IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON BIG DATA LEARNING AND DISCOVERY

Received June 17, 2018, accepted August 26, 2018, date of publication August 31, 2018, date of current version September 21, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2868098

Oriented Feature Selection SVM Applied to Cancer Prediction in Precision Medicine

YANG SHEN[1], CHUNXUE WU[1], CONG LIU[1], YAN WU[2], AND NAIXUE XIONG[3,4]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2]School of Public and Environmental Affairs, Indiana University, Bloomington, IN 47405, USA
[3]School of Computer Science and Technology, Tianjin University, Tianjin 300350, China
[4]Department of Mathematics and Computer Science, Northeastern State University, Tahlequah, OK 74464, USA

Corresponding author: Naixue Xiong (xiongnaixue@gmail.com)

This work was supported in part by the Shanghai Science and Technology Innovation Action Plan Project under Grant 16111107502 and Grant 17511107203, in part by the National Natural Science Foundation of China under Grant No.61502220, and in part by the Program for Tackling Key Problems in Henan Science and Technology under Grant 172102310636.

ABSTRACT Advances in the gene sequencing technology and the outbreak of artificial intelligence have made precision medicine a reality recently. Applying machine learning algorithms to cancer prediction using gene expression data helps to discover the link between genetic data and cancer, which will promote the development and application of precision medicine. Considering the natural order of genes, a new classification method that combines fused lasso and elastic net as regularization for linear support vector machine (SVM), which uses huberized hinge loss as the loss function, is proposed in this paper, which we name it oriented feature selection SVM (OFSSVM). Due to the characteristics of the elastic net and fused lasso, the OFSSVM can not only provide automatic feature selection, but also average the adjacent coefficients, resulting in a sparse and smooth solution. We demonstrate its effectiveness in both binary classification and multiclass classification in the sense of comprehensive evaluation that not only the classification accuracy but also the interpretability are considered. The experiments show that the OFSSVM is an appealing compromise between interpretability and classification accuracy, and is superior to other traditional methods in the sense of comprehensive evaluation.

INDEX TERMS Cancer prediction, elastic net, feature selection, fused lasso, gene expression data, machine learning, precision medicine, SVM.

I. INTRODUCTION

Precision medicine, a concept as a part of healthcare for a long time, reappears in the public view recent years due to the advances in gene sequencing technology and outbreaks of artificial intelligence.

According to the Precision Medicine Initiative that is announced by then President of the United States, Barack Obama, On January 30, 2015, precision medicine is ''an innovative approach that takes into account individual differences in people's genes, environments, and lifestyles [1].'' In contrast with traditional medical treatments that are designed for the ''average patient'', the precision medicine provides an alternative to patients who do not respond to those ''one-size-fits-all'' treatments. It emphasizes the goal to overcome still existing limitations of therapeutic interventions by tailoring medical treatments to individual patient characteristics [2]. Several powerful new discoveries [3]–[6] and new treatments [7]–[9] that are tailored to specific characteristics, such as a person's genetic makeup, or the genetic profile of an individual's tumor, have already been achieved by advances in precision medicine. Many of these treatments are drugs known as targeted therapies [10] that block the growth and spread of cancer by interfering with specific molecules (''molecular targets'') that are involved in the growth, progression, and spread of cancer. As a cornerstone of precision medicine, one targeted therapy only works on cancers that have a corresponding target, so a biopsy is needed to see if the genetic change targeted by the treatment is present in patient's cancer or to determine how DNA changes that may be causing the cancer to grow [11]. This is helping transform the way of treating cancer. For example, taking routine molecular testing as part of patient care enables physicians to select treatments that improve chances of survival and reduce the risk of adverse effects.

As the premise of precision medicine, genetic information that benefits from the gene sequencing technology can be

obtained easily over recent decades. DNA microarray, a collection of microscopic DNA spots attached to a solid surface, is a tool can measure the expression levels of large numbers of genes simultaneously or genotype multiple regions of a genome [12]. It can determine whether the DNA from a particular individual contains a mutation in genes. With a proven track record spanning nearly two decades in the lab, microarray platforms are still more economical and yield higher throughput, providing significant advantages when working with a large number of samples [13]. On the other hand, the rapid development of sequencing technology and the significant drop in costs have triggered the implementation of genome testing in patient care. Next generation sequencing (NGS) techniques have been used increasingly in clinical laboratories to scan the whole or part of the human genome in order to facilitate diagnosis and prognostics of genetic disease [14]. It allows researchers to generate more complete and scientifically accurate data than previously possible with microarrays.

Based on the abundant gene information, cancer prediction is the antecedent application in precision medicine, since judging whether there is a tumor or classifying different types of tumor will facilitate the application of targeted therapies, revealing great importance in cancer diagnosis and drug discovery [15]. Different from most previous cancer classification studies that are clinical-based and have limited diagnostic ability, gene expression data provides the systematic information related to cancer and makes a deep insight into how the cancer come into being possible. In general, the gene expression data has following characteristics [16]: (1) The dimensions of gene expression data are very high and usually contain thousands or even tens of thousands of genes. (2) The publicly available data set size is usually very small or very large due to the inclusion of noise data. (3) Most genes are irrelevant to cancer distinction. There is an unmet need to develop computational algorithms for cancer diagnosis, prognosis and therapeutics that can identify complex patterns and help in classifications based on plethora of emerging cancer research outcomes in public domain [17].

