minimizing $\|W\|_{1,2}$. Therefore, problem (12) can be changed to the following objective for unsupervised feature selection:

$$\min_{W^TW=I, G\in Ind, F} \frac{\left\|W^TX - FG^T\right\|_F^2}{Tr(W^TS_tW)} + \gamma\|W\|_{1,2}\,, \tag{13}$$

where γ is a regularization parameter which controls the row sparsity of the projection matrix W. The greater the γ is, the more sparse rows the projection matrix W has.

The optimal solution of the problem (13) can harness the discriminative power of the unsupervised trace ratio model, thus it tends to select discriminative features. Only those discriminative features would have non-zero weights in W, and thus each new projected feature is a linear combination of only these discriminative features. In this way, only discriminative information are retained.

5 Optimization Algorithm

We use the alternating optimization method to solve the problem (13). When W is fixed, the problem becomes problem (11), which can be solved by alternating optimization. Specifically, when G is fixed, the optimal F is:

$$F = W^TXG(G^TG)^{-1}\,; \tag{14}$$

when F is fixed, the optimal G is:

$$G_{ij} = \begin{cases} 1, & j = \arg\min\limits_k \left\|W^Tx_i - f_k\right\|_2^2 \\ 0, & other \end{cases} \tag{15}$$

As mentioned before, this update of F and G is exactly the K-means procedure.

When G and F are fixed, we substitute Eq. (14) into the problem (13) and thus the problem (13) becomes

$$\min_{W^TW=I} \frac{Tr(W^TS_wW)}{Tr(W^TS_tW)} + \gamma\|W\|_{1,2}\,, \tag{16}$$

where

$$S_w = (X - XG(G^TG)^{-1}G^T)(X - XG(G^TG)^{-1}G^T)^T. \tag{17}$$

Due to the trace ratio formulation, the above objective is difficult to optimize. The standard proximal gradient, Augmented Lagrange Multiplier, fixed point, proximal methods cannot efficiently optimize it. We will use the iterative re-weighted optimization strategy to solve this objective. Solving the above objective is equivalent to solve:

$$\min_{W^TW=I} \frac{Tr(W^TS_wW)}{Tr(W^TS_tW)} + \gamma Tr(W^TDW)\,, \tag{18}$$

where D is a diagonal matrix with the i-th diagonal element $d_i = \frac{1}{2\|\mathbf{w}^i\|_2}$. When $\|\mathbf{w}^i\|_2 = 0$, the original objective is not differentiable. Following [8], we can introduce a small perturbation to regularize the i-th diagonal element of D as $\frac{1}{2\sqrt{\|\mathbf{w}^i\|_2^2+\zeta}}$. Then it can be verified that the algorithm minimizes the following problem: $\frac{Tr(W^TS_wW)}{Tr(W^TS_tW)} + \gamma\sum_{i=1}^{d}\sqrt{\|\mathbf{w}^i\|_2^2+\zeta}$ is apparently reduced to problem Eq. (16) when $\zeta \to 0$.

In the following, we derive an iterative algorithm to solve the problem (18) with a similar trick used in [17]. The Lagrangian function of the problem (18) is:

$$\begin{aligned}\mathcal{L}(W,\Lambda) = &\frac{Tr(W^TS_wW)}{Tr(W^TS_tW)} + \gamma Tr(W^TDW)\\ &-Tr(\Lambda(W^TW - I))\,.\end{aligned} \tag{19}$$

By taking the derivative *w.r.t.* W to zero, we have

$$\left(S_w - \frac{Tr(W^TS_wW)}{Tr(W^TS_tW)}S_t + \gamma Tr(W^TS_tW)D\right)W = W\Lambda\,. \tag{20}$$

Thus, the optimal solution of W is the m smallest eigenvectors of the matrix:

$$S_w - \frac{Tr(W^TS_wW)}{Tr(W^TS_tW)}S_t + \gamma Tr(W^TS_tW)D. \tag{21}$$

We can iteratively update the D and the W such that the Eq. (20), *i.e.* KKT condition, is satisfied. Please notice that D is not a variable to optimize. In the iterative steps, we optimize Eq. (21) to get W, and then re-calculate Eq. (21), where D is only an intermediate value to help calculation.

In summary, the algorithm to solve the discriminative unsupervised feature selection problem (13) is outlined in Algorithm 1. Since our formulation is based on TRACe ratio and K-means formulations, we call this algorithm as TRACK for short.

