

# DNA Copy Number Selection Using Robust Structured Sparsity-Inducing Norms

Vangelis Metsis, Fillia Makedon, Dinggang Shen, and Heng Huang

**Abstract**—Array comparative genomic hybridization (aCGH) is a newly introduced method for the detection of copy number abnormalities associated with human diseases with special focus on cancer. Specific patterns in DNA copy number variations (CNVs) can be associated with certain disease types and can facilitate prognosis and progress monitoring of the disease. Machine learning techniques have been used to model the problem of tissue typing as a classification problem. Feature selection is an important part of the classification process, because many biological features are not related to the diseases and confuse the classification tasks. Multiple feature selection methods have been proposed in the different domains where classification has been applied. In this work, we will present a new feature selection method based on structured sparsity-inducing norms to identify the informative aCGH biomarkers which can help us classify different disease subtypes. To validate the performance of the proposed method, we experimentally compare it with existing feature selection methods on four publicly available aCGH datasets. In all empirical results, the proposed sparse learning based feature selection method consistently outperforms other related approaches. More important, we carefully investigate the aCGH biomarkers selected by our method, and the biological evidences in literature strongly support our results.

**Index Terms**—Feature evaluation and selection, biomarker detection, aCGH, DNA copy number variations, cancer classification.



## 1 INTRODUCTION

CHROMOSOMAL aberrations occur in many diseases. For example, in cancer, increases or decreases in DNA copy number can alter the expression levels of tumor suppressor genes and oncogenes resulting in tumor genesis. Array comparative genomic hybridization (aCGH) is a recently introduced technique for identifying chromosomal aberrations in human diseases throughout the human genome. aCGH can be used for detection and mapping of copy number abnormalities which can be associated with certain disease phenotypes. This, in turn, can facilitate the localization of critical genes related to specific diseases which can be used as biomarkers for disease diagnosis, prognosis and response to therapy [18], [30].

Machine Learning techniques can be used to discover patterns in DNA copy number variations associated with certain diseases. A set of chromosomal aberrations occurring consistently when a certain disease is observed can indicate that there is correlation between those aberrations and the observed disease. Such patterns have been utilized by researchers [1], [7], [9], [12], [17], [18], [28], [29], [31] for cancer detection and typing. In the general case, the task to accomplish is the classification of tissue samples as cancerous or non-cancerous, and

extensively their classification to a specific cancer type.

In the setting of supervised learning, the copy number changes of particular locations (probes) of the genome are used as features for training and classification. In general, the number of probes of a high-resolution CGH can span from hundreds to thousands. On the contrary, only a few genes are associated with most diseases. Moreover, the number of available samples to be used for training is usually only a few dozens. To reduce noise and avoid over-fitting a feature selection step is necessary before training and classification. An extra advantage of the feature selection process is that the majority of the irrelevant features are discarded and the few remaining can be indicators of possible biomarkers related to the observed disease.

Feature selection has already been shown to significantly benefit the classification accuracy of aCGH data [9], [12], [22]. Thus, it is important to design effective feature selection method for identifying the DNA copy number biomarkers. Previous works have shown the superior performance of sparsity regularization in dimensionality reduction and feature selection [4], [24]. The  $\ell_1$ -norm based sparse regularization term was used with regression or SVM models to perform feature selection by shrinking the coefficients of the irrelevant features to zero [34]. Such sparsity norms impose flat sparsity to shrink the coefficients without using structure information. Later the structured group sparsity was introduced to select group-wise features [45]. The sparse group Lasso model was proposed to combine both group  $\ell_1$ -norm regularization and  $\ell_1$ -norm regularization [11]. Both later methods [11], [45] group the features/variables using specific structures. Considering the multi-task or multi-class structures, the structured

- V. Metsis, F. Makedon and H. Huang are with the Department of Computer Science and Engineering, University of Texas at Arlington Arlington, TX 76019, USA.
- D. Shen is with the Department of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599, USA.
- Corresponding author: heng@uta.edu.

sparsity-inducing norm  $\ell_{2,1}$ -norm was proposed for feature selection by [2], [26]. Later Nie *et. al.* [25] proposed the use of joint  $\ell_{2,1}$ -norm minimization on both loss function and regularization. As a mixed norm, the  $\ell_{2,1}$ -norm couples feature selection across tasks, *i.e.* group the coefficients of the same feature cross all tasks/classes and impose the  $\ell_1$ -norm between all features. Thus, the  $\ell_{2,1}$ -norm is performed to select features across all tasks/classes with joint sparsity. That means that each feature has small scores for all or has large scores over all tasks/classes. This is different to group Lasso and sparse group Lasso, which group different features using  $\ell_2$ -norm. If we reduce the  $\ell_{2,1}$ -norm from matrix format to vector format, it becomes the flat  $\ell_1$ -norm. In this work, we introduce our newly developed feature selection method based on structured sparse regularization that produces higher accuracy compared to the methods that have been previously tested on aCGH data. Our method was inspired by the  $\ell_{2,1}$ -norm based multi-task learning and feature selection [2], [26].

To effectively select the DNA copy number biomarkers, we propose a hybrid regularization method which uses two separate regularization terms involving  $\ell_{2,1}$ -norm and  $\ell_1$ -norm. This is particularly important in multi-class classification problems which contain a big number of classes because a feature, for example, that is important for one class but not important for all others get a low total score (coefficient) to be lost in the feature selection process. Our method ensures that such features will at least get a high coefficient value for the classes that they are important to and have more chances to be included in the final set of selected features. Each regularization term is assigned a different weight according to the specifics of the dataset. More important, we introduce a new efficient optimization algorithm to solve the proposed objective with rigorously proved global convergence.

To test the performance of our proposed method we conducted experiments on four different, publicly available aCGH datasets. We compare with other methods that have been recently proposed for feature selection on aCGH biomarkers and present the classification accuracy results using SVM [8] and Logistic Regression [5], [16] as classifiers. In all empirical results, our hybrid structured sparse learning model consistently outperforms other related feature selection approaches. More important, we carefully investigate the aCGH biomarkers selected by our method, and the biological evidences in literature strongly support our results.

The remainder of this article is organized as follows. Section 2 introduces the theory behind the proposed feature selection method. In section 3 we describe the datasets we used in our experiments. Our experimental results are presented in Section 4. Finally, Section 5 gives the conclusions of our findings.

**Notations:** We summarize the notations and the definition of norms used in this paper. Matrices are written as uppercase letters and vectors are written as lowercase

letters. For matrix  $W = \{w_{ij}\}$ , its  $i$ -th row,  $j$ -th column are denoted as  $w^i$ ,  $w_j$ , respectively. The  $\ell_p$ -norm of the vector  $v \in \mathbb{R}^n$  is defined as  $\|v\|_p = \left(\sum_{i=1}^n |v_i|^p\right)^{\frac{1}{p}}$  for  $p > 0$ . The Frobenius norm of the matrix  $W \in \mathbb{R}^{d \times m}$  is defined as  $\|W\|_F = \sqrt{\sum_{i=1}^d \sum_{j=1}^m w_{ij}^2} = \sqrt{\sum_{i=1}^d \|w^i\|_2^2}$ . The  $\ell_{2,1}$ -norm of matrix  $W$  is defined as  $\|W\|_{2,1} = \sum_{i=1}^d \|w^i\|_2$  (in other related papers, people also used the notation  $\ell_1/\ell_2$ -norm).  $\ell_{1,1}$ -norm of matrix  $W$  is defined as  $\|W\|_{1,1} = \sum_{i=1}^d \sum_{j=1}^m |w_{ij}|$ . The  $\ell_{r,p}$ -norm of the matrix  $W \in \mathbb{R}^{n \times m}$  is defined as

$$\|W\|_{r,p} = \left( \sum_{i=1}^n \left( \sum_{j=1}^m |W_{ij}|^r \right)^{\frac{p}{r}} \right)^{\frac{1}{p}} = \left( \sum_{i=1}^n \|w^i\|_r^p \right)^{\frac{1}{p}}. \quad (1)$$

At last,  $Tr(W)$  means the trace operation for matrix  $W$ .

## 2 FEATURE SELECTION VIA ROBUST STRUCTURED SPARSITY-INDUCING NORMS

### 2.1 Hybrid Structured Sparsity Regularization

Feature selection methods can be divided into filter method [22], [27], [41], [43], wrapper method [19], and embedded method [2], [34], [37], [45]. Wrapper methods utilize the learning machine of interest as a black box to select the subset of features that give the best predictive accuracy, and usually have good performance but high computational cost. Filter methods select features based on discriminant criteria that rely on the characteristics of data, independent of any classification algorithm. Filter methods are limited in scoring the predictive power of combined features, and thus have shown to be less powerful in predictive accuracy as compared to wrapper methods, whereas wrapper methods are much slower and cannot be efficiently applied to large datasets. Embedded methods perform feature selection as part of the training process and are usually specific to given learning machines [14]. The embedded methods combine the advantages of both wrapper and filter methods, *i.e.* good performance and low computational cost.

In this work, to effectively identify the aCGH biomarkers, we will introduce an embedded feature selection method based on least square regression with  $\ell_2$ -norm minimization on the loss function and hybrid structured sparsity-inducing norms  $\ell_{2,1}$ -norm and  $\ell_1$ -norm regularization terms. The least square regression has been widely used for learning tasks, due to its simple loss function and fast optimization procedure. Because we are targeting to biomarkers selection, not final classification accuracy, it is efficient to use the least square regression as loss function.

