2011 International Conference on Digital Image Computing: Techniques and Applications

On The Optimality of Sequential Forward Feature Selection Using Class Separability Measure

Lei Wang*, Chunhua Shen†, and Richard Hartley‡
*School of Computer Science and Software Engineering
University of Wollongong, Wollongong, Australia 2522
Email: leiw@uow.edu.au
†School of Computer Science
University of Adelaide, Adelaide, Australia 5005
Email: chunhua.shen@adelaide.edu.au
‡Research School of Engineering
Australian National University, Canberra, Australia 0200
Email: richard.hartley@anu.edu.au

***Abstract*—This paper studies sequential forward feature selection that uses the scatter-matrix-based class separability measure. We find that by adding a scale factor to each iteration of the conventional sequential selection, a sequential selection that guarantees the global optimum can be attained. We give a thorough theoretical proof of its optimality via a novel geometric interpretation, and this leads to a unified framework including the optimal sequential selection, the conventional sequential selection and the best-individual-N selection. In addition, we show that with our formulation, feature selection can be treated as a linear fractional maximization problem, and it can be efficiently solved by algorithms well developed in the literature. This gives a non-sequential globally optimal feature selection algorithm. Both theoretical and experimental study demonstrate their efficiency.**

***Keywords*-sequential; feature selection; class separability;**

I. Introduction

Feature subset selection is a fundamental problem in computer vision and pattern recognition [1], [2], [3], [4]. It aims to identify k important features from the original d ones by using a certain selection criterion. By feature selection, the dimensions of an input space can be reduced. This helps the algorithms that suffer from "curse of dimensionality". Also, it can remove noisy and Feature subset selection is essentially a combinatorial problem. There are C_d^k possible ways to choose k features out of d ones. In this situation, the commonly used selection algorithms are the Sequential Forward Selection (SFS) or a simpler Best-Individual-N selection (BIN) method which individually selects the k features having larger criterion values [2], [5]. Both selection algorithms are regarded as being suboptimal in the literature. For example, SFS is often criticized as a greedy algorithm in each iteration and this prevents it from achieving the global optimality of a given selection criterion. However, to our best knowledge, there is little work in the literature studying how to achieve a global-optimum-guaranteed sequential feature selection. This paper reports our recent research work in this line, with the focus on feature selection using the scatter-matrix-based class separability measure.

As the first contribution of this work, we will answer the following questions in a systematical way: i) How to make a sequential feature selection optimal? We show that by simply adding a scale factor to each iteration of the conventional sequential selection, a globally optimal SFS algorithm can be attained for the class separability measure. It can swiftly find the optimal selection result for thousands or millions of features; ii) What is the relationship between the optimal SFS, the conventional SFS and BIN? We unify the three algorithms into a single framework and clearly reveal their differences; iii) In what case can the conventional SFS be optimal? In answering the second question, we find that the conventional SFS is guaranteed to be optimal only when the number of selected features k is no more than 2. All the above analysis is based on a novel geometric interpretation that we propose for feature subset selection.

As the second contribution, we show that feature selection with the class separability measure is essentially a linear fractional optimization problem. Based on the above geometric interpretation this becomes clear. This finding is important because it allows feature selection to be optimally solved by well-established algorithms for fractional optimization. In this work, we propose an efficient feature subset selection based on the Dinkelbach's algorithm [6], [7]. Also, we envision that many achievements in the literature of fractional optimization can be readily employed to solve more general and difficult feature selection problems. We will investigate along this line in our future work.

Before ending this section, we would like to emphasize that in the rest of the paper, i) *All the analysis is for feature selection using the scatter-matrix-based class separability measure as the selection criterion*. To keep brevity, this criterion is called "class separability measure" in short; ii) The "optimality" in this work means that the selected

978-0-7695-4588-2/11 $26.00 © 2011 IEEE
DOI 10.1109/DICTA.2011.41

k features can maximize a predefined feature selection criterion.

II. Related work

Sequential forward selection. Sequential forward selection (SFS) [2] starts from an empty set of selected features. In each iteration, one feature is transferred from a feature pool to the selected feature set. The transferred feature satisfies that when it is added into the selected set, the selection criterion computed with all the selected features can be maximized. This process repeats until the size of the selected feature set reaches a predefined number, k.