Algorithm 1. Algorithm to solve the objective of our TRACK method in (13).

Initialize D as an identity matrix.

repeat

1. Iteratively update F by Eq. (14) and update G by Eq. (15) till to converge.

2. Iteratively update the diagonal matrix D with the i-th diagonal element as $d_i = \frac{1}{2\|\mathbf{w}^i\|_2}$, and update W by the m eigenvectors corresponding to the m smallest eigenvalues of

$$S_w - \frac{Tr(W^T S_w W)}{Tr(W^T S_t W)} S_t + \gamma Tr(W^T S_t W) D \,,$$

till converge.

until Converges

5.1 Convergence Analysis

In Algorithm 1, the Step 1 is the K-means clustering procedure and converges to local optimal solution. Step 2 is the iterative re-weighted algorithm to solve problem (16), *i.e.* problem (18). In each iteration within Step 2, the objective value of problem (18) is decreased until the algorithm converges. The proof is similar to [1,16], and thus we omit it due to limited space. When the Step 2 converges, Eq. (20) is satisfied. Note that Eq. (20) is the KKT condition of problem (18), therefore the converged solution satisfies the KKT condition of problem (18), and thus is a local optimal solution to the problem (18).

It deserves to be mentioned that, based on our unified and unique objective for both steps, Step 1 and Step 2 in Algorithm 1 are guarantied to mutually benefit each other. On the one hand, the better clustering results in Step 1 will result in better scatter matrices, and thus results in more discriminative projection matrix in Step 2; On the other hand, the more discriminative projection matrix in Step 2 will make the data more separable, thus lead to better clustering results in Step 1.

5.2 Theoretical Analysis for Feature Selection

To guarantee the unsupervised feature selection process, we provide the following important theorems on the problem (13). First, we will show that our method guarantees to have m features for selection, *i.e.* the sparsity shrinkage won't over suppress the non-zero rows in W. Second, we will prove that using the $\ell_{1,2}$-norm regularization in our TRACK objective is equivalent to using the $\ell_{0,2}$-norm regularization, which is the ideal feature selection formulation.

Theorem 1. *The number of non-zero rows of the optimal solution to the problem (13) will not be less than* m.

Proof: Because $W \in \Re^{d \times m}$ and $W^T W = I$, the rank of W is m. Thus, the number of non-zero rows of any feasible solution to the problem (13) will not smaller than m, otherwise the rank of W is smaller than m. □

Theorem 1 indicates the selected feature number is at least m by solving the problem (13) with even a very large γ. This is important, because the sparse learning based

feature selection methods could over suppress the non-zero rows such that there are no enough features for selection.

Moreover, we have the following theorem, which indicates that minimizing the $\ell_{1,2}$-norm of W in our TRACK objective is equivalent to minimizing the $\ell_{0,2}$-norm of W under the constraint of $W^TW = I$.

Theorem 2. *Let $W \in \Re^{d\times m}$. The optimal solutions to the problem $\min_{W^TW=I} \|W\|_{1,2}$ and the optimal solutions to the problem $\min_{W^TW=I} \|W\|_{0,2}$ are the same.*

Proof: Obviously, the optimal solution W^* to the problem $\min_{W^TW=I} \|W\|_{0,2}$ is any matrix with only m non-zero rows, and the matrix with the m non-zero rows is an orthonormal matrix. Without loss of generality, suppose $W^* = \begin{bmatrix} W_1 \\ 0 \end{bmatrix}$, where $W_1 \in \Re^{m\times m}$ is an orthonormal matrix, then we have $\|W^*\|_{1,2} = m$. For any matrix $W \in \Re^{d\times m}$ with the constraint $W^TW = I$, we can construct an orthonormal matrix $\hat{W} = [W, W^\perp] \in \Re^{d\times d}$, then the i-th row of $\hat{W}$ has $\|\hat{\mathbf{w}}_i\|_2 = 1$, and then the i-th row of W has $\|\mathbf{w}_i\|_2 \leq 1$. So we have $\|\mathbf{w}_i\|_2 \geq \|\mathbf{w}_i\|_2^2$, and then:

$$\|W\|_{1,2} \geq \|W\|_F^2 = m = \|W^*\|_{1,2} \ . \tag{22}$$

Therefore, W^* is the optimal solution to the problem $\min_{W^TW=I} \|W\|_{1,2}$. □

Therefore, in our TRACK method, features selected by the $\ell_{1,2}$-norm regularization are the same as using the ideal $\ell_{0,2}$-norm regularization.