Given  $n$  training data  $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$  with  $d$  features and the associated class labels  $Y =$

$[y_1, y_2, \dots, y_c] = [z_1, z_2, \dots, z_n]^T \in \mathbb{R}^{n \times c}$  ( $z_i$  is the class label vector for data point  $x_i$ ), traditional least square regression solves the following optimization problem to obtain the projection matrix  $W \in \mathbb{R}^{d \times c}$  and the bias vector  $b \in \mathbb{R}^c$ :

$$\min_{W, b} \sum_{i=1}^n \|W^T x_i + b - z_i\|_2^2. \quad (2)$$

For simplicity, the bias  $b$  can be absorbed into  $W$  when the constant value 1 is added as an additional dimension for each data point  $x_i$ , ( $1 \leq i \leq n$ ). Thus the problem becomes:

$$\min_W \sum_{i=1}^n \|W^T x_i - z_i\|_2^2. \quad (3)$$

To control variance and prevent overfitting, researchers usually add one or more regularization terms to the above equation as:

$$\min_W \sum_{i=1}^n \|W^T x_i - z_i\|_2^2 + R(W), \quad (4)$$

where  $R(W)$  is the regularization term and has several options, such as:

$$R_1(W) = \|W\|^2, \quad R_2(W) = \sum_{j=1}^c \|w_j\|_1, \\ R_3(W) = \sum_{i=1}^d \|w^i\|_2^0, \quad R_4(W) = \sum_{i=1}^d \|w^i\|_2.$$

$R_1(W)$  is the ridge regularization which suffers from the existence of outliers in the dataset due to high variance. Moreover, the ridge regularization was not designed for feature selection, hence it is not suitable to be used for biomarker selection. The three other regularization terms are based on non-smooth sparsity-inducing norms.

$R_2(W)$  is the LASSO regularization term which has the desired property of giving different weights to a feature across different classes  $c$  but produces very sparse solutions, especially when the number of samples is small. The  $\ell_1$ -norm regularization term imposes the flat sparsity, and its optimization techniques include LARS [10], linear gradient search [20], and proximal methods [3].  $R_3(W)$  is the  $\ell_{2,0}$ -norm regularization term, and  $R_4(W)$  is the  $\ell_{2,1}$ -norm regularization term. They impose structured sparsity by penalizing all  $c$  regression coefficients corresponding to a single feature as a whole. Thus, the important biomarkers will have large weights in  $W$  for all/most classes, *i.e.* the important gene  $i$  will have large value on  $\|w^i\|_2$ .

There are some other similar structured sparsity regularization, such as group features/covariates detection [26], [38], [39], [45], joint vector sparsity [33], hierarchical group features [46], *etc.* In other communities, the structured sparsity is also called block sparsity [32]. The structured sparsity learning problems can be efficiently solved by methods in [21], [23], [25], [40].

Although the  $\ell_0$ -norm of  $R_3(W)$  is the most desirable [23], it is an NP-hard problem. Thus, in this paper, we use  $R_4(W)$  based regularization, which is a good approximation of the  $\ell_{2,0}$ -norm regularization. The  $\ell_{2,1}$ -norm regularization can help us select the important biomarkers, which are discriminative to all/most classes. However, some important biomarkers may only be discriminative to a small number of classes, not all of them. Such biomarkers may not be selected by the  $\ell_{2,1}$ -norm regularization, because their weights are shrunk to small values by the  $\ell_1$  additions along the feature direction in  $R_4(W)$  definition. To address this problem, we use the hybrid sparsity-inducing norms with adding one more  $\ell_1$ -norm regularization term, and the new objective is:

$$\min_W J(W) = \sum_{i=1}^n \|W^T x_i - z_i\|^2 + \gamma_1 R_2(W) + \gamma_2 R_4(W), \quad (5)$$

or

$$\min_W J(W) = \|X^T W - Y\|^2 + \gamma_1 \|W\|_{1,1} + \gamma_2 \|W\|_{2,1}, \quad (6)$$

where the  $\ell_{2,1}$ -norm regularization imposes the structured sparsity, and the  $\ell_1$ -norm regularization allows the features to have large weights for some classes, not all of them. Thus, the learned weights of features will more precisely show their discriminative abilities.

Although solving this problem seems difficult as both regularization terms are non-smooth, we will show in the next section that our objective can be efficiently solved. For short we will call this objective function HSSL (Hybrid Structured Sparse Learning model). The optimal value of the parameters  $\gamma_1$  and  $\gamma_2$  can be determined experimentally from the dataset. The resulting values in the projection matrix  $W$  will determine the optimal coefficient values for each attribute  $x_i$ . To select the best  $k$  features we can just sort the features by decreasing coefficient value and keep the top  $k$  of them. Figure 1 shows a visualization of the coefficient table  $W$  after the application of HSSL feature selection method on aCGH dataset 3 (please see section 3). In the visualized gray-scale heat-map, each row represents a class and each column represents a feature. The gray-scale color of each square represents the calculated coefficient value of the feature for the corresponding class. Lighter color means the coefficient has a positive value, darker color means negative coefficient value, and gray color means a value close to 0. Large absolute values for each square indicate strong correlation for the corresponding feature-class pair. The overall importance of each feature is measured by calculating the sum of the absolute values of the feature for all classes. In the figure, the features are sorted from left to right by the total importance values.

## 2.2 An Efficient Algorithm to Solve HSSL Model

Although our objective function is convex, it is difficult to be solved, because both regularization terms are non-smooth. It was generally felt that the  $\ell_{2,1}$ -norm minimization problem is much more difficult to solve than



Fig. 1. Visualization of the coefficient table  $W$  after the application of HSSL feature selection method on aCGH dataset 3. Each row represents a class, each column represents a feature. The grayscale color of each square represents the final coefficient value of the feature for the corresponding class. Lighter color means the coefficient has a positive value, darker color means negative coefficient value, and gray color means a value close to 0. The features are sorted from left to right by total importance value.

the  $\ell_1$ -norm minimization problem. Existing algorithms usually reformulate it as a second-order cone programming (SOCP) or semidefinite programming (SDP) problem, which can be solved by interior point method or the bundle method. However, solving SOCP or SDP is computationally expensive, which limits their use in practice. Here, we propose an efficient iterative algorithm to solve our objective function in Eq. (6).

The Eq. (6) can be written as:

$$\min_W \text{Tr}(X^T W - Y)^T (X^T W - Y) + \gamma_1 \|W\|_{1,1} + \gamma_2 \|W\|_{2,1}. \quad (7)$$

Taking the derivative w.r.t  $w_i$  ( $1 \leq i \leq c$ ), and setting it to zero, we have

$$X X^T w_i - X y_i + \gamma_1 D_i w_i + \gamma_2 \tilde{D} w_i = 0, \quad (8)$$

where  $D_i$  ( $1 \leq i \leq c$ ) is a diagonal matrix with the  $k$ -th diagonal element as  $\frac{1}{2|w_{ki}|}$ :

$$D_i = \begin{bmatrix} \frac{1}{2|w_{1i}|} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{2|w_{di}|} \end{bmatrix}, \quad (9)$$

and  $\tilde{D}$  is a diagonal matrix with the  $k$ -th diagonal element as  $\frac{1}{2\|w^k\|_2}$ :

$$\tilde{D} = \begin{bmatrix} \frac{1}{2\|w^1\|_2} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{2\|w^d\|_2} \end{bmatrix}. \quad (10)$$

Thus, we can get  $w_i$  as:

$$w_i = (X X^T + \gamma_1 D_i + \gamma_2 \tilde{D})^{-1} X y_i. \quad (11)$$

Note that  $D_i$  and  $\tilde{D}$  depend on  $W$ , and are also unknown variables. Therefore, we can iteratively solve them. First, we randomly initialize  $W$ . Second, we calculate  $D_i$  and  $\tilde{D}$ . Third, we compute  $w_i$  by Eq. (11). We will repeat the Second and Third steps till the result converges. Our algorithm is described in Algorithm 1.

When  $|w_{ki}| = 0$  or  $\|w^i\|_2 = 0$ , Eq. (6) is not differentiable. Following [13], we can introduce a small perturbation to regularize the  $k$ -th diagonal element of  $D_i$  as  $\frac{1}{2\sqrt{w_{ki}^2 + \zeta}}$ . Similarly, when  $\|w^i\|_2 = 0$ , the  $i$ -th

diagonal element of  $\tilde{D}$  can be regularized as  $\frac{1}{2\sqrt{\|w^i\|_2^2 + \zeta}}$ . Then it can be verified that the derived algorithm minimizes the following problem:  $\sum_{i=1}^n \|W^T x_i - y_i\|^2 + \gamma_1 \sum_{i=1}^c \sum_{k=1}^d \sqrt{w_{ki}^2 + \zeta} + \gamma_2 \sum_{i=1}^d \sqrt{\|w^i\|_2^2 + \zeta}$ , which is apparently reduced to problem Eq. (6) when  $\zeta \rightarrow 0$ .

---

### Algorithm 1: Algorithm

---

**Input:**  $X, Y$

Initialize  $W^{(1)} \in \mathbb{R}^{d \times c}$ ,  $t = 1$ ;

**while** not converge **do**

1. Calculate the diagonal matrices  $D_i^{(t)}$  ( $1 \leq i \leq c$ ) and  $\tilde{D}^{(t)}$ , where the  $k$ -th diagonal element of  $D_i^{(t)}$  is  $\frac{1}{2|w_{ki}^{(t)}|}$  as Eq. (9), and the  $k$ -th diagonal element of  $\tilde{D}^{(t)}$  is  $\frac{1}{2\|(w^{(t)})^k\|_2}$  as Eq. (10);
2. For each  $i$  ( $1 \leq i \leq c$ ),  $w_i^{(t+1)} = (X X^T + \gamma_1 D_i^{(t)} + \gamma_2 \tilde{D}^{(t)})^{-1} X y_i$ ;
3.  $t = t + 1$ ;

**Output:**  $W^{(t)} \in \mathbb{R}^{d \times c}$ .

---

### 2.3 Algorithm Analysis

We will prove that the above algorithm converges to the global optimum.

*Lemma 1:*  $\|w\|_2 - \frac{\|w\|_2^2}{2\|w_0\|_2} \leq \|w_0\|_2 - \frac{\|w_0\|_2^2}{2\|w_0\|_2}$

**Proof:** Obviously,  $-(\|w\|_2 - \|w_0\|_2)^2 \leq 0$ , thus we have

$$-(\|w\|_2 - \|w_0\|_2)^2 \leq 0 \quad (12)$$

$$\Rightarrow 2\|w\|_2 \|w_0\|_2 - \|w\|_2^2 \leq \|w_0\|_2^2 \quad (13)$$

$$\Rightarrow \|w\|_2 - \frac{\|w\|_2^2}{2\|w_0\|_2} \leq \|w_0\|_2 - \frac{\|w_0\|_2^2}{2\|w_0\|_2} \quad (14)$$

which completes the proof.  $\square$

*Theorem 1:* The Algorithm 1 decreases the objective value in each iteration till converges.