Best-Individual-N selection. Best-Individual-N (BIN) selection [2] may be the simplest selection algorithm. It computes the selection criterion by using each feature individually and selects the k features having larger criterion values.

Class separability measure. This measure involves *Within-class scatter matrix* ($\mathbf{S}_W$), *Between-class scatter matrix* ($\mathbf{S}_B$), and *Total scatter matrix* ($\mathbf{S}_T$). Let $(\mathbf{x}, y) \in (\mathbb{R}^n \times \mathcal{Y})$ denote a training sample, where $\mathbb{R}^n$ stands for an n-dimensional input space, and $\mathcal{Y} = \{1, 2, \cdots, c\}$ is the set of c class labels. The number of samples in the i-th class is n_i. Let $\mathbf{m}_i$ be mean vector of the i-th class and $\mathbf{m}$ be mean vector of all classes. The scatter matrices are defined as

$$\begin{aligned}
\mathbf{S}_W &= \sum_{i=1}^{c} \left[\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)^\top\right] \\
\mathbf{S}_B &= \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^\top \\
\mathbf{S}_T &= \sum_{i=1}^{c} \left[\sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \mathbf{m})(\mathbf{x}_{ij} - \mathbf{m})^\top\right]
\end{aligned}$$

Two properties of the matrices are i) $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$ and ii) all the matrices are Positive Semi-Definite (PSD). These properties will be used in this work. A large class separability means small within-class scattering but large between-class scattering. A combination of two of the matrices can be used as a measure. Here, we focus on the measure of $\mathrm{tr}(\mathbf{S}_B)/\mathrm{tr}(\mathbf{S}_T)$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. As shown in the following, it has a set of nice properties that makes a global-optimum-guaranteed sequential feature selection possible.

III. Our geometric view of feature selection

Feature selection with the class separability measure can be formally stated as follows: *Given n training samples from c classes, each of which is represented by d features, $x_1, x_2, \cdots, x_d$, select a subset of k ($k < d$) features that maximizes* $\mathrm{tr}(\mathbf{S}_B)/\mathrm{tr}(\mathbf{S}_T)$. In this work, a novel geometric interpretation of feature selection is proposed as follows.

Lemma 1. *For feature selection with the class separability measure, each of the d features can be mapped to a point in a 2D cartesian coordinate system. Feature selection becomes identifying a set of k points whose centroid has the largest slope with respect to the origin.*

Proof. Assume that k features are selected out of d features and are denoted by x_{r_1},x_{r_2},$\cdots$,x_{r_k}. Let $\mathbf{x}_i = (x_{ir_1}, \cdots, x_{ir_k})^\top$ $(i = 1, \cdots, n)$ be the i-th training sample represented by the k selected features. $\mathbf{K}$ is the Gram matrix defined by $\{\mathbf{K}\}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$. It can be shown that

$$\{\mathbf{K}\}_{ij} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{t=1}^{k} x_{ir_t} x_{jr_t} \triangleq \sum_{t=1}^{k} \{\mathbf{G}_{r_t}\}_{ij}, \qquad (1)$$

where $\mathbf{G}_{r_t}$ is the Gram matrix computed by using the selected feature x_{r_t}, where $t = 1, 2, \cdots, k$. It can be written in a matrix form as $\mathbf{G}_{r_t} = \mathbf{X}_{r_t} \mathbf{X}_{r_t}^\top$, where $\mathbf{X}_{r_t} = (x_{1r_t}, \cdots, x_{nr_t})^\top$. From Eq.(1), it can be obtained that

$$\mathbf{K} = \textstyle\sum_{t=1}^{k} \mathbf{G}_{r_t}. \qquad (2)$$

The two terms of $\mathrm{tr}(\mathbf{S}_B)/\mathrm{tr}(\mathbf{S}_T)$ can be expressed as

$$\mathrm{tr}(\mathbf{S}_B) = \sum_{i=1}^{c} \frac{\mathbf{1}^\top \mathbf{K}^i \mathbf{1}}{n_i} - \frac{\mathbf{1}^\top \mathbf{K} \mathbf{1}}{n} \qquad (3)$$

and

$$\mathrm{tr}(\mathbf{S}_T) = \mathrm{tr}(\mathbf{K}) - \frac{\mathbf{1}^\top \mathbf{K} \mathbf{1}}{n}, \qquad (4)$$

