Hippocampal Shape Classification Using Redundancy Constrained Feature Selection

Luping Zhou[1], Lei Wang[1], Chunhua Shen[2], and Nick Barnes[2]

[1] School of Engineering, The Australian National University
[2] Embedded Systems Theme, National ICT, Australia*

Abstract. Landmark-based 3D hippocampal shape classification involves high-dimensional descriptor space, many noisy and redundant features, and a very small number of training samples. Feature selection becomes critical in this situation, because it not only improves classification performance, but also identifies the regions that contribute more to shape discrimination. This work identifies the drawbacks of SVM-RFE, and proposes a novel class-separability-based feature selection approach to overcome them. We formulate feature selection as a constrained integer optimization and develop a new algorithm to efficiently and optimally solve this problem. Theoretical analysis and experimental study on both synthetic data and real hippocampus data demonstrate its superior performance over the prevailing SVM-RFE. Our work provides a new efficient feature selection tool for hippocampal shape classification.

1 Introduction

Identifying the morphological differences between anatomical shapes related to disorders is important for medical image analysis. However, this is very difficult because the data are often high-dimensional but training samples are scarce. For hippocampal shapes, it is common for the SPHARM-PDM, which represents shapes by corresponded landmarks from parameterized surfaces, to represent a hippocampus with more than 1,000 landmarks. Stacking their coordinates leads to a high-dimensional feature vector. However, the number of training data is commonly around 30-50 only. Even for the advanced classifiers such as the Support Vector Machines (SVMs), the presence of many irrelevant and noisy features can significantly deteriorate learning performance. Feature subset selection becomes a critical step in this situation.

Feature selection has been widely used in medical applications, for example, the well-known SVM-RFE (recursive feature elimination) method [1]. Despite its popularity in feature selection, SVM-RFE has three drawbacks: i) Because SVM maximizes the minimum margin between two groups, SVM-RFE is not robust against noisy data even with soft-margin SVM; ii) SVM-RFE cannot effectively

* National ICT Australia is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australia Research Council. The authors thank the OASIS team and NICTA AASEDP project for providing the data.

T. Jiang et al. (Eds.): MICCAI 2010, Part II, LNCS 6362, pp. 266–273, 2010.

avoid selecting highly correlated discriminative features; and iii) SVM-RFE cannot flexibly deal with group-based feature selection. In landmark-based 3D representation of hippocampus, due to its continuous and overall smooth surface the change within a small area is not drastic. As a result, the coordinates of the landmarks in the area are often strongly correlated. The existence of such feature redundancy causes problems for the k-best feature selection. In the extreme case, if the most discriminative feature is duplicated several times, all of them will be selected and consequently those less discriminative but complementary features may be missed. This could significantly degrade the classification performance. Moreover, to benefit the explanation of the difference between hippocampal groups, the selection of landmarks is needed, that is, to select x, y, z coordinates (the features in the shape descriptor) of the same landmark simultaneously. Such a task may be cumbersome for SVM-RFE that uses the backward sequential selection. Additional criteria need to be imposed to combine the selection of individual coordinates, which might not be a natural extension.

In this paper we propose a new approach to select discriminative features in the hippocampal shape study. To address noisy features, we use the trace-based class separability measure as the feature selection criterion. This criterion has been shown to be robust to the small sample problem and noisy features [2]. However, this criterion cannot identify redundant features either. To overcome this problem, we propose a new redundancy-constrained feature selection (RCFS). The basic idea is to formulate the feature selection problem as a 0-1 linear fractional programming problem and impose extra constraints to avoid selecting redundant features. To achieve efficient feature selection, we study the constraints that maintain the global solvability through the *totally unimodular* (TUM) condition in integer programming, and demonstrate that hierarchically clustering features can generate qualified redundancy constraints. In addition, due to its flexibility of adding linear constraints, RCFS can be easily extended to select the landmark points. Experiments show that the proposed RCFS method significantly outperforms SVM-RFE on the hippocampus data due to its more robust selection criterion, the capability in identifying and removing redundant features, and the flexible extension for landmark selection.