**Proof:** To prove this theorem, we will compare the object function values in the iterations  $t + 1$  and  $t$ . According to Step 2 in the algorithm, we have

$$\begin{aligned} W^{(t+1)} = & \min_W \text{Tr}(X^T W - Y)^T (X^T W - Y) \\ & + \gamma_1 \sum_{i=1}^c w_i^T D_i^{(t)} w_i + \gamma_2 \text{Tr}(W^T \tilde{D}^{(t)} W), \end{aligned} \quad (15)$$

therefore we have:

$$\begin{aligned} & \text{Tr}(X^T W^{(t+1)} - Y)^T (X^T W^{(t+1)} - Y) + \\ & \gamma_1 \sum_{i=1}^c (w_i^{(t+1)})^T D_i^{(t)} w_i^{(t+1)} + \gamma_2 \text{Tr}(W^{(t+1)T} \tilde{D}^{(t)} W^{(t+1)}) \\ & \leq \text{Tr}(X^T W^{(t)} - Y)^T (X^T W^{(t)} - Y) + \\ & \gamma_1 \sum_{i=1}^c (w_i^{(t)})^T D_i^{(t)} w_i^{(t)} + \gamma_2 \text{Tr}(W^{(t)T} \tilde{D}^{(t)} W^{(t)}) \end{aligned}$$

Based on the definitions of  $W$ ,  $D_i$ , and  $\tilde{D}$ , we can re-write the above inequality as:

$$\begin{aligned} & Tr(X^T W^{(t+1)} - Y)^T (X^T W^{(t+1)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c \frac{(w_{ij}^{(t+1)})^2}{2 |w_{ij}^{(t)}|} + \gamma_2 \sum_{k=1}^d \frac{\|(w^{(t+1)})^k\|_2^2}{2 \|(w^{(t)})^k\|_2} \\ \leq & Tr(X^T W^{(t)} - Y)^T (X^T W^{(t)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c \frac{(w_{ij}^{(t)})^2}{2 |w_{ij}^{(t)}|} + \gamma_2 \sum_{k=1}^d \frac{\|(w^{(t)})^k\|_2^2}{2 \|(w^{(t)})^k\|_2} \end{aligned}$$

We further re-write the inequality as:

$$\begin{aligned} & Tr(X^T W^{(t+1)} - Y)^T (X^T W^{(t+1)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c \left( \frac{(w_{ij}^{(t+1)})^2}{2 |w_{ij}^{(t)}|} - |w_{ij}^{(t+1)}| + |w_{ij}^{(t+1)}| \right) + \\ & \gamma_2 \sum_{k=1}^d \left( \frac{\|(w^{(t+1)})^k\|_2^2}{2 \|(w^{(t)})^k\|_2} - \|(w^{(t+1)})^k\|_2 + \|(w^{(t+1)})^k\|_2 \right) \\ \leq & Tr(X^T W^{(t)} - Y)^T (X^T W^{(t)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c \left( |w_{ij}^{(t)}| + \frac{(w_{ij}^{(t)})^2}{2 |w_{ij}^{(t)}|} - |w_{ij}^{(t)}| \right) + \\ & \gamma_2 \sum_{k=1}^d \left( \|(w^{(t)})^k\|_2 + \frac{\|(w^{(t)})^k\|_2^2}{2 \|(w^{(t)})^k\|_2} - \|(w^{(t)})^k\|_2 \right), \end{aligned}$$

and

$$\begin{aligned} & Tr(X^T W^{(t+1)} - Y)^T (X^T W^{(t+1)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c \left( |w_{ij}^{(t+1)}| \right) + \gamma_2 \sum_{k=1}^d \left( \|(w^{(t+1)})^k\|_2 \right) \\ \leq & Tr(X^T W^{(t)} - Y)^T (X^T W^{(t)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c \left( |w_{ij}^{(t)}| \right) + \gamma_2 \sum_{k=1}^d \left( \|(w^{(t)})^k\|_2 \right) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c \left( \left( \frac{(w_{ij}^{(t)})^2}{2 |w_{ij}^{(t)}|} - |w_{ij}^{(t)}| \right) - \right. \\ & \left. \left( \frac{(w_{ij}^{(t+1)})^2}{2 |w_{ij}^{(t)}|} - |w_{ij}^{(t+1)}| \right) \right) + \\ & \gamma_2 \sum_{k=1}^d \left( \left( \frac{\|(w^{(t)})^k\|_2^2}{2 \|(w^{(t)})^k\|_2} - \|(w^{(t)})^k\|_2 \right) - \right. \\ & \left. \left( \frac{\|(w^{(t+1)})^k\|_2^2}{2 \|(w^{(t)})^k\|_2} - \|(w^{(t+1)})^k\|_2 \right) \right) \end{aligned} \quad (16)$$

Applying the Lemma 1 to the above inequality, the last two items on the right hand side are less than zero.

Thus, we have:

$$\begin{aligned} & Tr(X^T W^{(t+1)} - Y)^T (X^T W^{(t+1)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c |w_{ij}^{(t+1)}| + \gamma_2 \sum_{k=1}^d \|(w^{(t+1)})^k\|_2 \\ \leq & Tr(X^T W^{(t)} - Y)^T (X^T W^{(t)} - Y) + \\ & \gamma_1 \sum_{i=1}^d \sum_{j=1}^c |w_{ij}^{(t)}| + \gamma_2 \sum_{k=1}^d \|(w^{(t)})^k\|_2 \end{aligned} \quad (17)$$

We can write the results into matrix formulations as:

$$\begin{aligned} & Tr(X^T W^{(t+1)} - Y)^T (X^T W^{(t+1)} - Y) + \\ & \gamma_1 \|W^{(t+1)}\|_{1,1} + \gamma_2 \|W^{(t+1)}\|_{2,1} \\ \leq & Tr(X^T W^{(t)} - Y)^T (X^T W^{(t)} - Y) + \\ & \gamma_1 \|W^{(t)}\|_{1,1} + \gamma_2 \|W^{(t)}\|_{2,1}. \end{aligned} \quad (18)$$

Therefore, the algorithm decreases the objective value in each iteration.  $\square$

In the convergence,  $W^{(t)}$ ,  $D_i^{(t)}$  ( $1 \leq i \leq c$ ) and  $\tilde{D}^{(t)}$  will satisfy the Eq. (8). As the problem (7) is a convex problem, satisfying the Eq. (8) indicates that  $W$  is a global optimum solution to the problem (7). Therefore, the Algorithm 1 will converge to the global optimum of the problem (7). Because we have closed form solution in each iteration, our algorithm converges very fast.

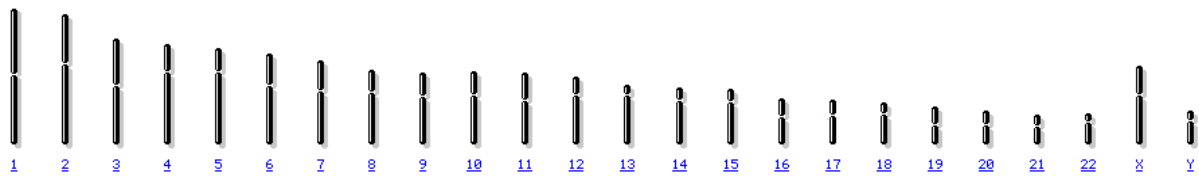
### 3 DATA DESCRIPTIONS

In order to assess the performance of our proposed method for feature selection in aCGH data, we conducted extensive classification experiments where we compared our method with other state-of-the-art feature selection methods that have been recently proposed for aCGH feature selection. For our experiments we used 4 different publicly available aCGH datasets.

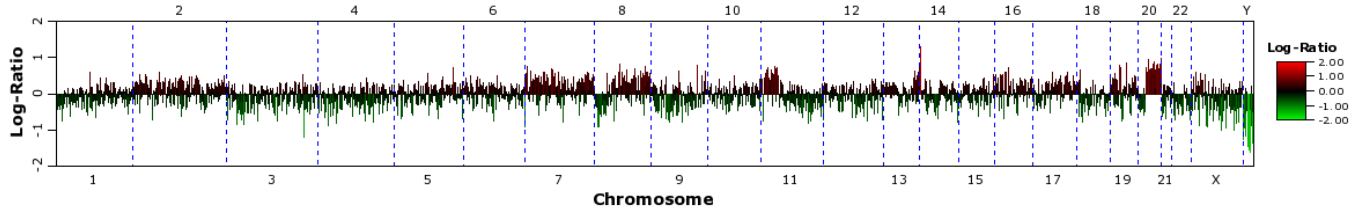
**Dataset 1:** The first dataset contains a total of 75 samples coming from subjects with oral squamous cell carcinoma (SCC) (14 TP53 mutant samples) and healthy subjects (61 wildtype samples). The dataset is available as part of the supplementary material of the publication [42]. Each CGH sample consists of 1979 probes.

**Dataset 2:** The second dataset has been collected by the authors of [44] to investigate the biological basis between aging and sporadic breast cancer incidence and prognosis. DNA samples from matched ER+ invasive breast cancers diagnosed in either young (<45) or old (>70) women were analyzed with aCGH. The datasets consists of 71 samples, 27 of them coming from young women and 44 from old women.