where $\mathbf{K}^i$ is the part of $\mathbf{K}$ computed by using training samples from class i. The similar definition applies to $\mathbf{G}^i$ used below. By combining Eq.(2) and (3), we obtain

$$\begin{aligned}
\mathrm{tr}(\mathbf{S}_B) &= \textstyle\sum_{i=1}^{c} \frac{\mathbf{1}^\top \sum_{t=1}^{k} \mathbf{G}_{r_t}^i \mathbf{1}}{n_i} - \frac{\mathbf{1}^\top \sum_{t=1}^{k} \mathbf{G}_{r_t} \mathbf{1}}{n} \\
&= \textstyle\sum_{t=1}^{k} \left(\sum_{i=1}^{c} \frac{\mathbf{1}^\top \mathbf{G}_{r_t}^i \mathbf{1}}{n_i} - \frac{\mathbf{1}^\top \mathbf{G}_{r_t} \mathbf{1}}{n} \right) \\
&\triangleq \textstyle\sum_{t=1}^{k} f(\mathbf{X}_{r_t}) = \sum_{t=1}^{k} f_{r_t}.
\end{aligned}$$

Similarly, combining Eq.(2) and (4) gives

$$\begin{aligned}
\mathrm{tr}(\mathbf{S}_T) &= \textstyle\sum_{t=1}^{k} \left(\mathrm{tr}(\mathbf{G}_{r_t}) - \frac{\mathbf{1}^\top \mathbf{G}_{r_t} \mathbf{1}}{n} \right) \\
&\triangleq \textstyle\sum_{t=1}^{k} g(\mathbf{X}_{r_t}) = \sum_{t=1}^{k} g_{r_t}
\end{aligned}$$

Hence, in the class separability measure the information of each feature is represented by (g_{r_t}, f_{r_t}), which corresponds to a point a 2D cartesian coordinate system. Following the results, the selection criterion can be written as

$$J = \frac{\mathrm{tr}(\mathbf{S}_B)}{\mathrm{tr}(\mathbf{S}_T)} = \frac{\sum_{t=1}^{k} f_{r_t}}{\sum_{t=1}^{k} g_{r_t}} = \frac{0 - \frac{1}{k} \sum_{t=1}^{k} f_{r_t}}{0 - \frac{1}{k} \sum_{t=1}^{k} g_{r_t}}. \qquad (5)$$

Viewing (g_{r_t}, f_{r_t}) as a point, $(\frac{1}{k} \sum_{t=1}^{k} g_{r_t}, \frac{1}{k} \sum_{t=1}^{k} f_{r_t})$ becomes the centroid of (g_{r_1}, f_{r_1}), (g_{r_2}, f_{r_2}), $\cdots$, and (g_{r_k}, f_{r_k}). J is just the slope of the centroid with respect to the origin $(0, 0)$. Hence, geometrically, feature selection here is to identify the k points whose centroid has the largest slope with respect to the origin. ■

IV. A GLOBALLY OPTIMAL SEQUENTIAL SELECTION

Firstly, we need to identify the area in a cartesian coordinate system where the d points (g_t, f_t) $(t = 1, \cdots, d)$ possibly reside. Let $\mathbf{S}_B^t$ and $\mathbf{S}_T^t$ denote between-class and total scatter matrices computed by using the t-th feature. Since both $\mathbf{S}_B^t$ and $\mathbf{S}_T^t$ are PSD, it can be shown that

$$\mathrm{tr}(\mathbf{S}_B^t) = \sum_{i=1}^{c} \frac{\mathbf{1}^\top \mathbf{G}_t^i \mathbf{1}}{n_i} - \frac{\mathbf{1}^\top \mathbf{G}_t \mathbf{1}}{n} = f_t \geq 0$$

and

$$\mathrm{tr}(\mathbf{S}_T^t) = \mathrm{tr}(\mathbf{G}_t) - \frac{\mathbf{1}^\top \mathbf{G}_t \mathbf{1}}{n} = g_t > 0^1.$$

Moreover, recall that there is $\mathbf{S}_T - \mathbf{S}_B = \mathbf{S}_W$ and $\mathbf{S}_W$ is PSD. Thus, we obtain

$$\mathrm{tr}(\mathbf{S}_W^t) = \mathrm{tr}(\mathbf{S}_T^t) - \mathrm{tr}(\mathbf{S}_B^t) = g_t - f_t \geq 0$$

Therefore, all the d points must reside in the first quadrant of the cartesian coordinate system and must not be above the line $f - g = 0$. Figure 1 illustrates the area where the points can distribute and the geometric interpretation.