2 Redundancy-Constrained Feature Selection (RCFS)

Let $(\boldsymbol{x}, y) \in (\mathbb{R}^n \times \mathcal{Y})$ be a training sample, where $\mathcal{Y} = \{1, 2, \cdots, s\}$ is the label set. Let l_i be the number of samples in class i, $\boldsymbol{m}_i$ the mean of class i and $\boldsymbol{m}$ the mean of all classes. The within-class, between-class and total scatter matrices are defined as

$$\mathbf{S}_W = \sum_{i=1}^{s} \sum_{j=1}^{l_i} (\boldsymbol{x}_{ij} - \boldsymbol{m}_i)(\boldsymbol{x}_{ij} - \boldsymbol{m}_i)^\top, \quad \mathbf{S}_B = \sum_{i=1}^{s} l_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^\top,$$
$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B = \sum_{i=1}^{s} \sum_{j=1}^{l_i} (\boldsymbol{x}_{ij} - \boldsymbol{m})(\boldsymbol{x}_{ij} - \boldsymbol{m})^\top.$$

When feature dimensionality is much larger than the number of training samples, which is the case of hippocampal shape classification, the scatter matrices are rank-deficient and the determinants become zero. Hence, the trace-based form, $\mathrm{tr}(\mathbf{S}_B)/\mathrm{tr}(\mathbf{S}_T)$, is used in this paper. It is not difficult to show that

$$\mathrm{tr}(\mathbf{S}_B) = \sum_{i=1}^{s} l_i (\boldsymbol{m}_i - \boldsymbol{m})^\top (\boldsymbol{m}_i - \boldsymbol{m}) = \sum_{t=1}^{n} \left(\sum_{i=1}^{s} l_i (m_{it} - m_t)^2 \right) \triangleq \sum_{t=1}^{n} f_t \quad (1)$$

where m_{it} and m_t are the t-th feature of $\boldsymbol{m}_i$ and $\boldsymbol{m}$, respectively. Similarly,

$$\mathrm{tr}(\mathbf{S}_T) = \sum_{t=1}^{n} \left(\sum_{i=1}^{s} \sum_{j=1}^{l_i} (x_{ijt} - m_t)^2 \right) \triangleq \sum_{t=1}^{n} g_t \quad (2)$$

where x_{ijt} is the t-th feature of $\boldsymbol{x}_{ij}$. We have proved in [3] that *if the most discriminative feature t, which has the maximal f_t/g_t, is duplicated k times, feature selection by maximizing* $\mathrm{tr}(\mathbf{S}_B)/\mathrm{tr}(\mathbf{S}_T)$ *will repetitively select it k times.* Similar results exist for sufficiently correlated features.

Basic Problem. To prevent selecting discriminative but mutually redundant features, we propose the redundancy-constrained feature selection (RCFS). Let $\boldsymbol{\omega} \in \{0,1\}^n$ be an n-dimensional binary selector ("1" for being select and "0" for not). Selecting k features can be expressed as finding the optimal $\boldsymbol{\omega}$,

$$\boldsymbol{\omega}^\star = \arg\max_{\boldsymbol{\omega}} \frac{f_1\omega_1 + \cdots + f_n\omega_n}{g_1\omega_1 + \cdots + g_n\omega_n} = \arg\max_{\boldsymbol{\omega}} \frac{\mathbf{f}^\top \boldsymbol{\omega}}{\mathbf{g}^\top \boldsymbol{\omega}} \quad (3)$$

$$\text{subject to } \boldsymbol{\omega} \in \{0,1\}^n, \ \ \boldsymbol{\omega}^\top \mathbf{1} = k, \text{ and } \ \boldsymbol{\omega} \in \boldsymbol{\Omega}.$$

$\boldsymbol{\Omega}$ contains the constraints used to avoid selecting redundant features. With the Dinkelbach's algorithm [4], solving Eq.(3) iteratively solve a subproblem,

$$\begin{aligned} z(\lambda) \triangleq \quad & \max_{\boldsymbol{\omega}} \ (\mathbf{f}^\top \boldsymbol{\omega} - \lambda \mathbf{g}^\top \boldsymbol{\omega}) \\ \text{subject to } \ & \boldsymbol{\omega} \in \{0,1\}^n, \ \boldsymbol{\omega}^\top \mathbf{1} = k, \ \boldsymbol{\omega} \in \boldsymbol{\Omega}. \end{aligned} \quad (4)$$

When $z(\lambda) = 0$, the optimal solution of (4) will be the optimal solution of (3).