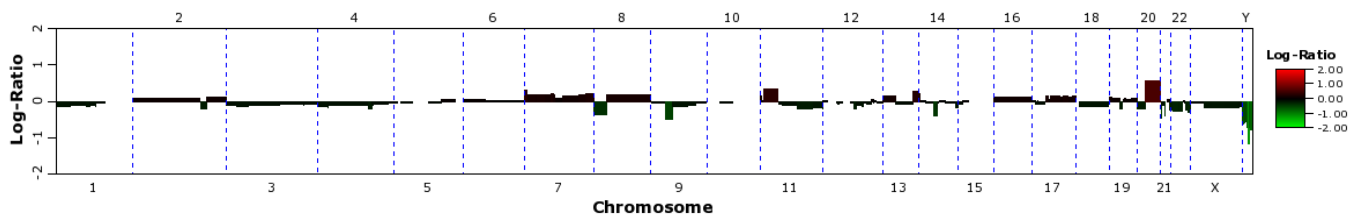
**Dataset 3:** Our third dataset, consists of 98 samples of aCGH profiles coming from 3 different types of primary colorectal cancer: metastasis-free, liver and peritoneal metastasis. 36 samples come from patients who developed liver metastasis, 37 come from patients who developed peritoneal metastasis and 25 from patients



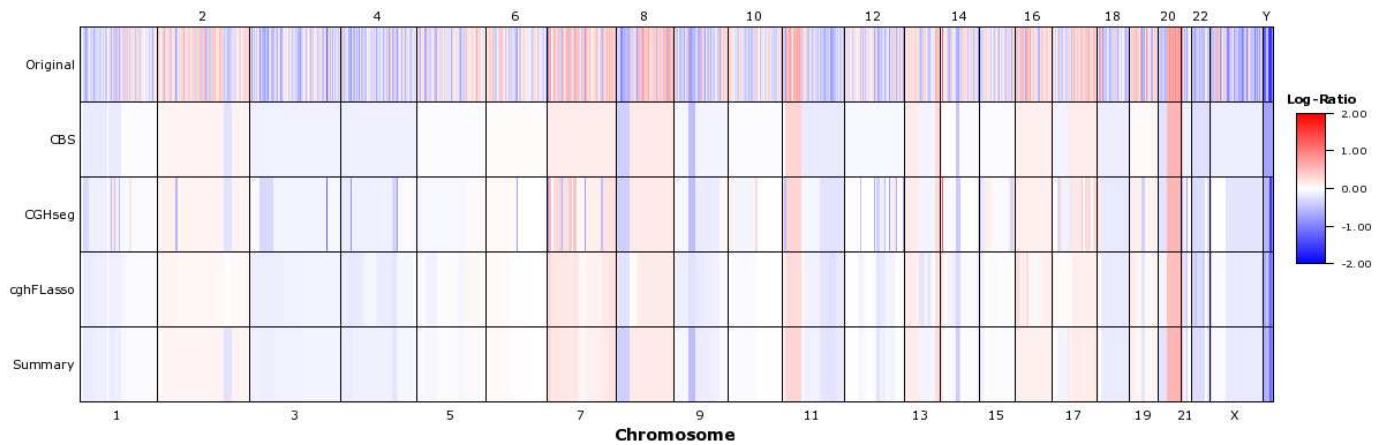
(a) Full male human genome.



(b) Original data. Chromosome numbers are given on top and bottom of the image. Log-ratios are indicated by both the y-axis and the color (green indicates regions of chromosomal loss and red indicates regions of chromosomal gain).



(c) Summary data (Pointwise averaging of all computed profiles).



(d) CNV Heatmap. The first line is the heatmap of the original log-ratios; the last is the heatmap of the averaged profile (pointwise averaging across the outputs of all algorithms); and the lines in the middle are the heatmaps corresponding to the data discretized and smoothed by different algorithms (CBS [36], CGHseg [28] and cghFLasso [35]).

Fig. 2. The images above visualize the CNVs of a sample of colorectal cancer with liver metastasis coming from our third dataset. To visualize the data we used the CGHweb tool (<http://compbio.med.harvard.edu/CGHweb/>).

who remained metastasis-free. The dataset can be found in NCBI GEO database with the code name "GSE20496".

**Dataset 4:** The fourth dataset consists of 101 samples coming from 5 different breast cancer subtypes (basal-like - 23 samples, luminal A - 43 samples, luminal B - 14 samples, ERBB2 - 15 samples, and normal breast-like - 6 samples). Each CGH sample consists of 2149 probes. The dataset can be found in the supplementary data of [7].

Figure 2 visualizes a sample coming from our third

aCGH dataset. That is a cancerous sample which contains colorectal cancer with liver metastasis. In 2b we can see the original log-ratios of the DNA copy number variations throughout the chromosome. In 2c we can see the pointwise averaging of all computed profiles after the sample has been segmented. During segmentation, each single-sample signal is divided into regions of constant copy number, called segments [9], [42]. In this work we do not apply segmentation before feature selection since we want to test our method at the probe

level and also evaluate its ability to identify consecutive important probes that may belong to the same segment. Finally, 2d shows 4 different heatmaps obtained from the same sample. The first line is the heatmap of the original log-ratios; the last is the heatmap of the averaged profile (pointwise averaging across the outputs of all algorithms); and the lines in the middle are the heatmaps corresponding to the data discretized and smoothed by different algorithms (CBS [36], CGHseg [28] and cghFLasso [35]).

## 4 EXPERIMENTAL RESULTS AND DISCUSSIONS

We performed our hybrid structured sparse learning model (shortly HSSL) on aCGH biomarker selection studies. To evaluate the performance of our proposed feature selection method we conducted experiments on 4 aCGH datasets, where we used and compared to five other feature selection methods (which are popularly used in bioinformatics research), including Maximum Influence Feature Selection (MIFS) [22], Relief-F [41], Information Gain (IG) and  $\chi^2$ -statistic (chi-squared) [43] as implemented in Weka [15], and Minimum Redundancy Maximum Relevance (mRMR) found in [27].

In our experiments we measured the performance of each of the above methods using SVMs [8] and Logistic Regression (LR) [5] for classification. For the needs of our experiments we used the LIBSVM [6] implementation of SVM with RBF kernel and the implementation of Logistic Regression found in Weka [15]. We evaluated the performance of each of the different feature selection methods on a range of different number of selected features (from 5 to 100). To assess the classification accuracy we performed 10-Fold cross validation (CV) applying each of the feature selection methods on the same data subsets and using the same SVM parameters, which have been determined in advance as appropriate for the target dataset, throughout the experiments. Furthermore, to eliminate the effect of randomness, we repeated each 10-Fold CV round 10 times, with different sample distributions every time, and we took the average accuracy. The classification accuracy results of all different feature selection and classification method combinations are shown in Figure 3. Figures 4 and 5 give a more in-depth view of the classification behavior of each method by visualizing the confusion matrices of the classification results for all feature selection and classification method combinations, using the top 50 features. The number 50 was selected based on the fact that in most of the experiments, selection of a bigger number of features does not yield a significant increase in classification accuracy. Figure 4 shows the original distribution of values where the sum of all numbers in each row is equal to the number of instances corresponding to that class. Figure 5 shows normalized percentage values. For each row  $i$  corresponding to an actual class, the normalized percentage values are calculated as  $Normalized\_m_{ij} =$

$(m_{ij} / \sum_{j=1}^n m_{ij}) \times 100$ , where  $n$  is the number of classes in the dataset. The normalized confusion matrices have been included to help the reader better understand the percentage accuracies in the confusion matrices.

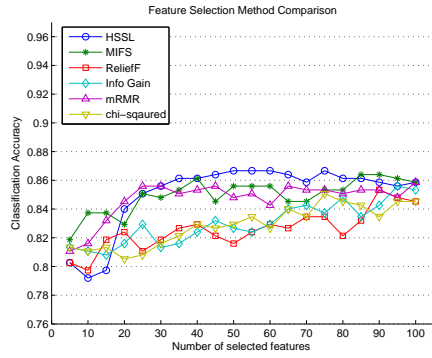
### 4.1 Classification Results Using Selected Biomarkers

The first dataset contains samples from only 2 different classes (oral squamous cell carcinoma vs. healthy tissue), thus forming a binary classification problem. In this dataset HSSL shows superior performance compared to the other feature selection methods for both SVM and Logistic Regression classifiers, especially when using between 30 and 50 features. The sample number imbalance between the two classes (14 versus 61 samples) appears to severely affect the classification accuracy (especially for SVM classifier), however, as it can be observed from the confusion matrices displayed (Figures 4 and 5), the features selected by HSSL mitigate the problem compared to the other feature selection methods. ReliefF seems to consistently have the overall poorest performance in this dataset. The rest of the feature selection methods display similar behavior negatively affected by the class imbalance.

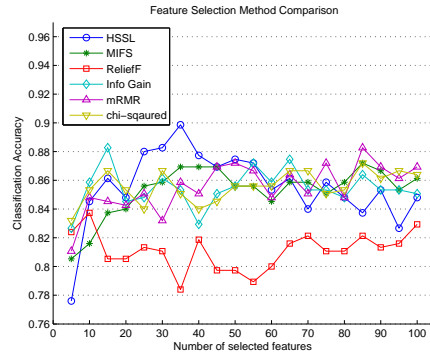
The second dataset is again a binary dataset (breast cancers diagnosed in either young (<45) or old (>70) women). In this dataset, when using SVM classifier, HSSL and MIFS compete for the first place, whereas the other feature selection methods lag far behind. With Logistic Regression as classifier, the overall performance of all methods is lower at smaller number of features and only when using 65 features and above, HSSL shows a clear advantage. As it appears, this is inherently a difficult dataset as there might not be enough biomarkers to differentiate between breast cancers in younger and older women. That leads to a low overall classification accuracy for all feature selection and classification methods.

The third dataset is the first multi-class dataset, containing samples from 3 different types of primary colorectal cancer. The number of samples in each class of this dataset is balanced and the overall classification accuracy for all methods is relatively high. However, HSSL is a clear winner, showing significantly higher performance compared to all other feature selection methods for both SVM and LR classifiers.