Machine learning, one of major branches of artificial intelligence and the most rapidly developing subfield of AI research, has been used for years in medical domain for the intelligent data analysis to make not only future predictions of outcome of certain treatment but also to find hidden relationships within the medical data [18]–[20]. It holds a great potential for pattern recognition in cancer datasets, as evident from recent literature survey [21]–[24]. Using machine learning algorithms for cancer prediction, such as determining whether to have cancer or classifying cancer into different type or subtype, has become a research hotspot. Many classification methods have been proposed by researchers. Chandrasekar and Meena [25] put forward to use fast extreme learning method with ANP (Analytic Network Process) for cancer classification, resulting in higer classification accuracies, less training time, and lower implementation complexity than traditional methods. Abdel-Ilah and Sahinbegovic [26] implemented a feed forward back propagation network (FFBPN) for classification of breast cancer cases to malignant or benign. With selecting the number of hidden layers, number of neurons in the hidden layer and the type of activation functions in hidden layers, an Artificial Neural Network (ANN) with high and acceptable level of accuracy can be obtained. Karabatak and Ince [27] presented an automatic diagnosis system for detecting breast cancer based on association rules and neural network. Begum *et al.* [28] proposed a method called ADASVM, which is a combination of AdaBoost and support vector machines (SVMs). Although all of these methods yield excellent classification accuracy, there has nothing to do with the biological significance since that which genes are related to the cancer or which gene is most likely the culprit is still fuzzy to the pathologist. In view of this, feature selection is taken as a routine operation before classification, that not only brings the biological significance but also better accuracy. Combined with feature selection, Akay [29] utilized SVM to diagnose breast cancer. The F-score was used as a measurement of feature selection; the larger the F-score was, the more likely this feature was more discriminative. The result showed that the highest classification accuracy (99.51%) was obtained for the SVM model that contained five features when applied this method to Wisconsin breast cancer dataset. Student and Fujarewicz [30] used partial least squares (PLS) to select genes and extended this idea to multiclass classification, resulting in a very high accuracy rate for different combinations of classification methods and giving very stable feature rankings at the same time. Other classifiers associated with different feature selection measurements, such as information gain [31], neighborhood granules and the entropy [32] were also proposed. Separating from the classifiers these procedures of feature selection are generally isolated. Benefiting from the characteristic of lasso and its variants which will be introduced in Section II-C.2, some automatic feature selection classifiers have been raised. Beyond the problem of cancer prediction, we wonder if the natural order of the genes makes some sense to the causes of cancer; taking this into account may produce significant output in cancer prediction.

In this paper, we propose a method named oriented feature selection SVM (OFSSVM) that inspired by the work of Wang *et al.* [33] and Rapaport *et al.* [34]. Using huberized hinge loss [35] as the loss function, OFSSVM combines fused lasso [36] and elastic net [37] as a measure of regularization. It turns out that the model is a generalized fused lasso problem, which is tricky task. To over come this trouble, many approaches had been studied, such as coordinate-wise descent algorithm [38], split Bregman iteration [39] and other methods [40]–[42]. We use alternating direction method of multipliers (ADMM) [43] to solve this model, which is inspired by Watanabe *et al.* [44], and test the model on various data sets to evaluate its performance on cancer prediction, such as determine whether to have a cancer or the subtype of certain cancer, and extend it to multiclass case. Not only the classification accuracy but also the interpretability

of classifier is taken into account to get a comprehensive evaluation. A comparison between our OFSSVM and linear SVM, EN-SVM [45], HHSVM [33] and fused SVM [34] are also presented. The experiments show that the OFSSVM is an appealing compromise between interpretability and classification accuracy, and is superior to others in the sense of comprehensive evaluation.

The rest of the paper is organized as follows. In Section II, we give a simple introduction to notation, data sets, related methods, our OFSSVM and the implementation of it. In Section III, experiments tested on different data sets and comparison between multiple classifiers are presented. Discussion around the performance based on classification accuracy and interpretability is also involved. Finally, we conclude the paper along with outlining future directions in Section IV.

II. MATERIALS AND METHODS

A. NOTATION

In general, a matrix $X \in \mathbf{R}^{n\times p}$ represents the sample set which contains n observations (or samples) and there are p predictors (or features) for each observation. Since the data sets studied in this paper are gene expression data, the matrix X denotes n patients (tumor samples) and p genes (expression level) for each patient accordingly. That is,

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{12} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

It can also be written as the form of vector, namely, $X = ({x_1}^T, {x_2}^T, \dots, {x_n}^T)^T$ where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ represents the i-th patient which has p genes. The corresponding responses are denoted as $y = (y_1, y_2, \dots, y_n)^T$, where y_i indicates the class to which x_i belongs.

B. MATERIALS

1) PROSTATE TUMOR DATASET

This dataset comes from the study about relationship between clinical behavior of prostate cancer and underlying gene expression differences by Singh *et al.* [46] This dataset consists of 52 prostate tumors and 50 nontumor prostate samples with 10509 genes available.

2) AML/ALL DATASET

This dataset comes from a proof-of-concept study published in 1999 by Golub *et al.* [47] The training set consists of 38 leukemia patients of which 11 suffer from acute myeloid leukemia (AML) and 27 from acute lymphoblastic leukemia (ALL). The test set consists of 34 patients of which 14 suffer from AML and 20 from ALL. The number of genes is 7128.

TABLE 1. GCM dataset.

Cancer class	Training Set	Test Set
Breast	8	4
Prostate	8	6
Lung	8	4
Colorectal	8	4
Lymphoma	16	6
Bladder	8	3
Melanoma	8	2
Uterus	8	2
Leukemia	24	6
Renal	8	3
Pancreas	8	3
Ovary	8	4
Mesothelioma	8	3
CNS	16	4
Total	144	54

3) GCM DATASET

This dataset comes from the work of Ramaswamy *et al.* [48] It involves a training set of 144 tumor samples, spanning 14 different types of cancer, and a test set of 54 samples as shown in Table 1. 16063 genes are available for Gene expression measurements.

C. METHODS

1) SVMS

The SVMs root in a technique named separating hyperplane. Define a hyperplane by

$$\{x : f(x) = x^T\beta + \beta_0\}, \tag{1}$$

then a classification rule induced by $f(x)$ is

$$G(x) = sign[x^T\beta + \beta_0], \tag{2}$$

where $G(x)$ denotes the class that observation x belongs to.