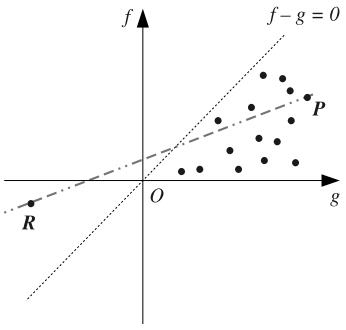

Figure 1. Illustration of our geometric interpretation of feature selection in Lemma 1. A black point $P(g, f)$ denotes each of the d features. R denotes the reference point in Corollary 1.

Based on above result, we derive the globally optimal sequential feature selection. The proofs can be found in the first author's homepage.

Case of $k = 1$. When $k = 1$, feature selection finds the point (g_{r_1}, f_{r_1}) having the maximal slope, f_{r_1}/g_{r_1}, with respect to the origin. In this case, simply selecting the point with the largest slope guarantees the optimum.

Case of $k \geq 2$. To consider the general case of $k \geq 2$, we first give Lemma 2.

Lemma 2. *Let $(g_{r_1}, f_{r_1}), \cdots, (g_{r_l}, f_{r_l})$ denote a set of l $(l \geq 1)$ points where $f_{r_i} \geq 0$ and $g_{r_i} > 0$. Given two points $(\hat{g}, \hat{f})$ and $(\check{g}, \check{f})$ that satisfy*

$$\frac{\hat{f}}{\hat{g}} \geq \frac{f_{r_i}}{g_{r_i}}, \qquad \frac{\check{f}}{\check{g}} \leq \frac{f_{r_i}}{g_{r_i}} \qquad \forall i = 1, \cdots, l,$$

it can be shown that

$$\frac{\hat{f}}{\hat{g}} \geq \frac{f_{r_1} + f_{r_2} + \cdots + f_{r_l}}{g_{r_1} + g_{r_2} + \cdots + g_{r_l}} \geq \frac{\check{f}}{\check{g}}.$$

Based on Lemma 2, we prove a key theorem in our work.

Theorem 1. *When class separability measure J is maximized, one of the k $(1 \leq k < d)$ selected points must be the point having the largest slope with respect to the origin.*

This result is interesting because it is a bit different from what we thought before. Commonly, when selecting multiple features, we often think that picking the feature having the largest criterion value may be greedy and that selecting two or more features with smaller criterion values together might be a better option. Theorem 1 shows that we need not worry this any more and the feature with the largest criterion value must be selected.

Based on Theorem 1, we can obtain Corollary 1. It gives a rule to optimally select the remaining $k - 1$ features.

Corollary 1. *Suppose that l $(1 \leq l < k)$ points have been selected as $(g_{r_1}, f_{r_1}), \cdots, (g_{r_l}, f_{r_l})$. To maximize the criterion J, one of the remaining $k - l$ selected points must be the point that has the largest slope with respect to a reference point $(-\frac{1}{k-l}\sum_{i=1}^{l} g_{r_i}, -\frac{1}{k-l}\sum_{i=1}^{l} f_{r_i})$.*

V. GEOMETRIC INTERPRETATION OF SFS AND BIN

The above geometric interpretation can be used to analyze the conventional SFS and BIN selection algorithms. Based on the analysis, we give a unified selection algorithm that accommodates the optimal SFS, conventional SFS and BIN. This allows us to clearly observe the difference among the three algorithms.

Given that l $(1 \leq l < k)$ points have been selected as $(g_{r_1}, f_{r_1}), \cdots, (g_{r_l}, f_{r_l})$, the conventional SFS selects the $(l + 1)$-th point as the one maximizing

$$J = \frac{f_{r_1} + \cdots + f_{r_l} + f_{r_{l+1}}}{g_{r_1} + \cdots + g_{r_l} + g_{r_{l+1}}}.$$

The above equation can be rewritten as

$$J = \frac{f_{r_{l+1}} - (-\sum_{i=1}^{l} f_{r_i})}{g_{r_{l+1}} - (-\sum_{i=1}^{l} g_{r_i})}.$$

Geometrically, the SFS chooses the point that has the largest slope with respect to a reference point $(-\sum_{i=1}^{l} g_{r_i}, -\sum_{i=1}^{l} f_{r_i})$. Compared with the optimal SFS developed above, the mere difference is the absence of scale factor $\frac{1}{k-l}$. Also, we find two interesting properties of the conventional SFS as follows.