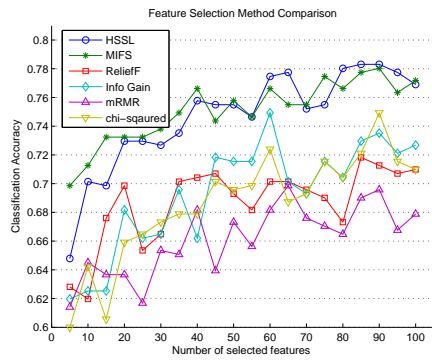
Finally, the fourth dataset contains samples from 5 different classes, thus forming another multiclass classification problem. In this dataset, again, HSSL shows superior performance for both SVM and Logistic Regression classifiers compared to the other feature selection methods, although as one can see, HSSL is a clear winner when using Logistic Regression classifier. The number of samples in this dataset is again not very well balanced among the different classes, since the second class (luminal A type cancer) contains 43 of 101 total samples, whereas the fifth class contains only 6 samples.



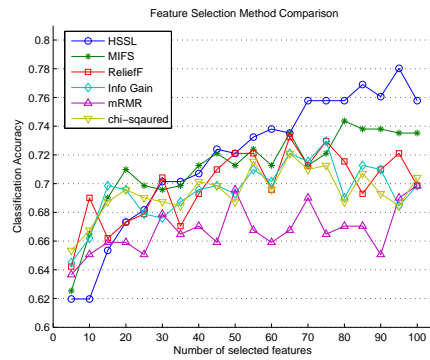
(a) Dataset 1 - SVM



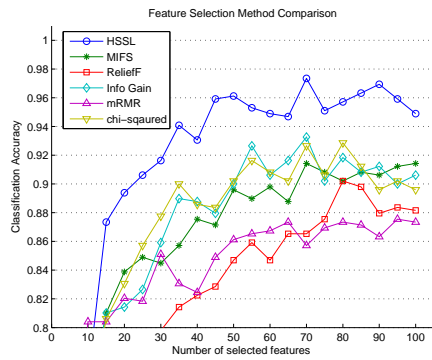
(b) Dataset 1 - Logistic Regression.



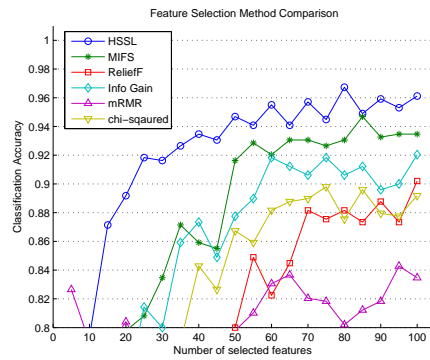
(c) Dataset 2 - SVM



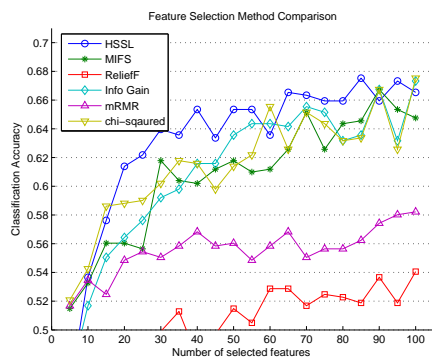
(d) Dataset 2 - Logistic Regression.



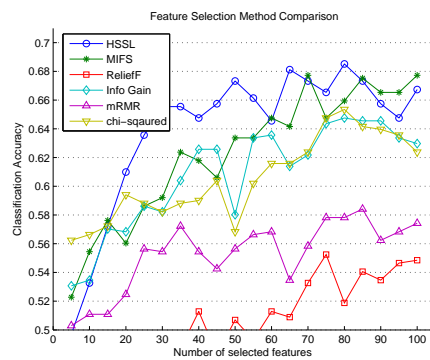
(e) Dataset 3 - SVM



(f) Dataset 3 - Logistic Regression.



(g) Dataset 4 - SVM



(h) Dataset 4 - Logistic Regression.

Fig. 3. Classification accuracy results for all 4 different datasets comparing our proposed method (HSSL) with 5 existing feature selection methods using SVM and Logistic Regression classifiers.



|           |            | HSSL            | MIFS            | Relief-F        | Info Gain       | mRMR            | chi-squared     |              |              |
|-----------|------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|--------------|--------------|
| Dataset 1 | LR         | Predicted class | Predicted class | Predicted class | Predicted class | Predicted class | Predicted class | Actual class |              |
|           |            | A B             | A B             | A B             | A B             | A B             | A B             |              | A B          |
|           | A          | 6 8             | 5 9             | 4 10            | 7 7             | 7 7             | 6 8             | Actual class |              |
|           | B          | 2 59            | 2 59            | 5 56            | 4 57            | 3 58            | 3 58            |              |              |
| SVM       | A B        | A B             | A B             | A B             | A B             | A B             | A B             | Actual class |              |
| A         | 5 9        | 0 14            | 1 13            | 2 12            | 4 10            | 2 12            | Actual class    |              |              |
| B         | 1 60       | 0 61            | 1 60            | 1 60            | 1 60            | 1 60            |                 |              |              |
| Dataset 2 | LR         | A B             | A B             | A B             | A B             | A B             | A B             | Actual class |              |
|           |            | A               | 11 16           | 10 17           | 12 15           | 12 15           | 9 18            |              | 12 15        |
|           | B          | 4 40            | 3 41            | 5 39            | 7 37            | 3 41            | 7 37            |              |              |
|           | SVM        | A B             | A B             | A B             | A B             | A B             | A B             | A B          | Actual class |
| A         | 11 16      | 13 14           | 11 16           | 13 14           | 10 17           | 12 15           | Actual class    |              |              |
| B         | 2 42       | 3 41            | 6 38            | 7 37            | 7 37            | 7 37            |                 |              |              |
| Dataset 3 | LR         | A B C           | A B C           | A B C           | A B C           | A B C           | A B C           | Actual class |              |
|           |            | A               | 35 1 0          | 33 3 0          | 30 4 1          | 32 4 0          | 31 5 1          |              | 31 5 0       |
|           | B          | 0 35 2          | 2 34 0          | 7 28 2          | 5 31 1          | 7 28 3          | 6 31 0          |              |              |
|           | C          | 0 2 23          | 0 2 22          | 4 1 20          | 1 1 23          | 2 3 20          | 2 0 23          |              |              |
| SVM       | A B C      | A B C           | A B C           | A B C           | A B C           | A B C           | A B C           | Actual class |              |
| A         | 36 0 0     | 28 8 1          | 31 4 1          | 33 3 0          | 33 3 0          | 33 3 0          | Actual class    |              |              |
| B         | 1 36 0     | 6 29 2          | 5 30 2          | 4 32 1          | 5 30 2          | 4 32 1          |                 |              |              |
| C         | 0 2 23     | 3 2 20          | 1 2 22          | 0 2 23          | 0 3 22          | 0 1 23          |                 |              |              |
| Dataset 4 | LR         | A B C D E       | A B C D E       | A B C D E       | A B C D E       | A B C D E       | A B C D E       | Actual class |              |
|           |            | A               | 17 4 1 1 0      | 15 5 2 1 0      | 10 11 2 1 0     | 17 4 2 1 0      | 12 6 4 1 0      |              | 15 5 2 1 0   |
|           | B          | 0 40 2 1 0      | 2 36 3 2 0      | 4 33 3 3 0      | 1 33 6 3 0      | 3 33 5 2 0      | 1 33 5 4 0      |              |              |
|           | C          | 3 6 3 1 0       | 3 4 5 2 0       | 2 3 7 2 0       | 3 6 4 1 0       | 3 6 5 1 0       | 3 5 4 1 0       |              |              |
| D         | 1 5 0 9 0  | 2 6 0 8 0       | 3 9 1 2 0       | 1 8 1 5 0       | 2 5 0 7 0       | 1 7 1 6 0       |                 |              |              |
| E         | 1 5 0 0 0  | 1 4 1 0 0       | 1 4 1 0 0       | 0 5 0 1 0       | 0 5 1 0 0       | 0 5 0 0 0       |                 |              |              |
| SVM       | A B C D E  | A B C D E       | A B C D E       | A B C D E       | A B C D E       | A B C D E       | A B C D E       | Actual class |              |
| A         | 14 6 1 1 0 | 5 18 0 0 0      | 8 12 3 0 0      | 15 6 1 1 0      | 8 11 3 1 0      | 13 7 1 1 0      | Actual class    |              |              |
| B         | 0 37 4 1 0 | 0 43 0 0 0      | 2 38 2 0 0      | 1 36 4 1 0      | 1 36 4 1 1      | 1 37 4 1 0      |                 |              |              |
| C         | 3 5 5 1 0  | 1 13 0 0 0      | 3 6 5 0 0       | 3 7 4 1 0       | 3 5 4 1 0       | 3 7 3 1 0       |                 |              |              |
| D         | 0 5 1 10 0 | 1 14 0 0 0      | 2 11 0 1 0      | 0 5 0 9 0       | 2 5 0 8 0       | 0 6 0 9 0       |                 |              |              |
| E         | 0 5 0 0 0  | 0 6 0 0 0       | 1 5 1 0 0       | 0 6 0 0 0       | 0 5 1 0 0       | 0 6 0 0 0       |                 |              |              |
|           |            | HSSL            | MIFS            | Relief-F        | Info Gain       | mRMR            | chi-squared     |              |              |

Fig. 4. Confusion Matrices of the classification results for SVM and Logistic Regression (LR) using the different feature selection methods. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The numbers within a row of a matrix show the distribution of predicted class for the instances belonging to that actual class. Note that each number is the average of 10 different runs of the classification experiment. For readability purposes the numbers have been rounded to the closest integer. This introduces some rounding error, e.g. an original value of 0.3 appears as 0 in the confusion matrix. Gray-scale intensities corresponding to the distribution of values within a row have been used to enhance readability.