The evolutionary process of the SVMs can be regarded as a path from the maximal margin classifier to SVMs via support vector classifier. The maximal margin classifier tries to find a separating hyperplane for which the margin is largest - that is, it is the hyperplane that has the farthest minimum distance to the training observations [49]; indeed, the maximal margin hyperplane depends directly on only a small subset of the observations which lie on the margin and called support vectors. Then an observation can be classified based on which side of the maximal margin hyperplane it lies due to (2). The maximal margin classifier suits for the situation that the observations can be perfectly separated. Nevertheless, in many cases the observations are not able to be separated exactly by a simple hyperplane; no separated hyperplane exists for these cases. The support vector classifier generalizes maximal margin classifier to suit for the non-separable cases by introducing the idea that allows a few observations to be misclassified; observations can lie on the wrong side of the margin or hyperplane.

In general, the solution of support vector classifier can be represented as the following optimization problem:

$$\begin{aligned} &\min_{\beta,\beta_0,\xi,t} \ \frac{1}{2}||\beta||_2^2 \\ &\text{subject to } y_i({x_i}^T\beta+\beta_0) \geq 1-\xi_i \\ &\qquad \sum_{i=1}^{n}\xi_i \leq t \\ &\qquad \xi_i \geq 0, \quad i=1,2,\ldots,n, \end{aligned} \tag{3}$$

where y_i has been coded as $\{-1,+1\}$ since the binary classification and t is a nonnegative tuning parameter while ξ_i $(i=1,2,\ldots,n)$ are slack variables that allow individual observations to be on the wrong side of the margin or the hyperplane. Computationally it is convenient to rephrased (3) in the equivalent form

$$\begin{aligned} &\min_{\beta,\beta_0} \ \frac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to } \xi_i \geq 0, \quad y_i({x_i}^T\beta+\beta_0) \geq 1-\xi_i \quad \forall i, \end{aligned} \tag{4}$$

where there is a one-to-one correspondence between the parameters C and t in (3); the separable case corresponds to $C=\infty$.

After performing some algebra, it reveals that the solution function $f(x)$ can be given as

$$\begin{aligned} f(x) &= x^T\beta+\beta_0 \\ &= \sum_{i=1}^{n}\alpha_i y_i\langle x,x_i\rangle + \beta_0, \end{aligned} \tag{5}$$

where α_i $(i=1,2,\ldots,n)$ is nonzero only for the support vectors in the solution. Note that the support vectors in support vector classifier is a little bit different from which in maximal margin classifier; in this case the support vectors include the points lie on the edge of the margin ($\xi_i=0$, $0<\alpha_i<C$) and those locate in the inside of margin ($\xi_i>0, \alpha_i=C$).

The support vector classifier still produces a linear decision boundaries for classification. However, in practice there are many cases that the decision boundaries are non-linear; the classes overlap and any linear classifier will perform poorly. In order to catch the non-linear relationship between the features and the responses, there is a straightforward method that adds the power of the original features or the products of each pair of features into the feature space, resulting in an enlarged feature space. Using basis expansions, an enlarged feature space can be achieved. Define the basis function $h_m(x)$, $m=1,\ldots,M$, the support vector classifier is fitted using input features $h(x_i)=(h_1(x_i),h_2(x_i),\ldots,h_M(x_i))$, $i=1,\ldots,n$, and produce the function $f(x)=h(x)^T\beta+\beta_0$ [50]. Although it seems that it is still a linear function, a non-linear function can be seen from the perspective of the original feature space. Not specify the exact form of basis function, SVMs use kernel K to realize the goal of obtaining an enlarged feature space which has a very large, even infinite dimensions. Given the kernel function $K(x,x')=\langle h(x),h(x')\rangle$, the equation (5) can be rephrased as

$$\begin{aligned} f(x) &= h(x)^T\beta+\beta_0 \\ &= \sum_{i=1}^{n}\alpha_i y_i\langle h(x),h(x_i)\rangle + \beta_0 \\ &= \sum_{i=1}^{n}\alpha_i y_i K(x,x_i) + \beta_0. \end{aligned} \tag{6}$$

K should be a symmetric positive (semi-) define function. The general choice for K are dth-degree polynomial $K(x,x')=(1+\langle x,x'\rangle)^d$, radial basis $K(x,x')=exp(-\gamma||x-x'||^2)$ and neural network $K(x,x')=tanh(k_1\langle x,x'\rangle+k_2)$.

2) REGULARIZATION ON SVMS

With $f(x)=h(x)^T\beta+\beta_0$, equation (4) can also be rephrased as the form *loss* + *penalty* [50], [51]

$$\min_{\beta_0,\beta}\sum_{i=1}^{n}[1-y_i f(x_i)]_+ + \frac{\lambda}{2}||\beta||_2^2, \tag{7}$$

where the subscript "+" indicates positive part and $\lambda=1/C$.

Generalize the form *loss* + *penalty* which is wildly used in regression and classification problems by defining the solution as

$$\hat{\beta}(\lambda)=\arg\min_{\beta\in\mathrm{R}^p} L(y,f(X))+\lambda J(\beta), \tag{8}$$

where $L(y,f(x))$ is the loss function and $J(\beta)$ denotes the penalty on coefficients. When tuning parameter $\lambda\to\infty$, the L_2 norm penalty $||\beta||_2^2$ tends to shrink the estimates of β_i towards zero but never exactly equal zero, resulting in a significant decrease in variance. In the same vein, the L_1 norm (lasso) penalty $||\beta||_1$ is introduced in the procedure of regularization. Distinguish from L_2 penalty the L_1 penalty will shrink some coefficients to exactly zero when λ is sufficiently large, namely conduct feature selection or lead to sparsity. Bradley and Mangasarian [52] proposed the L_1-norm SVM, but it still suffered some troubles; that is, the number of features obtained from the L_1 regularization is at most $\min(n,p)$ for all values of λ [37], [53] and the L_1 penalty tends to select only one feature among a set of strong but correlated features [37].