Corollary 2. *The conventional sequential forward feature selection will be globally optimal for selecting k features out of d ones when $k \leq 2$.*

[1]We exclude any feature whose value is identical for all training samples regardless of class labels. Such a feature has no discriminative ability and can be easily detected by data preprocessing. Hence, g_t will not be zero.

Table I
A UNIFIED ALGORITHM FOR GLOBALLY OPTIMAL SFS, CONVENTIONAL SFS AND BIN

Input: n training samples represented as $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{1, \cdots, c\}$,
d is the total number of features and y_i is the class label of $\mathbf{x}_i$, k is the number of the features to be selected, where $1 \leq k < d$.
Output: a set of k selected points and the corresponding k features.
Initialization:
compute (g_i, f_i) $(i = 1, 2, \cdots, d)$ and store them in memory, **Initialize** the reference point (g_{ref}, f_{ref}) as $(0, 0)$
Feature selection:
for $t = 1, \cdots, k$
(1) From the remaining $d - t + 1$ points, **select** the one having the largest slope with respect to (g_{ref}, f_{ref})
(2) **Remove** the selected point from the feature pool

(3.1) Optimal SFS: **Update** (g_{ref}, f_{ref}) to $(-\frac{1}{k-t}\sum_{i=1}^{t} g_{r_i}, -\frac{1}{k-t}\sum_{i=1}^{t} f_{r_i})$
(3.2) SFS: **Update** (g_{ref}, f_{ref}) to $(-\sum_{i=1}^{t} g_{r_i}, -\sum_{i=1}^{t} f_{r_i})$
(3.3) BIN: Always **Fix** (g_{ref}, f_{ref}) as $(0, 0)$
end

However, for the case of $k \geq 3$, in the second iteration the reference point of SFS is $R(-g_1, -f_1)$, whereas the reference point of optimal SFS is $R^\star(-\frac{1}{k-1}g_1, -\frac{1}{k-1}f_1)$. Assume that for two points $P_i(g_i, f_i)$ and $P_j(g_j, f_j)$, the line $\overline{P_iP_j}$ passes between R and $R^\star$. Connecting R and $R^\star$ to P_i and P_j respectively, it can be known that the point having the large slope is different to R and $R^\star$. The two algorithms will choose different points in this iteration. It becomes difficult to judge the optimality of the conventional SFS in this situation.

Corollary 3. *The conventional sequential forward feature selection will be globally optimal for selecting k features from d ones if its $k-1$ previously selected features are identical to those selected by the optimal SFS.*

The Best-Individual-N selection (BIN) algorithm simply selects the k features that have larger criterion values. Geometrically, BIN just selects the k points having larger slope with the origin $(0, 0)$. Thinking the BIN as a sequential forward selection, it can be known that the BIN always uses the origin $(0, 0)$ as the "reference point". Thus, we can put the optimal SFS, SFS and BIN into a unified algorithm in Table I. As shown, *based on the geometric interpretation, the difference among the optimal SFS, the conventional SFS and the BIN is merely the different ways to set the reference point.*

VI. AN OPTIMAL NON-SEQUENTIAL SELECTION

In developing the optimal sequential feature selection, we also find that our formulation in Eq.(5) turns feature selection to be a 0-1 linear fractional programming problem. It can be readily solved by well-developed algorithms in the literature. Linear fractional programming (or ratio optimization) aims to maximize or minimize the ratio of two linear functions. It is widely used to find the best ratio of cost/time or return/profit and is also used to solve the problems of minimum ratio cycles, minimum ratio spanning trees, minimum ratio pathes, etc [8], [9]. There has been a large body of analysis and algorithms on fractional programming. Particularly, the 0-1 fractional programming problem has been the well-studied one. We follow the description in [7] and apply it to our feature selection problem. Rewrite our formulation in Eq.(5) as linear fractional programming problem,