6 Experimental Results

In this section, we compare the proposed TRACK feature selection algorithm with other state-of-the-art unsupervised feature selection algorithms: Maximum-Variance (MaxVar), Laplacian Score (LS) [9], SPEC [27] and MCFS [2], and ldaKm [6].

6.1 Brief Descriptions of Comparison Methods

We briefly describe the comparison methods in this section. MaxVar is the simplest unsupervised feature selection algorithm, which just select top ranked features with maximum variance. Although selected features are representative for data variance, they are not guaranteed to be discriminant for classification [9].

Laplacian Score selects features that can preserve the local manifold structure of data, and such features are supposed to be discriminative. It computes the score for each feature as $S_i = \frac{\hat{f}_i{}^T \mathcal{L} \hat{f}_i}{\hat{f}_i{}^T D \hat{f}_i}$, where $\mathcal{L}$ is the graph Laplacian, and $\hat{f}_i = f_i - \frac{\hat{f}_i{}^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}}\mathbf{1}$.

SPEC algorithm selects features that are most consistent with the graph structure of data. It computes the score for each feature as $S_i = \hat{f}_i{}^T \mathcal{L} \hat{f}_i$, where $\hat{f}_i = \frac{D^{\frac{1}{2}} f_i}{\|f_i\|}$.

MCFS algorithm first performs regression using the eigenvector of graph Laplacian, and then selects features with maximum spectral regression coefficients. The regression

problem is formulated as $\min_{a_k} \left\| y_k - X^T a_k \right\|_F^2$, where y_k is the k_{th} eigenvector of the graph Laplacian matrix, a_k is the spectral regression coefficients. The score for the i_{th} feature is defined as $S_i = \max_k |a_{k,i}|$.

LdaKm is an adaptive dimensionality reduction method that integrates K-means clustering and LDA. The ldaKm alternatively performs the following two steps: (1) perform K-means clustering on projected space; (2) perform traditional ratio trace LDA to get the projection matrix. Following our approach, $\ell_{1,2}$-norm regularization is used to select features for ldaKm method.

6.2 Data Sets and Evaluation Metrics

Four real world data sets are used to validate the effectiveness of our TRACK feature selection algorithm: MSRC-V1, ORL, JAFFE, and XM2VTS.

MSRC-V1 database is from Microsoft Research in Cambridge. This data set contains coarse pixel-wise labeled images, and it is commonly used for full scene segmentation.

ORL database contains a set of face images taken between April 1992 and April 1994 at the ATT lab. Ten different images are taken for each of the 40 distinct subjects. For some subjects, the images were taken at different times, with different light condition, facial expressions (*i.e.*: smiling or not smiling, open or closed eyes). All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position.

JAFFE (Japanese Female Facial Expression) database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models, which were taken at the Psychology Department in Kyushu University. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects.

XM2VTS (Extended Multi Modal Verification for Teleservices and Security applications) database is a large multi-modal database which was captured onto high quality digital video. It contains four recordings of 295 subjects taken over a period of four months. Sets of data taken from this database are available including high quality color images, 32 KHz 16-bit sound files, video sequences and a 3d Model.

Important statistics of the data sets are summarized in Table 1.

Table 1. Data set descriptions

	sample #	feature #	class #
MSRC-V1	210	1302	7
ORL	400	644	40
JAFFE	213	1024	10
XM2VTS	1180	1024	295

Three measures are used to evaluate the clustering performance of all methods: accuracy, normalized mutual information (NMI) and purity.

Accuracy is the percentage of correct predicted label. Because the real label of each cluster is unknown, the Hungarian algorithm [14] is used to get the best map to the real label. Let C denotes the ground truth label, C' denotes the label obtained from a clustering algorithm, the mutual information (MI) is defined as:

$$MI(C, C') = \Sigma_{c_i \in C, c'_j \in C'} p(c_i, c'_j) log \frac{p(c_i, c'_j)}{p(c_i) p(c'_j)} \tag{23}$$

where $p(c_i), p(c'_j)$ are the probability of a arbitrarily selected sample belongs to cluster c_i, c'_j, respectively. $p(c_i, c'_j)$ is the probability of a arbitrarily selected sample belongs to both cluster c_i and c'_j.