The elastic net penalty [37] is a compromise between L_1 and L_2 regularization, and has the form

$$J(\alpha,\beta)=\sum_{j=1}^{p}\left(\alpha|\beta_j|+(1-\alpha){\beta_j}^2\right). \tag{9}$$

The second term encourages highly correlated features to be averaged, while the first term encourages a sparse solution in the coefficients of these averaged features. It can perform like the L_1 norm doing feature selection and automatically include the whole groups of features that have highly correlation once one feature among them is selected, namely group effect.

Take the order of features into account, Tibshirani *et al.* [36] proposed the fused lasso penalty

$$J(\lambda_1, \lambda_2, \beta) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|. \quad (10)$$

The first penalty encourages sparsity in the coefficients; the second penalty encourages sparsity in their differences; that is, flatness of the coefficient profiles β_j as a function of the index set j.

3) RELATED METHODS

Wang *et al.* [33] proposed an method named HHSVM that combines the elastic net penalty and huberized squared hinge loss [35]. It has the form

$$\min_{\beta_0,\beta} \sum_{i=1}^{n} L\left(y_i(\beta_0 + {x_i}^T\beta)\right) + \lambda_1||\beta||_1 + \frac{\lambda_2}{2}||\beta||_2^2, \quad (11)$$

where $\lambda_1, \lambda_2 \geq 0$ are regularization parameters and the loss function has the form

$$L(yf) = \begin{cases} 0, & \text{for } yf > 1, \\ (1-yf)^2/2\delta, & \text{for } 1-\delta < yf \leq 1, \\ 1 - yf - \delta/2, & \text{for } yf \leq 1-\delta, \end{cases} \quad (12)$$

where $\delta \geq 0$ is a pre-specified constant.

Rapaport *et al.* [34] proposed a method named fused SVM to classify the arrayCGH data by adding the fused lasso penalty instead of a L_2 penalty. It has the form

$$\begin{aligned} &\min_{\beta} \sum_{i=1}^{n} \max(0, 1 - y_i {x_i}^T\beta) \\ &\text{subject to } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq \mu \\ &\sum_{j=1}^{p} |\beta_j| \leq \lambda, \end{aligned} \quad (13)$$

where λ and μ are two parameters that control the relative trade-off between fitting the training data, enforcing sparsity of the solution (small λ) and enforcing the solution to be piecewise constant (small μ).

4) OFSSVM

Instead of using a plain L_1 penalty in the combination of elastic net, the method we propose in this paper choose a regularization inspired by fused lasso and the model has the form

$$\begin{aligned} \min_{\lambda_1,\lambda_2,\lambda_3,\beta,\beta_0} &\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda_1 \sum_{j=1}^{p} |\beta_j| \\ &+ \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| + \frac{\lambda_3}{2} \sum_{j=1}^{p} {\beta_j}^2, \end{aligned} \quad (14)$$

where $L(y_i, f(x_i))$ is the huberized hinge loss function as in (12).

It can be written in a more compacter form

$$\min_{\lambda_1,\lambda_2,\lambda_3,\beta,\beta_0} L(y, f(X)) + \lambda_1||\beta||_1 + \lambda_2||D\beta||_1 + \frac{\lambda_3}{2}||\beta||_2^2, \quad (15)$$

where $L(y, f(X)) = \sum_{i=1}^{n} L(y_i, f(x_i))$ and $D \in \mathbf{R}^{(p-1)\times p}$ with $D_{j,j} = -1$, $D_{j,j+1} = 1$, and the rest elements are zero.

Due to the gene expression data has a sufficiently large dimension, a basis expansion or kernel has not been adopted and the $f(x)$ has the form in (1), namely, $f(x) = x^T\beta + \beta_0$.

5) IMPLEMENTATION OF OFSSVM

The ADMM [43] solves problem in the form

$$\begin{aligned} &\min\ f(x) + g(z) \\ &\text{subject to } Ax + Bz = c \end{aligned} \quad (16)$$

with variables $x \in \mathbf{R}^n$ and $z \in \mathbf{R}^m$, where $A \in \mathbf{R}^{p\times n}$, $B \in \mathbf{R}^{p\times m}$, $c \in \mathbf{R}^p$, $f : \mathbf{R}^n \to \mathbf{R}\bigcup\{+\infty\}$ and $g : \mathbf{R}^m \to \mathbf{R}\bigcup\{+\infty\}$ are closed convex function. To solve above optimization problem, it consists of the iteration

$$\begin{aligned} x^{k+1} &= \arg\min_x \left(f(x) + (\rho/2)||Ax + Bz^k - c + u^k||_2^2\right) \\ z^{k+1} &= \arg\min_x \left(g(z) + (\rho/2)||Ax^{k+1} + Bz - c + u_k||_2^2\right) \\ u^{k+1} &= u^k + Ax^{k+1} + Bz^{k+1} - c \end{aligned} \quad (17)$$

where u is the scaled dual variable, $\rho > 0$ is called the penalty parameter and the superscript of each variable is the iteration counter. The convergence of ADMM has been explored by many authors, including Gabay [54] and Eckstein and Bertsekas [55].

Assume that the predictors are all standardized to have mean 0 and variance 1, that is, $\sum_i x_{ij}/n = 0$, $\sum_i x_{ij}^2 = 1$, then β_0 is typically chosen as zero since the β_0 is expected value when there is no input. In order to make our model suit for the form of ADMM, we rewrite the hinge loss function

$$\mathcal{L}(YX\beta) = L(y, f(X)) = \sum_{i=1}^{n} L(y_i, f(x_i)), \quad (18)$$

where $Y = diag\{y_1, y_2, \ldots, y_n\}$ is a $n \times n$ diagonal matrix and the ith element of $\{YX\beta\}^{n\times 1}$ is $y_i {x_i}^T\beta$. Then our model can be represented as

$$\min_{\lambda_1,\lambda_2,\lambda_3,\beta} \mathcal{L}(YX\beta) + \lambda_1||\beta||_1 + \lambda_2||D\beta||_1 + \frac{\lambda_3}{2}||\beta||_2^2. \quad (19)$$