|           |              | HSSL            | MIFS            | Relief-F        | Info Gain       | mRMR            | chi-squared     |              |              |
|-----------|--------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|--------------|--------------|
| Dataset 1 | LR           | Predicted class | Predicted class | Predicted class | Predicted class | Predicted class | Predicted class | Actual class |              |
|           |              | A B             | A B             | A B             | A B             | A B             | A B             |              |              |
|           | A            | 44 56           | 37 63           | 26 74           | 49 51           | 53 47           | 44 56           | Actual class |              |
|           | B            | 3 97            | 3 97            | 8 92            | 6 94            | 5 95            | 5 95            |              |              |
| SVM       | A B          | A B             | A B             | A B             | A B             | A B             | Actual class    |              |              |
| A         | 36 64        | 0 100           | 10 90           | 11 89           | 29 71           | 13 87           | Actual class    |              |              |
| B         | 2 98         | 0 100           | 2 98            | 1 99            | 2 98            | 1 99            |                 |              |              |
| Dataset 2 | LR           | A B             | A B             | A B             | A B             | A B             | A B             | Actual class |              |
|           |              | A               | 41 59           | 37 63           | 44 56           | 44 56           | 32 68           |              | 44 56        |
|           | B            | 9 91            | 8 92            | 10 90           | 15 85           | 7 93            | 16 84           | Actual class |              |
|           | SVM          | A B             | A B             | A B             | A B             | A B             | A B             |              | Actual class |
| A         | 41 59        | 47 53           | 41 59           | 50 50           | 39 61           | 46 54           | Actual class    |              |              |
| B         | 4 96         | 6 94            | 13 87           | 15 85           | 15 85           | 16 84           |                 |              |              |
| Dataset 3 | LR           | A B C           | A B C           | A B C           | A B C           | A B C           | A B C           | Actual class |              |
|           |              | A               | 96 4 0          | 92 8 0          | 84 12 4         | 88 12 0         | 85 13 2         |              | 87 13 0      |
|           | B            | 0 96 4          | 6 93 1          | 19 75 6         | 14 84 2         | 18 75 7         | 17 83 1         | Actual class |              |
|           | C            | 1 8 91          | 1 10 90         | 17 2 81         | 4 4 92          | 6 14 80         | 6 2 92          |              |              |
| SVM       | A B C        | A B C           | A B C           | A B C           | A B C           | A B C           | Actual class    |              |              |
| A         | 99 1 0       | 77 22 2         | 85 12 3         | 91 9 0          | 91 9 0          | 92 8 0          | Actual class    |              |              |
| B         | 3 96 1       | 16 79 5         | 12 82 5         | 10 87 3         | 14 81 5         | 11 86 3         |                 |              |              |
| C         | 0 8 92       | 10 10 80        | 6 6 88          | 1 6 93          | 0 13 87         | 1 6 94          |                 |              |              |
| Dataset 4 | LR           | A B C D E       | A B C D E       | A B C D E       | A B C D E       | A B C D E       | A B C D E       | Actual class |              |
|           |              | A               | 73 17 3 6 0     | 65 23 9 3 0     | 43 47 7 3 0     | 73 16 7 4 0     | 51 27 17 4 0    |              | 63 23 8 5 0  |
|           | B            | 0 92 5 2 0      | 4 85 7 4 0      | 10 76 8 6 0     | 3 76 13 8 0     | 7 76 11 6 0     | 3 77 11 8 0     | Actual class |              |
|           | C            | 23 46 21 10 0   | 24 29 36 11 0   | 17 21 49 13 0   | 20 43 30 7 0    | 19 40 34 7 0    | 24 39 27 10 0   |              |              |
| D         | 9 32 1 57 0  | 11 39 0 51 0    | 21 61 5 12 0    | 5 55 8 32 0     | 16 36 3 45 0    | 9 45 7 39 0     | Actual class    |              |              |
| E         | 10 80 7 3 0  | 10 73 10 7 0    | 13 73 13 0 0    | 3 87 0 10 0     | 7 83 10 0 0     | 3 90 0 7 0      |                 |              |              |
| SVM       | A B C D E    | A B C D E       | A B C D E       | A B C D E       | A B C D E       | A B C D E       | Actual class    |              |              |
| A         | 61 28 6 4 1  | 22 78 0 0 0     | 34 54 12 0 0    | 64 28 3 4 0     | 37 46 13 4 0    | 58 32 5 4 0     | Actual class    |              |              |
| B         | 1 86 9 3 0   | 0 100 0 0 0     | 5 89 6 1 0      | 3 85 10 2 0     | 3 84 8 3 1      | 3 85 9 3 0      |                 |              |              |
| C         | 23 33 37 7 0 | 7 91 1 0 0      | 23 43 33 1 0    | 19 47 27 7 0    | 24 39 29 9 0    | 21 50 21 7 0    | Actual class    |              |              |
| D         | 0 31 4 65 0  | 9 91 0 0 0      | 16 75 0 9 0     | 1 36 1 61 0     | 13 31 1 55 0    | 1 37 1 60 0     |                 |              |              |
| E         | 7 90 3 0 0   | 0 100 0 0 0     | 10 80 10 0 0    | 3 93 3 0 0      | 7 83 10 0 0     | 3 93 3 0 0      |                 |              |              |

Fig. 5. Normalized Confusion Matrices of the classification results for SVM and Logistic Regression (LR) using the different feature selection methods. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The numbers within a row of a matrix show the distribution of predicted class for the instances belonging to that actual class. Note that each number is the average of 10 different runs of the classification experiment. The numbers have been normalized to show percentage values ( $Normalized_{m_{ij}} = (m_{ij} / \sum_{j=1}^n m_{ij}) \times 100$ ). For readability purposes the numbers have been rounded to the closest integer. Gray-scale intensities corresponding to the distribution of values within a row have been used to enhance readability.

Out of all the feature selection methods evaluated, MIFS appears to be the most easily affected by class imbalance (as we also saw in dataset 1), whereas HSSL is the least affected method.

In total, we see that HSSL shows top performance in all different datasets and classification methods used. Especially when we are dealing with multi-class problems, such as in datasets 3 and 4, we see that HSSL has a clear advantage compared to existing feature selection methods due to its ability to identify features that may be important for one class but insignificant for the rest of them.

## 4.2 Biomarker Analysis

Apart from classifying the tumor tissue samples based on their aCGH analysis, it is of great importance to identify what genetic abnormalities cause the disease itself. In other words we are interested in identifying the biomarkers that may connect certain properties of the genotype with their corresponding effects on the phenotype. Those connections are already known for some disease types. For example, in Figure 6 we can see the connection between certain genes and diseases as listed in Entrez Genome NCBI Database<sup>1</sup>. The visualizations are made using the on-line Entrez Map Viewer Software<sup>2</sup>. However, for many disease types, their connection to certain genomic functionalities is yet to be discovered. The BAC/PAC clones used to form the aCGH datasets can help towards this direction. BAC (Factor-based Bacterial Artificial Chromosome) and PAC (P1-derived Artificial Chromosome) are cloning systems specifically designed at cloning DNA fragments in excess of 100 - 300 kb. In aCGH analysis, BAC/PAC clones are used to measure areas of the genome with increased or decreased DNA copy numbers compared to the normal/control levels. Each clone region can contain one or more genes. Over- or under-expression of such genes can lead to cell abnormalities such as tumor genesis. Therefore, CNVs that occur consistently for a certain disease in the genomic area covered by a specific clone can be an indication that the associated genes existing in that area could be related to the disease itself.

Our feature selection method allows us to automatically analyze aCGH data and find clones who's CNVs are related to specific cancer types. The clones are ranked in order of importance based on their predictive power with regard to the examined cancer classes of each dataset. For example, in dataset 3, the clone RP11-47L3 is ranked as the most important with regard to its ability to differentiate between the three different cancer

1. Entrez Genome NCBI Database organizes information on genomes including maps, chromosomes, assemblies, and annotations (<http://www.ncbi.nlm.nih.gov/sites/genome>).

2. The Map Viewer provides special browsing capabilities for a subset of organisms in Entrez Genomes. Map Viewer allows you to view and search an organism's complete genome, display chromosome maps, and zoom into progressively greater levels of detail, down to the sequence data for a region of interest.

| Dataset 1 |             |  |
|-----------|-------------|--|
|           | Clones      | Genes  |
| 1         | RP11-43B19  | LPAL2  |
| 2         | RP11-42A17  | GABRG1   |
| 3         | GS1-174H8   | BBS9   |
| 4         | CTB-1O12    | FHIT   |
| 5         | RP11-110I16 | RP11-110I16  |
| 6         | RP11-59E12  | LAMA3  |
| 7         | RP11-52B21  | LRCH1, ESD   |
| 8         | RP11-14I14  | JMJD1C   |
| 9         | RP11-130N6  | N/A  |
| 10        | RP11-283M20 | RPS15A, ARL6IP1, SMG1  |
| 11        | RP11-109D4  | RP11-109D4, ARL6IP1, SMG1  |
| 12        | RP11-119N7  | LOC645481  |
| 13        | RP11-70F16  | N/A  |
| 14        | RP11-221G13 | MAMLD1   |
| 15        | RP11-34J24  | VOPP1  |
| 16        | RP11-162F2  | RPS27AP11  |
| 17        | RP11-88B16  | EFCAB5   |
| 18        | RP11-160L9  | CDK2AP2, CABP2, GSTP1, LOC100505621, NDUFV1, LOC390213, NUDT8, TBX10, ACY3 |
| 19        | RP11-94J8   | IL13RA2, LOC100419790, YAP1P2  |
| 20        | RP11-97P11  | LANCL2, VOPP1  |

TABLE 1

The 20 most important BAC/PAC clones of Dataset 1 and the corresponding genes found in the genomic area covered by each clone.

types of the dataset. Increased copy number of the clone shows a strong correlation with colorectal cancer type 1 (liver metastasis), whereas decreased copy number shows strong correlation with colorectal cancer type 2 (peritoneal metastasis). The CNVs of the clone does not show strong correlation with class type 3 (metastasis-free colorectal cancer), (see Figure 1). The RP11-47L3 comes from locus AC022706 of Homo Sapiens chromosome 17 (see Figure 7). In the same region lies the gene SLFN5 (schlafen family member 5) which encodes a protein believed to have a role in hematopoietic cell differentiation. Therefore, this gene and the corresponding encoded protein may be related to the metastasis type developed by the examined patients. Tables 1, 2, 3 and 4 list the top 20 clones of each dataset and the corresponding genes found in the genomic area covered by each clone. Where "N/A" appears instead of a gene, it means that there is no known gene in the covered area according to NCBI Genome Entrez Database.