NMI is the normalized MI as following:

$$NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))} \tag{24}$$

where $H(C) and H(C')$ are the entropies of C and C', respectively.

Purity is computed by assigning the label of a cluster to the most frequent class. More formally, it is defined as:

$$purity(C, C') = \frac{1}{N} \Sigma_j \max_i (c'_j \cap c_i) \tag{25}$$

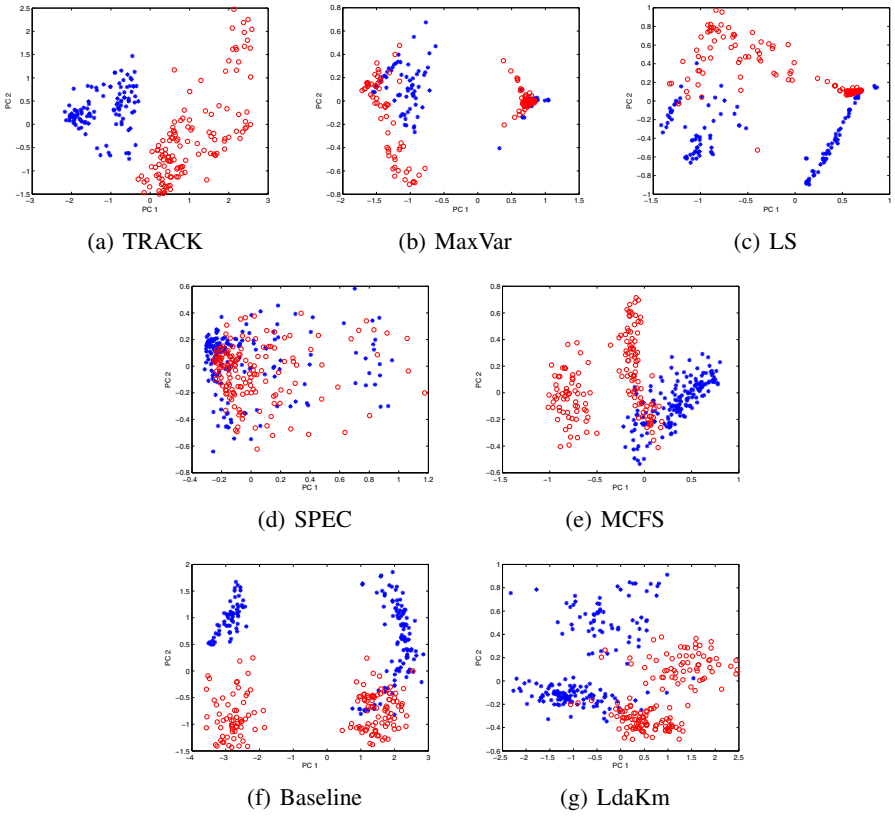
6.3 Demonstration of Discriminant Power of Selected Features

In this section, we show the discriminant power of selected features by various algorithms. We use different unsupervised feature selection algorithms to select top 30 features on the MSRC-V1 data set. Then selected features are used to perform principle component analysis (PCA), and data samples are projected onto the first 2 principle components (PC), as shown in Figure 1 (PCA performed using top 30 features). For the baseline method, all features are used to perform PCA.

From Figure 1, we can see that: The TRACK algorithm separates data much better than other feature selection algorithms. The MCFS and ldaKm algorithms perform slightly better than the remaining algorithms. Data are much more entangled with each other using the MaxVar and SPEC algorithm. This shows that: the TRACK algorithm can harness the discriminant power of trace ratio formulation, therefore, features selected by the TRACK algorithm are much more discriminant than those selected by other algorithms, and using those discriminant features can separate data from different classes well.