Notice that equation (19) is a unconstrained problem, auxiliary variables should be introduced to convert it to an equivalent constrained optimization problem. The form induced by introducing auxiliary variables shows as

$$\begin{aligned} \min_{\lambda_1,\lambda_2,\lambda_3,\beta} &\ \mathcal{L}(v_1) + \lambda_1||v_2||_1 + \lambda_2||v_3||_1 + \frac{\lambda_3}{2}||v_2||_2^2 \\ \text{s.t. } &YX\beta - v_1 = 0,\ \beta - v_2 = 0, \\ &Dv_4 - v_3 = 0,\ \beta - v_4 = 0, \end{aligned} \quad (20)$$

where v_1, v_2, v_3, v_4 are auxiliary constraint variables. Then the corresponding component with (16) is as follows:

$$f(x) = \lambda_2||v_3||_1, \quad g(z) = \mathcal{L}(v_1) + \lambda_1||v_2||_1 + \frac{\lambda_3}{2}||v_2||_2^2$$
$$A = \begin{pmatrix} YX & 0 \\ I & 0 \\ 0 & I \\ I & 0 \end{pmatrix} \in \mathbf{R}^{(n+3p-1)\times\,(2p-1)}, \quad x = \begin{pmatrix} \beta \\ v_3 \end{pmatrix} \in \mathbf{R}^{2p-1},$$
$$B = \begin{pmatrix} -I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & -D \\ 0 & 0 & -I \end{pmatrix} \in \mathbf{R}^{(n+3p-1)\times\,(n+2p)},$$
$$z = \begin{pmatrix} v_1 \\ v_2 \\ v_4 \end{pmatrix} \in \mathbf{R}^{n+2p}, \quad c = \mathbf{0}. \tag{21}$$

Now the x-update in (17) splits into two separate problems

$$\begin{aligned} \beta^{k+1} &= \arg\min_{\beta}(||YX\beta - v_1^k + u_1^k||_2^2 + ||\beta - v_2^k + u_2^k||_2^2 \\ &\quad + ||\beta - v_4^k + u_4^k||_2^2) \\ v_3^{k+1} &= \arg\min_{v_3}\left(\lambda_2||v_3||_1 + \frac{\rho}{2}||v_3 - Dv_4^k + u_3^k||_2^2\right). \end{aligned} \tag{22}$$

Accordingly, the z-update splits into three separate problems

$$\begin{aligned} v_1^{k+1} &= \arg\min_{v_1}\left(\mathcal{L}(v_1) + \frac{\rho}{2}||YX\beta^{k+1} - v_1 + u_1^k||_2^2\right) \\ v_2^{k+1} &= \arg\min_{v_2}(\lambda_1||v_2||_1 + \frac{\lambda_3}{2}||v_2||_2^2 \\ &\quad + \frac{\rho}{2}||\beta^{k+1} - v_2 + u_2^k||_2^2) \\ v_4^{k+1} &= \arg\min_{v_4}(||v_3^{k+1} - Dv_4 + u_3^k||_2^2 \\ &\quad + ||\beta^{k+1} - v_4 + u_4^k||_2^2). \end{aligned} \tag{23}$$

The above (u_1, u_2, u_3, u_4) are dual variables corresponding to (v_1, v_2, v_3, v_4) respectively, and $u_i \in \mathbf{R}^{n_i}$, $\sum_{i=1}^4 n_i = n+3p-1$. Then the u-update can be formulated as

$$\begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ u_3^{k+1} \\ u_4^{k+1} \end{pmatrix} = \begin{pmatrix} u_1^k \\ u_2^k \\ u_3^k \\ u_4^k \end{pmatrix} + \begin{pmatrix} YX\beta^{k+1} - v_1^{k+1} \\ \beta^{k+1} - v_2^{k+1} \\ v_3^{k+1} - Dv_4^{k+1} \\ \beta^{k+1} - v_4^{k+1} \end{pmatrix} = u^{k+1}. \tag{24}$$

Using the direct method the β-update in (22) can be represented equivalently as a closed form

$$\begin{aligned} \beta^{k+1} = (X^TX + 2I_p)^{-1}(X^TY^T(v_1^k - u_1^k) \\ + v_2^k - u_2^k + v_4^k - u_4^k), \end{aligned} \tag{25}$$

where $X^TX + 2I_p$ is always invertible while the subscript of identity matrix I denotes number of dimensions. In other words, computing the β-update amounts to solving a linear system with positive definite coefficient matrix $X^TX + 2I_p$ and righthand side $X^TY^T(v_1^k - u_1^k) + v_2^k - u_2^k + v_4^k - u_4^k$. Note that $X^TX + 2I_p$ is a $p \times p$ matrix. In our model the p denotes the number of genes; p can be prohibitively large and the inverse of $X^TX + 2I_p$ is to be a challenge. As we show below, an appropriate use of numerical linear algebra can exploit this fact and substantially improve performance.

Theorem 1 (Woodbury Matrix Identity): The following identity holds:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}, \tag{26}$$

where A, U, C and V are matrices of sizes $n \times n$, $n \times k$, $k \times k$ and $k \times n$.

Using the theorem above, we can translate the inverse of $X^TX + 2I_p$ to following form

$$(X^TX + 2I_p)^{-1} = \frac{1}{2}I_p - \frac{1}{4}X^T(I_n + \frac{1}{2}XX^T)^{-1}X \tag{27}$$

by setting $A = 2I_p$, $U = X^T$, $C = I_n$, $V = X$; resulting in solving an inverse of $n \times n$ matrix instead of $p \times p$ matrix. Now the β-update can be solved efficiently.