$$\begin{aligned} \text{maximize} \quad & J = \frac{\mathrm{tr}(\mathbf{S}_B)}{\mathrm{tr}(\mathbf{S}_T)} = \frac{\sum_{t=1}^{d} w_t f_t}{\sum_{t=1}^{d} w_t g_t} = \frac{\mathbf{f}^\top \mathbf{w}}{\mathbf{g}^\top \mathbf{w}} \\ \text{subject to} \quad & \mathbf{w} \in \{0,1\}^d, \quad \mathbf{w}^\top \mathbf{1} = k \end{aligned} \tag{6}$$

where $\mathbf{f} = (f_1, f_2, \cdots, f_d)^\top$ and $\mathbf{g} = (g_1, g_2, \cdots, g_d)^\top$ are coefficient vectors and $\mathbf{w}$ is a 0-1 column vector to be optimized. Note that the condition of $\mathbf{g}^\top \mathbf{w} > 0$ has been justified in footnote 1. This fractional programming problem can be solved via a sequence of 0-1 integer programming sub-problems. In our case, the sub-problem can be easily solved by ranking and its optimality is guaranteed. Denoting $\lambda^\star$ the maximum value of the problem in Eq.(6), it is known that

$$\frac{\mathbf{f}^\top \mathbf{w}}{\mathbf{g}^\top \mathbf{w}} \leq \lambda^\star \Longleftrightarrow \mathbf{f}^\top \mathbf{w} - \lambda^\star \mathbf{g}^\top \mathbf{w} \leq 0,$$

where the equality is achieved when $\mathbf{w} = \mathbf{w}^\star$. For a given real number λ, the sub-problem solves

$$\begin{aligned} z(\lambda) \quad &\triangleq \quad \max_{\mathbf{w}} \ \left(\mathbf{f}^\top \mathbf{w} - \lambda \mathbf{g}^\top \mathbf{w}\right) \\ \text{subject to} \quad & \mathbf{w} \in \{0,1\}^d, \quad \mathbf{w}^\top \mathbf{1} = k. \end{aligned} \tag{7}$$

Here $z(\lambda)$ is the optimal function value of this maximization problem. $z(\lambda)$ is a piece-wise linear, convex and strictly decreasing function. It have been shown in [7] that i) $z(\lambda) = 0$ if and only if $\lambda = \lambda^*$ and ii) when $\lambda = \lambda^*$, the optimal solution to the sub-problem in Eq.(7) is also optimal to the original linear fractional problem in Eq.(6). Thus, solving the original problem becomes finding the λ^* satisfying $z(\lambda) = 0$. A bunch of algorithms have been proposed to generate a sequence of λ that can quickly converge to $\lambda^\star$, and the Dinkelbach's algorithm often demonstrates excellent performance. We are not going to elaborate this algorithm in this paper and the readers are referred to [6], [7]. In addition, a recent paper [10] which elegantly develops an optimal feature selection for a set of

trace-ratio-based criteria essentially uses the Dinkelbach's algorithm too. Formulating our feature selection as a 0-1 linear fractional programming problem, Table II describes a globally optimal non-sequential feature selection using the Dinkelbach's algorithm. Introducing such a non-sequential algorithm allows us to take advantage of the well-established literature on linear fractional programming, for example, ample theoretical analysis and smart algorithms, to develop more efficient feature selection algorithms. It is worth noting that we have improved the above Dinkelbach's algorithm to handle feature selection with the presence of redundant features in our recent work [4].

VII. Computational complexity

Our work proposes two different optimal feature selection algorithms for class separability measure. One works in a sequential mode and the other does not. They can be selected for use based on the problems in practice. The optimal SFS performs k iterations, in each of which one of the remaining features is selected. Its computational complexity is $\mathcal{O}(dk)$. According to [7], the total number of iterations of the Dinkelbach's algorithm in the worst case is $\mathcal{O}(\log(dM))$, where $M \geq 1$. In each iteration it finds the top k features having larger $(f_i - \lambda g_i)$ values. Hence, its worst-case computational complexity is about $\mathcal{O}(d\log(dM))$[2]. The Dinkelbach's algorithm is often very efficient in practice and it terminates in a few iterations. In terms of space complexity, both algorithms need store g_i and f_i $(i = 1, 2, \cdots, d)$ and there is no much difference.