6.4 Clustering Performance Comparison

We select top 10 till to top 100 features using different methods, and perform K-means using the selected features to evaluate the clustering performance. Since K-means clustering is sensitive to initialization, we perform 20 trials and record the average clustering metric. The result of using all features is also reported as a baseline. The



(a) TRACK (b) MaxVar (c) LS

(d) SPEC (e) MCFS

(f) Baseline (g) LdaKm

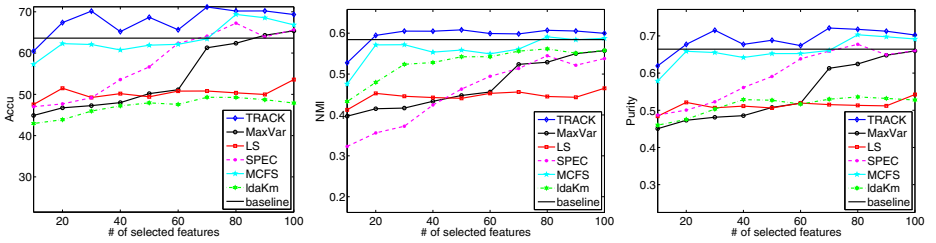
Fig. 1. Projection on first two principle components (PC) using top 30 features selected by various feature selection algorithms on the MSRC-V1 data set. The horizontal axis is the score of the first principle component, and the vertical axis is the score of the second principle component. Different shape or color mark samples from different classes.

regularization parameter is tuned from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ for both the TRACK algorithm and the ldaKm algorithm. The reduced dimension m in our method is set as: $m = c - 1$ if $d <= n$, and $m = c - 1 + d - n$ if $d > n$, as suggested in the paper [25]. Clustering accuracy, NMI, purity on the four data sets are reported in Figures 2- 5.

From those figures, we can conclude that:

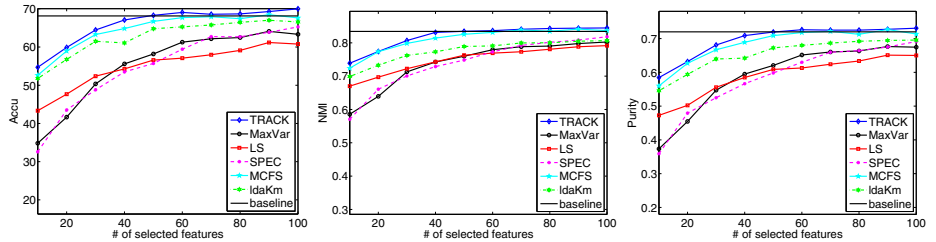
(1) On all the four data sets, our method can outperform other state-of-the-art unsupervised feature selection algorithms on all evaluation metrics. The TRACK algorithm can outperform the baseline (using all features) using just 20 to 50 features, which justifies that the TRACK algorithm is able to select the most discriminant features.

(2) Generally, clustering performance becomes better when more features are selected.