## 5 CONCLUSIONS

In this paper we introduced an efficient embedded feature selection method and compared its performance with existing state-of-the-art feature selection methods. The proposed method, which utilizes hybrid structured sparsity-inducing norms to determine the optimal coefficients for the initial set of features, consistently showed superior performance compared to other feature selection methods when used for feature selection in aCGH data. Especially in multi-class problems our method manages to significantly outperform the competitive feature selection methods. Our method is independent of the algorithm to be used during the classification process

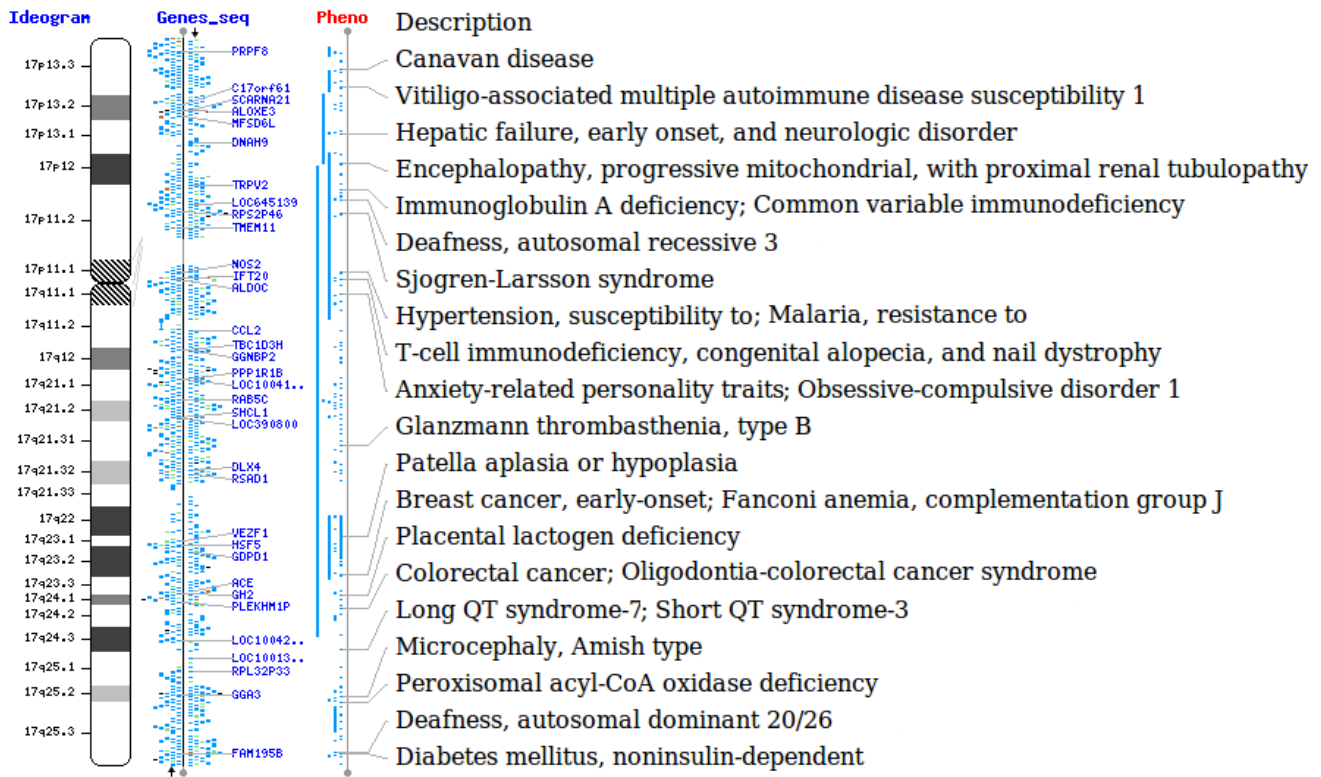


Fig. 6. Genotype-Phenotype mapping of well known genes and diseases on Chromosome 17, extracted from Entrez Genome NCBI Database.

| Dataset 2 |             |  |
|-----------|-------------|--|
| Clones    | Genes       |  |
| 1         | RP11-145B20 | SLC1A2   |
| 2         | RP11-568F15 | OR10V1, OR10Y1P, OR10V3P, OR10V2P, STX3, FABP5L7, MRPL16, GIF, TCN1  |
| 3         | RP11-49D19  | ZBTB3, POLR2G, TAF6L, TMEM179B, TMEM223, NXF1, STX5, WDR74, RNU2-2, SNHG1, SNORD22, SNORD25-SNORD31, SLC3A2, CHRM1 |
| 4         | RP11-729B4  | MS4A14, MS4A5, MS4A1, MS4A12, MS4A13   |
| 5         | RP11-77M17  | SERPING1, MIR130A, LOC100507106, YPEL4, CLP1, ZDHHC5, MED19, LOC100507231, TMX2, C11orf31, BTBD18                  |
| 6         | RP11-129G17 | VN1R55P, RNLS  |
| 7         | RP11-45L17  | C10orf68, ITGB1, LOC100288319  |
| 8         | RP11-35F11  | HRASLS5, LGALS12, TMSL5, RARRES3, HRASLS2  |
| 9         | RP11-181I11 | N/A  |
| 10        | RP11-61G7   | SPAG8, HINT2   |
| 11        | RP11-40G3   | DLG2   |
| 12        | RP11-48K2   | BOD1   |
| 13        | RP11-206I1  | RP11-206I1, LOC100507338, LOC100419850   |
| 14        | RP11-287G20 | CCDC147  |
| 15        | GS1-54J22   | C1GALT1, LOC100505904  |
| 16        | RP11-39C2   | GPR116, GPR110   |
| 17        | RP11-160A13 | PAQR9, LOC100289361, SR140   |
| 18        | RP11-1L22   | GPR39  |
| 19        | RP11-215H8  | ODZ4   |
| 20        | RP11-39I6   | CLTA   |

TABLE 2

The 20 most important BAC/PAC clones of Dataset 2 and the corresponding genes found in the genomic area covered by each clone.

| Dataset 3 |             |   |
|-----------|-------------|---|
| Clones    | Genes       |   |
| 1         | RP11-47L3   | SLFN5   |
| 2         | RP11-202L1  | N/A   |
| 3         | RP11-213G21 | N/A   |
| 4         | RP11-339F13 | EGFR, LOC100507500, LOC100130121, CALM1P2                               |
| 5         | RP11-338H14 | N/A   |
| 6         | CTC-263A14  | LOC100131520  |
| 7         | RP11-359H18 | LOC100131479, RPS27P29, VN1R93P, ZNF675, VN1R94P, ZNF681                |
| 8         | RP11-219A15 | LOC266619, LOC353194, LOC400578, LOC147228, LOC339186, CLPSMCR, TBC1D27 |
| 9         | RP4-552K20  | MAGEC3, LOC100420249, MAGEC1  |
| 10        | RP11-447J13 | CADM2, LOC100422711   |
| 11        | RP11-767J14 | N/A   |
| 12        | RP11-164K24 | LOC100506669, LOC283710   |
| 13        | RP11-125I23 | GTF3A, MTIF3  |
| 14        | RP11-122N14 | DMD   |
| 15        | RP11-326G21 | PDE4DIP, LOC100505971   |
| 16        | RP11-27C22  | RP1-27C22   |
| 17        | RP11-67F24  | IL12A, LOC730109, BRD7P2  |
| 18        | RP11-187L3  | CRYL1   |
| 19        | RP11-426J23 | EPHB6, TRPV6, TRPV5, C7orf34  |
| 20        | RP11-182H20 | TTY8, TTY7B, TTY21, TTY2, TTY1  |

TABLE 3

The 20 most important BAC/PAC clones of Dataset 3 and the corresponding genes found in the genomic area covered by each clone.

| Dataset 4 |             |  |
|-----------|-------------|--|
| Clones    | Genes       |  |
| 1         | RP11-48I18  | ZNF423, MRPS21P7, MRPS21P8   |
| 2         | RP11-58M3   | MARVELD3, PHLPP2, SNORA70D, SNORD71  |
| 3         | CTA-799F10  | SHANK3   |
| 4         | RP11-52K17  | RPL5P26, COL13A1   |
| 5         | RP11-14G23  | TDRG1, LRFN2, LOC100505697   |
| 6         | RP11-105E14 | LIX1L, RBM8A, GNRHR2, PEX11B, ITGA10, ANKRD35, PIAS3, NUDT17, POLR3C, RNF115 |
| 7         | RP11-204D12 | PCSK1  |
| 8         | RP11-44N11  | LOC392265, LOC100507001, ZHX2  |
| 9         | RP11-15L8   | LRFN4, PC, RNU7-23P, MIR3163, C11orf86, SYT12                                |
| 10        | RP11-116F9  | RPL5P22, PNOC, ZNF395  |
| 11        | RP11-249H15 | CDK18  |
| 12        | RP11-16A21  | LOC100131036, SPIRE1   |
| 13        | RP11-141N1  | LOC100132126   |
| 14        | CTB-23D20   | TAX1BP1, JAZF1   |
| 15        | RP11-208E21 | VPS13B, LETM1P3  |
| 16        | RP11-33J8   | SFMBT2   |
| 17        | RP11-35I11  | N/A  |
| 18        | RP11-125O21 | LOC100131849, KCNS2, STK3  |
| 19        | RP11-177M14 | EYA4   |
| 20        | RP11-45B19  | ZFAT, ZFATAS   |

TABLE 4

The 20 most important BAC/PAC clones of Dataset 4 and the corresponding genes found in the genomic area covered by each clone.