Global Solvability. When $\boldsymbol{\omega} \in \{0,1\}^n$, adding $\boldsymbol{\Omega}$ could make Eq.(4) very difficult to solve, even if $\boldsymbol{\Omega}$ only contains linear constraints and (4) becomes an integer linear program (ILP). ILP is much more difficult than LP, and there are no general polynomial-time algorithms. Nevertheless, if satisfying the *totally unimodular* (TUM) condition [5] , an ILP problem will reduce to an LP problem which can be easily solved. Relaxing $\boldsymbol{\omega} \in \{0,1\}^n$ to $[0,1]^n$, Eq.(4) becomes an LP problem with the feasible region defined as

$$R(\boldsymbol{\omega}) = \{\boldsymbol{\omega} : \mathbf{A}\boldsymbol{\omega} \leq \mathbf{b}, \boldsymbol{\omega} \geq 0\}. \quad (5)$$

Geometrically, $R(\boldsymbol{\omega})$ is a polyhedron. According to [5], for each integral vector $\mathbf{b}$, $R(\boldsymbol{\omega})$ is an *integral* polyhedron if and only if the matrix $\mathbf{A}$ is TUM. Because the optimal solution of an LP problem is always at one of the vertices of the polyhedron, the optima of the ILP and LP problems coincide with each other. Hence, to efficiently solve Eq.(4), $\mathbf{A}$ in (5) has to be TUM. A TUM matrix is a matrix with the determinants of all of its square submatrices being $+1$, -1, or 0. It has the following properties. (**P1**): TUM is preserved when permuting rows or columns or taking transpose; (**P2**): TUM is preserved when multiplying

a row or column by -1 or repeating a row or column; (**P3**): If $\mathbf{A}$ is TUM, $[\mathbf{A}\ \mathbf{I}]$ is TUM, where $\mathbf{I}$ is an identity matrix.

Although it is restrictive for $\mathbf{A}$ to be TUM, we show that the constraints obtained by feature clustering gives a qualified $\mathbf{A}$. Let $x_1, x_2, \cdots, x_n$ be the n features of $\boldsymbol{x}$. We define $d(x_i, x_j)$ as the "distance" between x_i and x_j that reflects their independence or complementary. It can be correlation coefficients, mutual information, or any criterion on feature redundancy. We define $d(x_i, x_j) = 1 - |\rho(x_i, x_j)|$, where ρ is Pearson correlation. Let $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_m$ be m clusters, forming a mutually exclusive and complete partition of the n features,

$$\{x_1, x_2, \cdots, x_n\} = \mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_m \ \text{ and } \ \mathcal{C}_i \cap \mathcal{C}_j = \emptyset, \ \ 1 \le i < j \le m. \tag{6}$$

We enforce that at most p_i ($p_i \ge 1$) features can be selected from $\mathcal{C}_i$,

$$\sum_{x_j \in \mathcal{C}_i} \omega_j \le p_i, \quad \forall i = 1, 2, \cdots, m. \tag{7}$$

Let $(x_{r_1}, \cdots, x_{r_n})$ be a rearrangement of $(x_1, \cdots, x_n)$ according to their appearing in $\mathcal{C}_1, \cdots, \mathcal{C}_m$ and this applies to $\boldsymbol{\omega}$ too. Let $\mathbf{I}_{n\times n}$ be an identity matrix and $\mathbf{1}_{1\times c_i}$ be a row vector of 1's. $\mathbf{A}\boldsymbol{\omega} \le \mathbf{b}$ in (5) can be explicitly written as

$$\begin{pmatrix} & \mathbf{1}_{1\times n} & & \\ & -\mathbf{1}_{1\times n} & & \\ -- & -- & -- & -- \\ \mathbf{1}_{1\times c_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{1\times c_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1}_{1\times c_m} \\ -- & -- & -- & -- \\ & \mathbf{I}_{n\times n} & & \end{pmatrix} \begin{pmatrix} \omega_{r1} \\ \omega_{r2} \\ \vdots \\ \vdots \\ \vdots \\ \omega_{rn} \end{pmatrix} \le \begin{pmatrix} k \\ -k \\ -- \\ p_1 \\ \vdots \\ p_m \\ -- \\ \mathbf{1}_{n\times 1} \end{pmatrix}. \tag{8}$$

The middle part of $\mathbf{A}$ is an *interval* matrix. It contains "0" and "1" only and has consecutive 1's in each row. Each interval matrix is TUM [5]. It can be proved that the whole $\mathbf{A}$ is TUM by using (**P1**), (**P2**) and (**P3**). Thus, the subproblem in Eq.(4) can be efficiently solved thanks to the equivalence of ILP and LP.