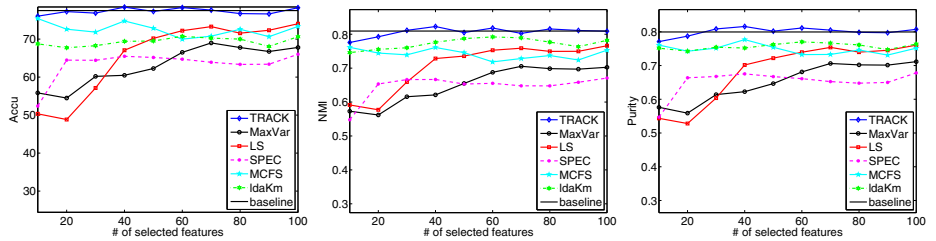
(a) Accuracy (b) NMI (c) Purity

Fig. 2. Clustering performance on MSRC-V1 data set



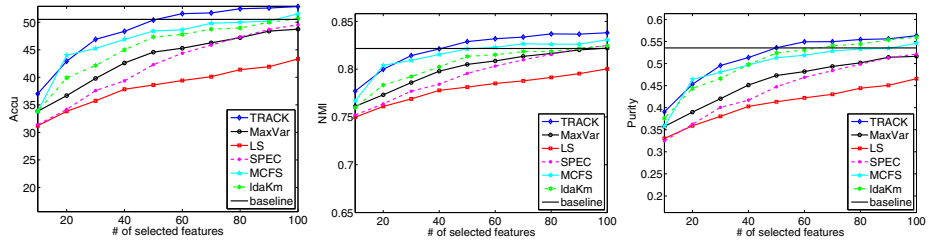
(a) Accuracy (b) NMI (c) Purity

Fig. 3. Clustering performance on ORL data set



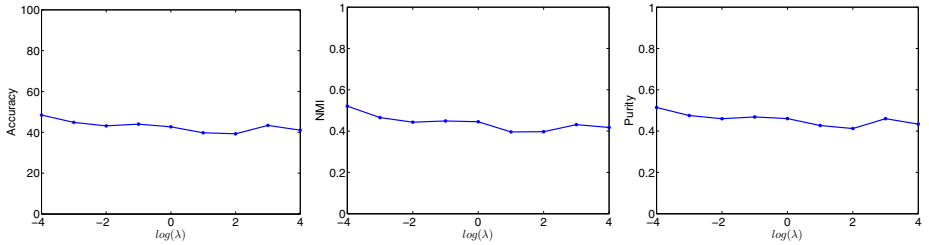
(a) Accuracy (b) NMI (c) Purity

Fig. 4. Clustering performance on JAFFE data set



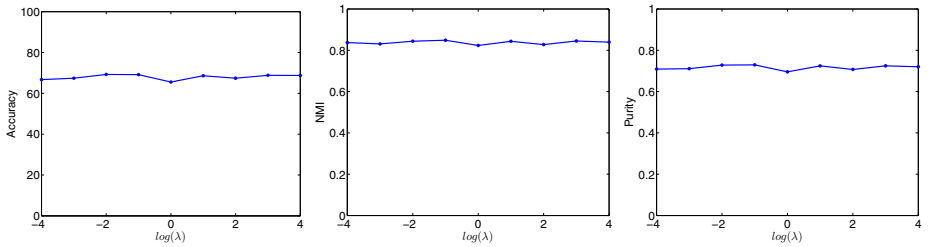
(a) Accuracy (b) NMI (c) Purity

Fig. 5. Clustering performance on XM2VTS data set



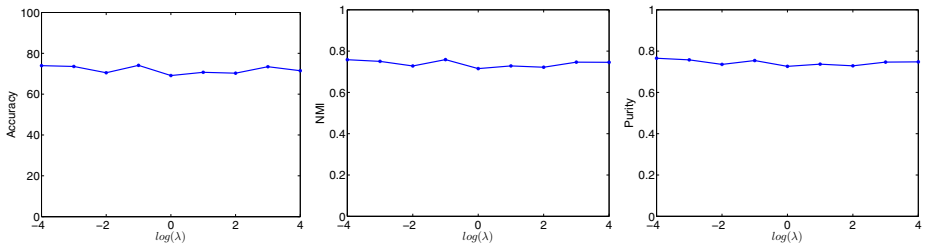
(a) Accuracy (b) NMI (c) Purity

Fig. 6. Clustering performance versus the regularization parameter on MSRC-V1 data set



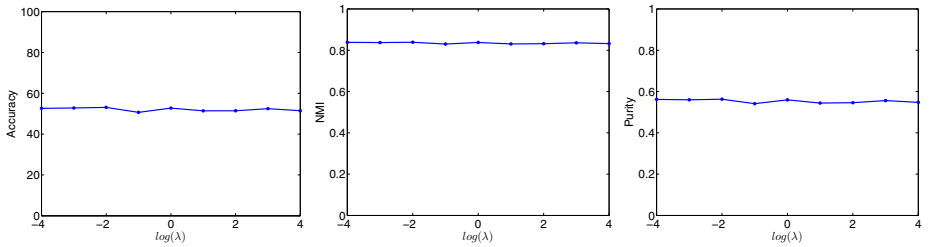
(a) Accuracy (b) NMI (c) Purity

Fig. 7. Clustering performance versus the regularization parameter on ORL data set



(a) Accuracy (b) NMI (c) Purity

Fig. 8. Clustering performance versus the regularization parameter on JAFFE data set



(a) Accuracy (b) NMI (c) Purity

Fig. 9. Clustering performance versus the regularization parameter on XM2VTS data set

(3) The MCFS algorithm performs the second best among the rest feature selection algorithms on all four data sets. Especially on ORL data set, the performance of MCFS is quite close to our TRACK algorithm.

6.5 Parameter Sensitivity

To study the sensitivity of our algorithm, we plotted the classification performance with different regularization parameters, as shown in Figure 6 to 9. From these figures, we can see that: our algorithm is not very sensitive to the regularization parameter. Therefore, the parameter is easy to be tuned.