In the same vein, the v_4-update in (23) can be solved similarly, which has a closed form

$$v_4^{k+1} = (D^TD + I_p)^{-1}\left(D^T(v_3^{k+1} + u_3^k) + \beta^{k+1} + u_4^k\right), \tag{28}$$

where the inverse of $D^TD + I_p$ can be translated to

$$(D^TD + I_p)^{-1} = I_p - D^T(I_{p-1} + DD^T)^{-1}D. \tag{29}$$

But $I_{p-1} + DD^T$ is a $(p-1) \times (p-1)$ matrix, it turns out that the conversion can not lend the same incredibly efficient operation as (27). Nevertheless, take the structure of $I_{p-1} + DD^T$ into account, a fast algorithm based on discrete Fourier transform (DFT) can be proposed to solve the inverse of $D^TD + I_p$. The detail can be found in Appendix.

Because of the non-smoothness of $\mathcal{L}$, a closed form solution associated with v_1-update can not be given directly. To address this issue, the proximal operator [56], [57] is introduced; indeed, ADMM is a special case of the proximal point algorithm.

Definition 1 (Proximal operator): Let $f : \mathbf{R}^n \to \mathbf{R} \bigcup \{+\infty\}$ be a closed convex function, the proximal operator $\mathbf{prox}_f : \mathbf{R}^n \to \mathbf{R}^n$ of f is defined by

$$\mathbf{prox}_f(v) = \arg\min_x(f(x) + \tfrac{1}{2}||x - v||_2^2), \tag{30}$$

while a proximal operator of the scaled function λf where $\lambda > 0$ can be expressed as

$$\mathbf{prox}_{\lambda f}(v) = \arg\min_x(f(x) + \tfrac{1}{2\lambda}||x - v||_2^2). \tag{31}$$

The function minimized on the righthand side is strongly convex and not everywhere infinite, so it admits a unique solution for every $v \in \mathbf{R}^n$. In proximal algorithms, the base operation is evaluating the proximal operator of a function, which involves solving a small convex optimization problem. These subproblems can be solved with standard methods, but they often admit closed form solutions or can be solved very

quickly with simple specialized methods. The proximal operator corresponding to huberized hinge loss has the form [44]

$$\mathbf{prox}_{\tau L}(t) = \begin{cases} t, & \text{if } t > 1 \\ \dfrac{t + \tau/\delta}{1 + \tau/\delta}, & \text{if } 1 - \delta - \tau \le t \le 1 \\ t + \tau, & \text{if } t < 1 - \delta - \tau, \end{cases} \tag{32}$$

then the v_1-update has an elementwise closed form

$$[v_1^{k+1}]_i = \mathbf{prox}_{L/\rho}\left([YX\beta^{k+1} + u_1^k]_i\right). \tag{33}$$

Even though the L_1 norm is not differentiable, a simple closed form solution to such problem can be computed by using subdifferential calculus. We list v_3- and v_2-update with a closed form blow

$$\begin{aligned} v_3^{k+1} &= S_{\lambda_2/\rho}(Dv_4^k - u_3^k), \\ v_2^{k+1} &= S_{\lambda_1/(\lambda_3+\rho)}(\beta^{k+1} + u_2^k), \end{aligned} \tag{34}$$

where the soft thresholding operator S is defined as

$$S_\kappa(a) = \begin{cases} a - \kappa & a > \kappa \\ 0 & |a| \le \kappa \\ a + \kappa & a < -\kappa. \end{cases} \tag{35}$$

Indeed, soft thresholding is the proximity operator of the L_1 norm.

The necessary and sufficient optimality conditions for the ADMM algorithm are primal feasibility and dual feasibility. As Boyd *et al.* [43] addressed that u-update dual feasibility always holds and the primal residual $r^{k+1} = Ax^{k+1} + Bz^{k+1} - c$ and z-update dual residual $s^{k+1} = \rho A^T B(z^{k+1} - z^k)$ at iteration $k+1$ in (16) converge to zero as ADMM proceeds. Then a reasonable stopping criterion for our model can be proposed, i.e.,

$$\begin{aligned} r^k &= Ax^k + Bz^k \quad \text{and } ||r^k||_2 \le \epsilon^{pri}, \\ s^k &= \rho A^T B(z^k - z^{k-1}) \quad \text{and } ||s^k||_2 \le \epsilon^{dual}, \end{aligned} \tag{36}$$

where A, B, x, z are variables in (21) while $\epsilon^{pri} > 0$ and $\epsilon^{dual} > 0$ are feasibility tolerances for the primal and dual feasibility, respectively. The tolerances can be chosen using an absolute and relative criterion, such as

$$\begin{aligned} \epsilon^{pri} &= \sqrt{n + 3p - 1}\epsilon^{abs} + \epsilon^{rel} \max\{||Ax^k||_2, ||Bz^k||_2\}, \\ \epsilon^{dual} &= \sqrt{2p - 1}\epsilon^{abs} + \epsilon^{rel}\rho||A^T u^k||_2, \end{aligned} \tag{37}$$

where $\epsilon^{abs} > 0$ and $\epsilon^{rel} > 0$ are absolute and relative tolerance, respectively.

III. RESULTS AND DISCUSSION

In this section, we examine our model on the datasets described in the previous section and present the results based on both classification accuracy and interpretability. Since the binary classification case, the Receiver Operating Characteristic (ROC) curve or the area under curve (AUC) is adopted to evaluate the performance of classifiers. In order to approach comprehensive evaluation, not only the performance for individual classifier but also the biological significance for model is taken into account. We compare the performance between OFSSVM and linear SVM, EN-SVM [45], HHSVM [33] and fused SVM [34] which are all solved by ADMM, although the variable splitting and the optimization steps vary slightly from OFSSVM. All experiments are carried out on Matlab 2014b platform.

A. DETERMINE WHETHER HAVE A PROSTATE CANCER

Splitting the prostate tumor data set into a training set and a test set, where the test set contained 15 normal samples and 16 tumor samples, accounting for nearly 30% of the entire sample, we train all the classifiers on the training set and choose the parameters by the procedure of 10-fold cross validation. Typically, ϵ^{pri} and ϵ^{dual} are both chosen as 10^{-3} for the stopping criterion of ADMM iteration, and set $\rho = 1$, $\delta = 0.2$ in (22)(23) and (32), respectively. We use this as default setting if no explicit setting refers in the following experiments. Test the classifiers on the test set, and count the classification accuracy. Surprisingly, all the classifiers we compare here achieve the same accuracy of 96.77%; only one sample is misclassified. A more amazing fact is that the linear SVM is the best one among these classifiers as shown in Figure 1. It turns out that EN-SVM, HHSVM and fused SVM have an identical ROC curve as OFSSVM, and are all considered to be suboptimal compared with linear SVM.