Constraints Generation. The above method has many algorithmic parameters, including m and $p_1, \cdots, p_m$. Optimally setting them is impractical. We propose *agglomerative hierarchical clustering* to handle it. Starting with the n features, two features (or subclusters) are merged at each level until only k clusters are left, giving a hierarchy of $n - k + 1$ levels. Then, the constrained feature selection is applied to *each* level of this hierarchy with all p_i in (8) being 1. Multi-fold cross-validation is used to identify the best selection from different levels. In doing so, i) we do not need to preset m. Instead, features are clustered at different degrees of redundancy in this hierarchy; ii) we only need to set $p_i = 1$. Because one cluster at a given level is formed by multiple clusters at preceding levels, the case of $p_i > 1$ can be implicitly approximated by a group of $p_j = 1$ in preceding levels; iii) the matrix $\mathbf{I}$ in $\mathbf{A}$ can be ignored; iv) this will not significantly slow down feature selection because only LP problems are solved and the Dinkelbach's algorithm usually terminates in a few iterations.

Table 1. Proposed redundancy-constrained feature selection (RCFS)

Input: l training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$ and the value of k,
Output: optimal binary selector $\boldsymbol{\omega}$ and corresponding k selected features.
Initialization:
hierarchically cluster n features (or 3D points) with correlation coefficient ρ,
establish linear constraints $\boldsymbol{\Omega}$ accordingly,
compute g_i and f_i $(i = 1, 2, \cdots, n)$ for each feature,
initialize k components of $\boldsymbol{\omega}$ as "1" and the remaining as "0",
Feature selection on each level with the Dinkelbach's algorithm:
(1) **Set** $\lambda = \mathbf{f}^\top \boldsymbol{\omega} / \mathbf{g}^\top \boldsymbol{\omega}$,
(2) **Solve** the maximization problem in Eq. (4)
(3) If $\mathbf{f}^\top \boldsymbol{\omega} - \lambda \mathbf{g}^\top \boldsymbol{\omega} < \xi$ (e.g., 10^{-4}), $\boldsymbol{\omega}$ is optimal. Otherwise, go to (1).
Cross-validation is used to identify the best selection from different levels

Landmark Selection. 3D landmark selection for hippocampal shapes is very useful for medical diagnosis and clinical interpretation. The SPHARM-PDM representation of hippocampal surfaces stacks the x, y, z coordinate values of all landmarks as a long vector. A straightforward feature selection chooses the individual coordinates instead of a 3D point as a whole. It is highly likely that, for a point, one of its three coordinates is selected but the other two are not, bringing difficulty in interpreting the selection result. A landmark-based selection is needed, in which the three coordinates of each point are selected (or not selected) together. Our proposed RCFS can handle this case effortlessly by assigning the same ω_i to the three coordinates. It can be shown that the matrix $\mathbf{A}$ is still TUM in this case. In contrast, SVM-RFE, as a backward sequential selection, cannot handle landmark selection naturally. It needs to incorporate additional criteria to evaluate the importance of a landmark as a whole. This is not as seamless as our RCFS.

3 Experiments

Synthetic data. A synthetic data set is used to illustrate the efficacy of RCFS on redundancy removal. Only 2 (x_1 and x_2) out of 52 features are statistically relevant to class labels, whereas the others are noises. x_2 is more discriminative than x_1. Two classes are sampled from $\mathcal{N}((2, 0.25)^\top, \boldsymbol{\Sigma})$ and $\mathcal{N}((2.5, 3)^\top, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (.24\ .38;\ .38\ .81)$. x_1 and x_2 are duplicated with random noise respectively. Assuming that $k = 2$ is known, we test RCFS, SVM-RFE, and the non-constrained feature selection (NCFS) on 30 training and test groups (100 vs. 500 samples). It is found that RCFS successfully selects (x_1, x_2) on 28 groups. In contrast, SVM-RFE only succeeds on 2 groups and keeps selecting x_2 and its duplicate on other groups. NCFS never succeeds and always selects x_2 and its duplicate. With all 54 features, a linear SVM obtains the test error rate $8.01 \pm 2.18\%$. With the 2 features selected by RCFS, SVM-RFE and NCFS, a linear SVM obtains $1.47 \pm 1.48\%$, $5.22 \pm 1.39\%$ and $5.45 \pm 0.82\%$, respectively. As shown, RCFS outperforms both SVM-RFE and NCFS.