VIII. Experimental results

The global optimum of the two proposed algorithms have adequately been justified through the above theoretical analysis. The following experiment only acts as a "sanity check" and an auxiliary demonstration of their global optimum. It is carried out in two ways. First, an exhaustive search is used to find the true optimal feature selection result. The results of the two proposed algorithms are checked to see whether they are identical to the true optimum. The "identical" means that both the Id's of the selected features and the optimal criterion values are same. Because this process depends on an exhaustive search, it can only be applied to a small-sized feature pool, for example, $d = 15$. On larger-sized feature pools, we demonstrate the optimality by verifying that the two proposed selection algorithms always give identical selection results. The logic behind this lies at that the optimality of the two algorithms have been theoretically justified with *different* ways. The consistent coincidence of their selection results can be seen as a strong sign of their optimality. Three data sets are selected from NIPS feature selection challenge[3], where the number of features is no less than 10,000, as shown in Table III.

[2]Here we consider the complexity of each iteration as $\mathcal{O}(d)$ for both algorithms, although the complexity for the non-sequential algorithm is actually marginally larger.

[3]http://www.clopinet.com/isabelle/Projects/NIPS2003/

Table III
THREE NIPS FEATURE SELECTION DATA SETS

Data set	Feature type	Training size	Feature number
ARCENE	Non sparse	100	10,000
DEXTER	Sparse integer	300	20,000
DOROTHEA	Sparse binary	800	100,000

Consistence with the exhaustive search. From each NIPS data set, we select the top 15 features to form a small-sized feature pool. The number of features to be selected, k, is increased from 1 to 14 one by one, and the two proposed algorithms are compared with the exhaustive search. The result is in Table IV. The first part reports the comparison on ARCENE data set. As seen, all of the three algorithms attain the same optimal criterion values (the absolute difference among them is less than 1e-10). Also, the selected features are compared with each other. It is found that on ARCENE and DEXTER the features selected by the three algorithms are completely identical. The features selected on DOROTHEA are different sometimes. We find that this is because different feature subsets may produce same optimal criterion values and they are essentially all "optimal" in terms of the selection criterion. The three algorithms may pick different ones from them. Since the exhaustive search gives true optimal result, this comparison verifies the optimality of the proposed selection algorithms. This well aligns with the previous theoretical analysis.

Consistence between two proposed algorithms. Practical applications often involve thousands or millions of features. The optimality of the proposed algorithms has to be tested in this situation. Since the exhaustive search becomes computationally intractable, we check whether the two proposed algorithms can give an identical result—a necessary condition for both of them to be optimal. On the three NIPS data sets, we gradually increase the number of selected feature from 1 to 100 and compare the obtained optimal criterion values. The results are plotted in Fig.2. The two algorithms are named OSFS (Optimal Sequential Feature Selection) and LFP (Linear Fractional Programming) and their curves are labeled by "+" and "o", respectively. As plotted, on all of the three data sets the two curves are completely overlapped to each other in the whole course. This verifies the consistence of two algorithms and their optimality in some sense.

IX. Future Work

This work mainly focuses on the discussion of the "optimality" of feature selection for class separability measure. For the future work, we will be particularly interested in exploring if these optimal feature selection algorithms can really outperform their suboptimal counterpart (say, the conventional sequential feature selection). For example, if simply pursuing the optimum of a selection criterion will

Table II
GLOBALLY OPTIMAL NON-SEQUENTIAL FEATURE SELECTION USING LINEAR FRACTIONAL PROGRAMMING

Same **Input** and **Output** as in Table I.
Initialization:
compute g_i and f_i $(i = 1, 2, \cdots, d)$ for each feature and store them in memory
Initialize randomly set k components of $\mathbf{w}$ as "1" and the remaining as "0",
Feature selection:
do
(1) **Set** $\lambda = \mathbf{f}^\top \mathbf{w} / \mathbf{g}^\top \mathbf{w}$,
(2) **Compute** $(f_i - \lambda g_i)$ for each of the d features and **find** the k ones with larger values. **Set** the corresponding k components of $\mathbf{w}$ as "1" and others as "0". (This solves the maximization problem in Eq.(7))
(3) If $\mathbf{f}^\top \mathbf{w} - \lambda \mathbf{g}^\top \mathbf{w} < \xi$ (e.g., 10^{-4}), the current $\mathbf{w}$ is optimal and all **done**. Otherwise, go to step (1).
while (1)