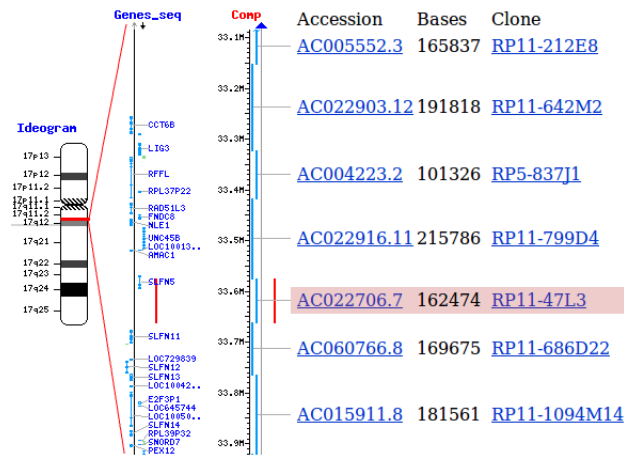


Fig. 7. Clone-Gene mapping in the region 33,080K-34,650K bp of Chromosome 17. In the genomic area covered by the examined clone (RP11-47L3) we find the gene SLFN5.

which makes it ideal for use in combination with different classification methods. Experimenting with four publicly available datasets, containing samples of different cancer types, we showed how our method can be used for biomarker detection and we also presented the top 20 biomarker genes found to be the most related with the examined cancer types for each dataset. Although in this work we examine the performance of our proposed method on aCGH data, it can be also applied to a variety of different data types where feature selection is useful.

## ACKNOWLEDGMENTS

This work was partially supported by the U.S. National Science Foundation (NSF) under award numbers CNS-0923494, CNS-1035913, IIS-1117965, IIS-1302675, IIS-1344152.

## REFERENCES

- [1] C. F. Aliferis, D. Hardin, and P. P. Massion. Machine learning models for lung cancer classification using array comparative genomic hybridization. page 7. American Medical Informatics Association, 2002.
- [2] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.
- [3] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [4] P. S. Bradley and O. L. Mangasarian. Feature selection via concave minimization and support vector machines. In *Machine Learning Proceedings of the Fifteenth International Conference (ICML'98)*, pages 82–90. Citeseer, 1998.
- [5] S. L. Cessie and J. C. V. Houwelingen. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(1):191–201, 1992.
- [6] C. C. Chang and C. J. Lin. LIBSVM: a library for support vector machines. 2001.
- [7] K. Chin, S. DeVries, J. Fridlyand, P. T. Spellman, R. Roydasgupta, W. L. Kuo, A. Lapuk, R. M. Neve, Z. Qian, and T. Ryder. Genomic and transcriptional aberrations linked to breast cancer pathophysiology. *Cancer cell*, 10(6):529–541, 2006.
- [8] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge Univ Pr, 2000.
- [9] A. Daemen, O. Gevaert, K. Leunen, E. Legius, I. Vergote, and B. D. Moor. Supervised classification of array cgh data with hmm-based feature selection. page 468. World Scientific Pub Co Inc, 2009.
- [10] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.
- [11] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. <http://arxiv.org/abs/1001.0736>.
- [12] T. Gambin and K. Walczak. A new classification method using array comparative genome hybridization data, based on the concept of limited jumping emerging patterns. *BMC bioinformatics*, 10(Suppl 1):S64, 2009.
- [13] I. Gorodnitsky and B. Rao. Sparse signal reconstruction from limited data using focus: A re-weighted minimum norm algorithm. *Signal Processing, IEEE Transactions on*, 45(3):600–616, 1997.
- [14] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- [16] D. W. Hosmer and S. Lemeshow. *Applied logistic regression*, volume 354. Wiley-Interscience, 2000.
- [17] J. Huang, A. Salim, K. Lei, K. O’Sullivan, and Y. Pawitan. Classification of array CGH data using smoothed logistic regression model. *Statistics in medicine*, 28(30):3798–3810, 2009.
- [18] K. Y. Kim, J. Kim, H. J. Kim, W. Nam, and I. H. Cha. A method for detecting significant genomic regions associated with oral squamous cell carcinoma using aCGH. *Medical and Biological Engineering and Computing*, 48(5):459–468, 2010.
- [19] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
- [20] J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In *SIGKDD09*, pages 547–556, 2009.
- [21] J. Liu, S. Ji, and J. Ye. Multi-task feature learning via efficient  $l_{2,1}$ -norm minimization. In *UAI2009*, 2009.
- [22] J. Liu, S. Ranka, and T. Kahveci. Classification and feature selection algorithms for multi-class CGH data. *Bioinformatics*, 24(13):i86, 2008.
- [23] D. Luo, C. Ding, and H. Huang. Towards structural sparsity: an explicit  $l_2/l_0$  approach. *ICDM*, 2010.

- [24] A. Y. Ng. Feature selection,  $l_1$  vs.  $l_2$  regularization, and rotational invariance. page 78. ACM, 2004.
- [25] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint  $l_{2,1}$ -norms minimization. *NIPS 2010*, 2010.
- [26] G. Obozinski, B. Taskar, and M. Jordan. Multi-task feature selection. Citeseer, 2006.
- [27] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.
- [28] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin. A statistical approach for array cgh data analysis. *BMC bioinformatics*, 6(1):27, 2005.
- [29] D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37:S11–S17, 2005.
- [30] D. Pinkel, R. Seagraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, and Y. Zhai. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature genetics*, 20:207–211, 1998.
- [31] S. Riccadonna, G. Jurman, S. Merler, S. Paoli, A. Quattrone, and C. Furlanello. Supervised classification of combined copy number and gene expression data. *Journal of Integrative Bioinformatics*, 4(3):74, 2007.
- [32] M. Stojnic.  $l_2/l_1$ -optimization in block-sparse compressed sensing and its strong thresholds. *IEEE Journal of Selected Topics in Signal Processing*, 2009.
- [33] L. Sun, J. Liu, J. Chen, and J. Ye. Efficient recovery of jointly sparse vectors. In *Advances in Neural Information Processing Systems 22*, pages 1812–1820, 2009.
- [34] R. Tibshirani. Regression shrinkage and selection via the LASSO. *J. Royal. Statist. Soc B.*, 58:267–288, 1996.
- [35] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18, 2008.
- [36] E. S. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, 23(6):657, 2007.
- [37] H. Wang, F. Nie, H. Huang, S. Kim, K. Nho, S. L. Risacher, A. J. Saykin, L. Shen, et al. 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.
- [38] H. Wang, F. Nie, H. Huang, S. Risacher, C. Ding, A. J. Saykin, and L. Shen. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *ICCV 2011*, pages 557–562. IEEE, 2011.
- [39] H. Wang, F. Nie, H. Huang, S. L. Risacher, A. J. Saykin, L. Shen, et al. Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics*, 28(12):i127–i136, 2012.
- [40] H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer’s disease progression prediction. In *NIPS*, pages 1286–1294, 2012.
- [41] Y. Wang and F. Makedon. Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. page 498, 2004.
- [42] H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21(22):4084, 2005.
- [43] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. Morgan Kaufmann Publishers, 1997.
- [44] C. Yau, V. Fedele, R. Roydasgupta, J. Fridlyand, A. Hubbard, J. W. Gray, K. Chew, S. H. Dairkee, D. H. Moore, and F. Schittulli. Aging impacts transcriptomes but not genomes of hormone-dependent breast cancers. *Breast Cancer Res*, 9(5):R59, 2007.
- [45] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.
- [46] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.



**Vangelis Metsis** is a Research Assistant Professor at the Department of Computer Science and Engineering (CSE) of the University of Texas at Arlington (UTA). Dr. Metsis earned his B.S. degree in Computer Science in 2005 from the Department of Informatics of Athens University of Economics and Business in Greece, and his Ph.D. in 2011 from the Department of Computer Science and Engineering of the University of Texas at Arlington. During the years 2006–2007, he has worked as a Research Associate for the E.C. funded project MedIEQ at the Department of Informatics and Telecommunications of the National Center for Scientific Research (NCSR) "Demokritos", Greece. His research interests include machine learning, data mining, bioinformatics, and computer vision with applications in healthcare.



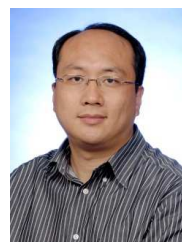
**Fyllia Makedon** is Jenkins-Garrett Professor and Department Head of Computer Science and Engineering at the University of Texas at Arlington (UTA). She received her Ph.D. in Computer Science from Northwestern University in 1982. Between 1991–2006, she was professor of computer science at Dartmouth College where she founded and directed the Dartmouth Experimental Visualization Laboratory (DEVLAB). In 2005–2006, she was Program Director at the National Science Foundation. Prior to Dartmouth, Prof.

Makedon was Assistant and Associate Professor at the Univ. of Texas at Dallas (UTD), where she directed the Computer LEARNING Research Center (CLEAR). She has supervised over 20 Ph.D. theses and numerous Masters Degree theses.



**Dinggang Shen** is a Professor of Radiology, Biomedical Research Imaging Center (BRIC), Computer Science, and Biomedical Engineering in the University of North Carolina at Chapel Hill (UNC-CH). He is currently directing the Center for Image Informatics, the Image Display, Enhancement, and Analysis (IDEA) Lab in the Department of Radiology, and also the medical image analysis core in the BRIC. He was a tenure-track assistant professor in the University of Pennsylvania (UPenn), and a faculty member

in the Johns Hopkins University. Dr. Shen's research interests include medical image analysis, computer vision, and pattern recognition. He has published more than 450 papers in the international journals and conference proceedings. He serves as an editorial board member for four international journals. He also serves in the Board of Directors, The Medical Image Computing and Computer Assisted Intervention (MICCAI) Society.



**Heng Huang** received the B.S. degree in Automation and the M.S. degree in Information Measurement Technology and Instruments from Shanghai Jiao Tong University, Shanghai, China, in 1997 and 2001, respectively. He received the Ph.D. degree in Computer Science from Dartmouth College in 2006. Since 2012, he has been an associate professor in computer science and engineering department at University of Texas at Arlington. From 2007 to 2012, he was an assistant professor at the same department. His research interests include machine learning, data mining, bioinformatics, computer vision, and health informatics.