Table 2. Comparison of classification *with* and *without* RCFS feature selection

k	RCFS wins (groups)	RCFS loses (groups)	Mean (test errors %)		p-value (one tailed)
			RCFS	Use all features	(paired t-test)
2500	18	8	38.11	39.31	0.1421
2000	14	9	38.93	39.31	0.3533
1500	17	12	36.98	39.31	**0.0189**
1000	18	8	**36.42**	39.31	**0.0022**
500	18	11	37.23	39.31	0.0687
$\sum$	84	48	-		

Hippocampi in OASIS. We apply our RCFS method to improving the discrimination of hippocampal shapes between AD and the normal control. Subjects aged from 60 to 79 in the OASIS data set (http://www.oasis-brains.org/) are used. We categorize subjects with a non-zero CDR rating into the AD group and the rest into normal control. There are 103 samples for the left and right hippocampi respectively. Each shape is represented by x, y, z coordinates of 1002 landmarks (3006 features in total) obtained from SPHARM-PDM representation with degree 15. Experimental results are reported only for the left hippocampi[1]. Samples are randomly partitioned into 30 training and test (50 vs. 53 samples) groups. With all 3006 features used, a linear SVM attains an average error rate of 39.31%. Due to the complexity of data and the scarcity of training samples, the test error rates of different groups vary significantly: from 26% to 55%. This inter-group variation may hide the true difference between different methods. To give a fair and accurate evaluation, we report the number of groups on which RCFS wins or loses in addition to the average test error rates. More importantly, we conduct a paired t-test to test the statistical difference between two methods. By pairing the test error rates, each time the two methods are compared on the same data set, which mitigates the influence of the inter-group variation.

The paired t-test is first used to detect the statistical difference between the test error rates from a linear SVM using the RCFS-selected features and all 3006 features, respectively. As shown in Table 2, significant difference is detected at the level of 0.05 on 30 test groups for $k = 1000$ and 1500, at the level of 0.1 for $k = 500$. This verifies that when a suitable number of features are selected, employing RCFS can significantly improve classification accuracy. For example, using only 1/3 of the original features can reduce the average test error rate from 39.31% to 36.42%.

The paired t-test is then used to detect the statistical difference between the test error rates from a linear SVM using the features selected by RCFS and SVM-RFE, respectively. As shown in Table 3, RCFS wins much more often than SVM-RFE does. The lowest average error rate 36.42% is achieved by RCFS when

[1] Results for the right hippocampi (with higher classification accuracy than the left) are omitted here due to the limit of pages. The hypothesis test shows that the performance of RCFS statistically equals that of SVM-RFE on the right hippocampi.

Table 3. Performance comparison of RCFS and SVM-RFE

k	RCFS wins (groups)	RCFS loses (groups)	Mean (test errors %)		p-value (one tailed)
			RCFS	SVM-RFE	(paired t-test)
2500	14	8	38.11	39.06	0.1218
2000	18	9	38.93	39.31	0.3710
1500	17	8	36.98	38.99	**0.0276**
1000	18	7	**36.42**	39.75	**0.0001**
500	15	9	37.23	38.87	0.0958
$\sum$	82	41	-		

$k = 1000$, as shown in bold. More importantly, the paired t-test indicates that, RCFS and SVM-RFE are significantly different at the level of 0.001 on the 30 test groups when $k = 1000$, at the level of 0.05 when $k = 1500$, and at the level of 0.1 When $k = 500$. It can be expected that the improvement of RCFS over SVM-RFE becomes less obvious when feature selection gains little from selecting too many or too few features. Even though, RCFS has never performed worse than SVM-RFE, in terms of number of wins and average test error.