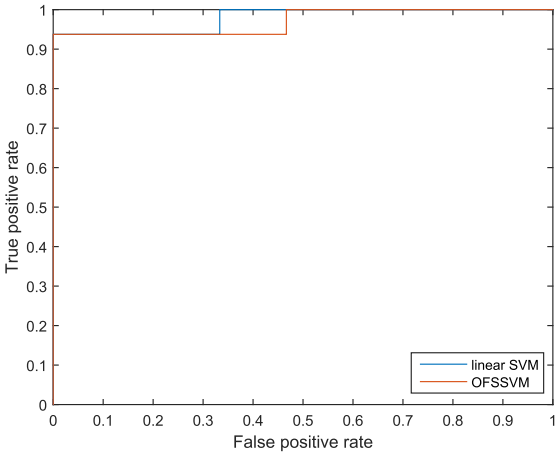


FIGURE 1. ROC for classification between OFSSVM and linear SVM. The linear SVM and OFSSVM achieve a AUC of 0.9792 and 0.9708, respectively. The EN-SVM, HHSVM and fused SVM have a identical curve as OFSSVM.

Rank the features similar to Guyon *et al.* [58] described in SVM-RFE, that arrange all features according to the magnitude of corresponding coefficients ignoring the sign, and then classify the test samples with limited number of features that rank ahead among others. Figure 2 shows the classification accuracy as a function of number of features. Linear SVM exhibits a tendency that the classification accuracy increases monotonically with the number of features. It shows that the linear SVM has poor interpretive capability since no feature selection occurs; linear SVM can not reach a high accuracy with several features that really contribute to the response. In order to get a highest accuracy, it uses even more

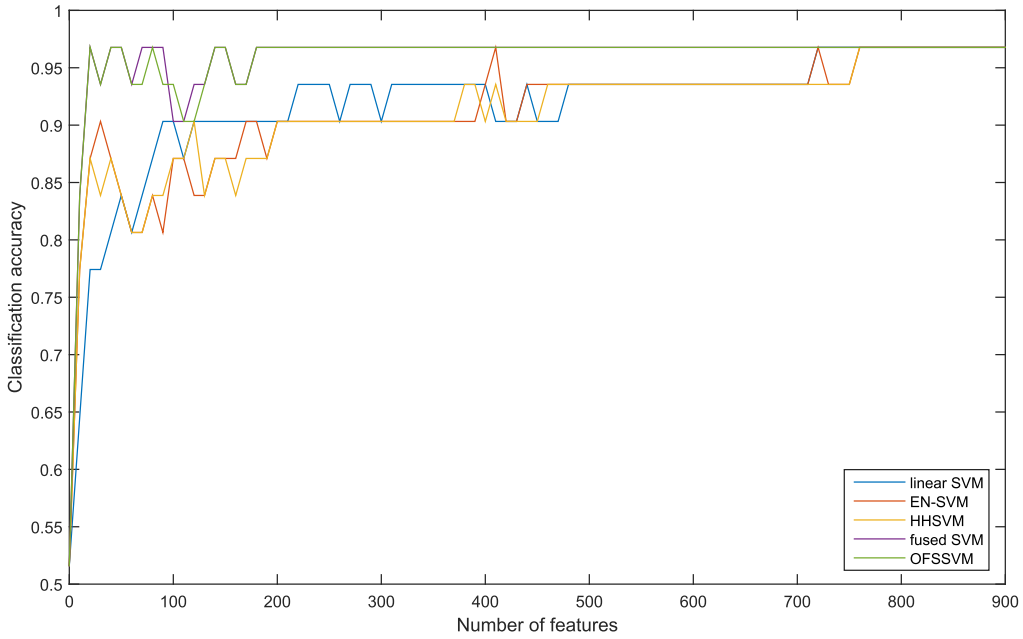


FIGURE 2. The curve of classification accuracy for different number of features. All features are ranked according to the magnitude of corresponding coefficients ignoring the sign for different classifiers. Train each classifier with the limited features that rank ahead, and plot classification accuracy as a function of number of features.

than 700 genes. Conversely, both OFSSVM and fused SVM use only 20 genes to achieve the highest accuracy revealing excellent interpretation ability which is of great significance for pathologist to investigate genes which are related to the cancer and make the choice of targeted therapy. Due to the application of L_1 regularization, EN-SVM and HHSVM also have an ability of feature selection, resulting in an compromise interpretability between OFSSVM and linear SVM.

B. CLASSIFICATION BETWEEN AML AND ALL

Applied our model to AML/ALL dataset without standardizing the pridictors, comparisons between OFSSVM and linear SVM, EN-SVM, HHSVM and fused SVM are presented in Table 2.

TABLE 2. Performance comparison between different classifiers.

Classifier	OFSSVM	linear SVM	EN-SVM	HHSVM	fused SVM
Accuracy(%)	97.06	97.06	97.06	97.06	97.06
AUC	0.9929	0.9964	0.9929	0.9964	0.9964

It can be seen that all classifiers get the same classification accuracy in this case. Moreover, the OFSSVM which we proposed have the same performance as EN-SVM and is considered as a suboptimal model comparing with linear SVM, HHSVM and fused SVM in view of AUC. But when take the interpretability into account, it turn out that the OFSSVM is superior to others, which is illustrated in Table 3.