Table IV
TEST THE **CONSISTENCE** OF TWO PROPOSED FEATURE SELECTION ALGORITHMS WITH AN EXHAUSTIVE SEARCH

Data set	k=1	2	3	4	5	6	7	8	9	10	11	12	13	14
ARCENE (EXH)	0.1411	0.1378	0.1260	0.1207	0.1125	0.1034	0.0960	0.0896	0.0842	0.0789	0.0732	0.0669	0.0613	0.0557
ARCENE (OSFS)	0.1411	0.1378	0.1260	0.1207	0.1125	0.1034	0.0960	0.0896	0.0842	0.0789	0.0732	0.0669	0.0613	0.0557
ARCENE (LFP)	0.1411	0.1378	0.1260	0.1207	0.1125	0.1034	0.0960	0.0896	0.0842	0.0789	0.0732	0.0669	0.0613	0.0557
DEXTER (EXH)	0.0067	0.0066	0.0063	0.0062	0.0060	0.0058	0.0056	0.0054	0.0052	0.0050	0.0049	0.0048	0.0048	0.0047
DEXTER (OSFS)	0.0067	0.0066	0.0063	0.0062	0.0060	0.0058	0.0056	0.0054	0.0052	0.0050	0.0049	0.0048	0.0048	0.0047
DEXTER (LFP)	0.0067	0.0066	0.0063	0.0062	0.0060	0.0058	0.0056	0.0054	0.0052	0.0050	0.0049	0.0048	0.0048	0.0047
DOROTHEA (EXH)	0.0116	0.0116	0.0116	0.0088	0.0074	0.0057	0.0049	0.0037	0.0029	0.0025	0.0020	0.0016	0.0013	0.0011
DOROTHEA (OSFS)	0.0116	0.0116	0.0116	0.0088	0.0074	0.0057	0.0049	0.0037	0.0029	0.0025	0.0020	0.0016	0.0013	0.0011
DOROTHEA (LFP)	0.0116	0.0116	0.0116	0.0088	0.0074	0.0057	0.0049	0.0037	0.0029	0.0025	0.0020	0.0016	0.0013	0.0011

EXH: exhaustive search; OSFS: optimal sequential forward selection; LFP: the non-sequential selection based on linear fractional programming

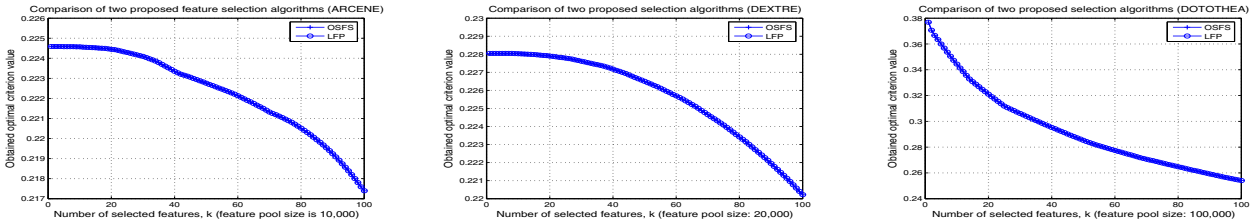

Figure 2. Test the **consistence** of two proposed feature selection algorithms

cause overfitting and adversely affect classification performance? The result in this work can effectively help us study this issue for class separability based feature selection.

REFERENCES

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[2] A. Webb, *Statistical Pattern Recognition (Second Edition)*. John Wiley and Sons, 2002.

[3] L. Wang, "Feature selection with kernel class separability," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1534–1546, 2008.

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

[5] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[6] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, no. 7, pp. 492–498, 1967.

[7] T. Matsui, Y. Saruwatari, and M. Shigeno, "An analysis of dinkelbach's algorithm for 0-1 fractional programming problems," *METR92-14, Dept. of Mathematical Engineering and Information Physics, University of Tokyo*, Dec 1992. [Online]. Available: http://www.misojiro.t.u-tokyo.ac.jp/ tomomi/TRs/TRs.html

[8] N. Megiddo, "Combinatorial optimization with rational objective functions," *Mathematics of Operations Research*, vol. 4, no. 4, pp. 414–424, 1979.

[9] E. Lawler, *Combinatorial Optimization: Networks and Matroids*. Holt, Rinehart, Winston, 1976.

[10] F. Nie, S. Yang, Y. Jia, C. Zhang, and S. Yan, "Trace ratio criterion for feature selection," in *AAAI*, 2008, pp. 671–676.