Discriminative landmark selection. The following shows 3D landmark selection by RCFS, and the visual explanation of the obtained shape difference. Note that SVM-RFE cannot automatically deal with this problem. The landmark selection is conducted by selecting $k = 250$ and $k = 125$ landmarks respectively on 30 training and test groups for both left and right hippocampi. For $k = 250$, a linear SVM obtains the lowest test error rate 26.42% (left) and 24.53% (right) among the 30 test groups. For $k = 125$, the two lowest error rates

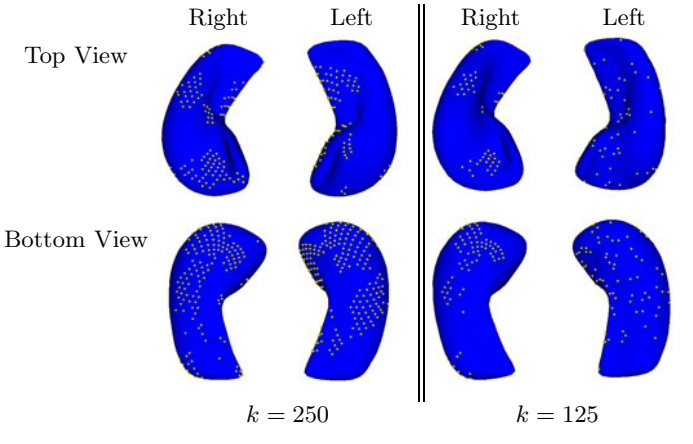

Fig. 1. Discriminative landmarks are selected in cases of $k = 250$ (left) and $k = 125$ (right) respectively. The selected landmarks are overlaid as the yellow balls on the mean shapes of the left and right hippocampi.

Table 4. Comparison of test error rates (%)of RCFS and NCFS for landmark selection

Test Error Rate (%)	left		right	
	$k = 250$	$k = 125$	$k = 250$	$k = 125$
NCFS	30.19	33.96	22.64	26.42
RCFS	26.42	26.42	24.53	22.64

becomes 26.42% and 22.64% respectively. The selected landmarks are overlaid on the mean shapes of the left and right hippocampi respectively, as shown in Fig. 1, to reveal the essential shape discrimination. By cross-referencing the results of $k = 250$ and $k = 125$, we can see that the majority of the identified differences locate in CA1 and subiculum surface zones, especially for the inferior part (bottom view). This observation agrees with some findings in the literature [6]. The sparsity of the selected landmarks is automatically determined by the RCFS algorithm. For example, the selected 125 landmarks of the left hippocampi are very sparse, while those in other cases are visually more gathered. However, as shown in Table 4, compared with NCFS where no redundancy constraints are imposed, RCFS achieves clearly better classification performance, except for the right hippocampi when 250 landmarks are selected. This demonstrates the advantage of RCFS.

4 Conclusion

SVM-RFE has been a fairly standard feature selection method used in many research fields. In this paper, we propose a constrained feature selection method that shows superior selection performance over SVM-RFE when noisy and redundant features exist. We apply it to identifying essential hippocampal shape difference between AD and the control. The proposed method can be efficiently solved as we carefully design the constraints and preserve its global solvability.

References

1. Vemuri, P., Gunter, J., et al.: Alzheimer's disease diagnosis in individual subjects using structural mr images: Validation studies. Neuroimage 39, 1186–1197 (2008)
2. Wang, L.: Feature selection with kernel class separability. IEEE Trans. Pattern Analysis and Machine Intelligence 30(9), 1534–1546 (2008)
3. Zhou, L., Wang, L., Shen, C.,: Feature selection with redundancy-constrained class separability. IEEE Trans. Neural. Networks 21(5), 853–858 (2010)
4. Dinkelbach, W.: On nonlinear fractional programming. Management Science 13(7) (1967)
5. Schrijver, A.: Theory of Linear and Integer Programming. John Wiley and Sons, Chichester (1986)
6. Csernansky, J., Wang, L., Swank, J., Miller, J., Gado, M., McKeel, D., Miller, M., Morris, J.: Preclinical detection of Alzheimer's disease: hippocampal shape and volume predict dementia onset in the elderly. Neuroimage 25, 783–792 (2005)