7 Conclusion

In this paper, we first rigorously prove that the unsupervised trace ratio formulation is the unified and unique objective of both trace ratio LDA and K-means clustering. Then we propose an unsupervised feature selection method using unsupervised trace ratio formulation regularized by $\ell_{1,2}$-norm of the projection matrix. The proposed method can harness the discriminant power of trace ratio LDA, thus it tends to select discriminative features. We derive an efficient algorithm to solve the proposed model with proved convergence. Four real world data sets are used to evaluate the effectiveness of the proposed method. Empirical results show that the proposed method outperforms other sate-of-the-art unsupervised feature selection algorithms on all three valuation metrics.

References

1. Argyriou, A., Evgeniou, T., Pontil, M.: Multi-task feature learning. In: NIPS, pp. 41–48 (2007)
2. Cai, D., Zhang, C., He, X.: Unsupervised feature selection for multi-cluster data. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 333–342. ACM (2010)
3. Cai, X., Nie, F., Huang, H., Ding, C.: Feature selection via l2,1-norm support vector machine. In: IEEE International Conference on Data Mining (2011)
4. Chen, C.H., Pau, L.F., Wang, P.S.P.: Handbook of pattern recognition and computer vision. World Scientific (2010)
5. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology 3(02), 185–205 (2005)
6. Ding, C., Li, T.: Adaptive dimension reduction using discriminant analysis and k-means clustering. In: International Conference on Machine Learning, pp. 521–528 (2007)
7. Forman, G.: An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research 3, 1289–1305 (2003)
8. Gorodnitsky, I., Rao, B.: Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. IEEE Transactions on Signal Processing 45(3), 600–616 (1997)
9. He, X., Cai, D., Niyogi, P.: Laplacian score for feature selection. Advances in Neural Information Processing Systems 18, 507 (2006)
10. Jia, Y., Nie, F., Zhang, C.: Trace ratio problem revisited. IEEE Transactions on Neural Networks 20(4), 729–735 (2009)

11. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
12. Kong, D., Ding, C., Huang, H., Zhao, H.: Multi-label relieff and f-statistic feature selections for image annotation. In: The 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2352–2359 (2012)
13. Kononenko, I.: Estimating attributes: analysis and extensions of relief. In: Bergadano, F., De Raedt, L. (eds.) ECML 1994. LNCS, vol. 784, pp. 171–182. Springer, Heidelberg (1994)
14. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2(1-2), 83–97 (1955)
15. Masaeli, M., Fung, G., Dy, J.G.: From transformation-based dimensionality reduction to feature selection. In: ICML, pp. 751–758 (2010)
16. Nie, F., Huang, H., Cai, X., Ding, C.: Efficient and robust feature selection via joint l2,1-norms minimization. Advances in Neural Information Processing Systems 23, 1813–1821 (2010)
17. Nie, F., Xiang, S., Jia, Y., Zhang, C.: Semi-supervised orthogonal discriminant analysis via label propagation. Pattern Recognition 42(11), 2615–2627 (2009)
18. Nie, F., Xiang, S., Jia, Y., Zhang, C., Yan, S.: Trace ratio criterion for feature selection. In: AAAI, pp. 671–676 (2008)
19. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
20. Raileanu, L.E., Stoffel, K.: Theoretical comparison between the gini index and information gain criteria. Ann. Math. Artif. Intell. 41(1), 77–93 (2004)
21. Wang, C., Caob, L., Miao, B.: Optimal feature selection for sparse linear discriminant analysis and its applications in gene expression data. Computational Statistics and Data Analysis 66, 140–149 (2013)
22. Wang, D., Nie, F., Huang, H., Yan, J., Risacher, S.L., Saykin, A.J., Shen, L.: Structural brain network constrained neuroimaging marker identification for predicting cognitive functions. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 536–547. Springer, Heidelberg (2013)
23. Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S.L., Saykin, A.J., Shen, L.: Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. Bioinformatics 28(2), 229–237 (2012)
24. Wang, H., Nie, F., Huang, H., Risacher, S., Ding, C., Saykin, A.J., Shen, L.: ADNI: Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In: IEEE Conference on Computer Vision (2011)
25. Xiang, S., Nie, F., Zhang, C.: Learning a mahalanobis distance metric for data clustering and classification. Pattern Recognition 41(12), 3600–3612 (2008)
26. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of The Royal Statistical Society Series B 68(1), 49–67 (2006)
27. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In: Proceedings of the 24th International Conference on Machine Learning, pp. 1151–1157. ACM (2007)