Due to the high precise of Matlab, none of coefficient is forced to be exactly zero. We use a threshold T to bound the features which contribute to the output and classify the test samples based on the limited features. Since an identical threshold applied to different classifiers will lead to different number of features, and makes it hard to interpret the ability of classification, we only set a threshold to OFSSVM and test other classifiers with the same limited number of features as the former. As shown in Table 3, when a threshold $T = 1 \times 10^{-2}$ is chosen, the OFSSVM reaches the maximal classification accuracy 97.06% with 90 genes, so does the HHSVM and fused SVM. The EN-SVM is only a bit lower compared with the previous three classifiers, reaching an accuracy of 94.12%. Meanwhile, the linear SVM has an accuracy of 64.71%, which is considered meaningless. In fact, linear SVM is the worst classifier among the others in consideration of interpretability just as discussed in Section III-A. When set $T = 3 \times 10^{-2}$, the OFSSVM uses only 14 genes to achieve an accuracy of 82.35%, showing obvious advantages over other classifiers since the EN-SVM, HHSVM and fused

TABLE 3. Interpretability comparison between different classifiers. Setting different thresholds for the OFSSVM can result in different numbers of features. Test all classifiers using the same number of features that have ranked.

T	feature	Accuracy(%) for				
		OFSSVM	linear SVM	EN-SVM	HHSVM	fused SVM
3×10^{-2}	14	82.35	58.82	76.47	76.47	79.41
2×10^{-2}	32	88.24	58.82	88.24	85.29	85.29
1×10^{-2}	90	97.06	64.71	94.12	97.06	97.06
9×10^{-3}	96	97.06	64.71	97.06	97.06	94.12
8×10^{-3}	101	97.06	64.71	97.06	97.06	97.06
7×10^{-3}	116	97.06	73.53	97.06	94.12	97.06
6×10^{-3}	127	97.06	73.53	97.06	97.06	97.06
5×10^{-3}	148	97.06	73.53	94.12	97.06	94.12

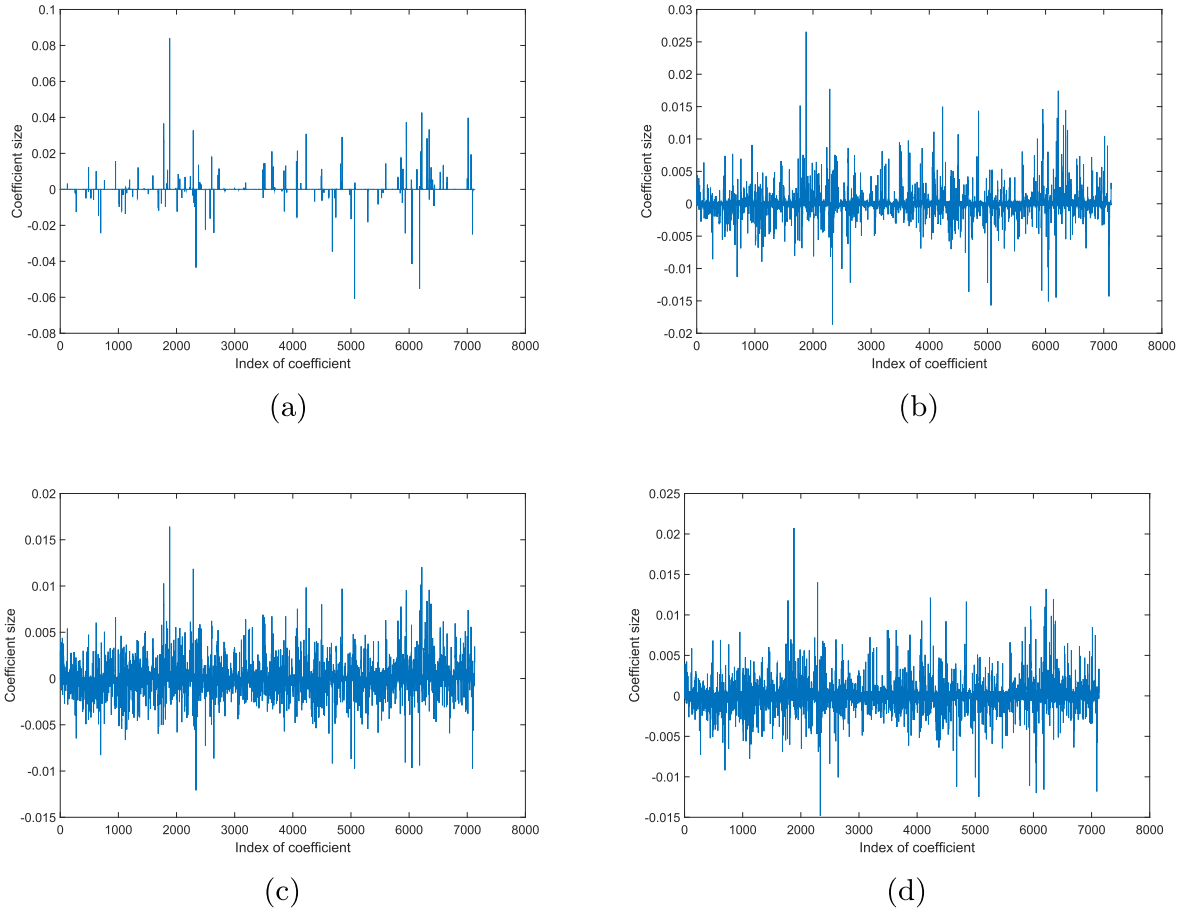


FIGURE 3. Coefficient for different classifiers when applied to AML/ALL dataset. (a) Coefficient of OFSSVM. (b) Coefficient of fused SVM. (c) Coefficient of EN-SVM. (d) Coefficient of HHSVM.

SVM only get an accuracy of 76.47%, 76.47% and 79.41%, respectively. This is of great significance when investigate the most helpful individual gene.

Moreover, it also shows that the OFSSVM is capable of resulting stable features; the classification accuracy tends to keep stable generally when more features are added to the model. On the contrary, the EN-SVM, HHSVM and fused SVM are more easily twisted in such case as shown in Table 3, that their accuracy drop from 97.06% to 94.12% when more features are supplied. This is the cause of smoothness of the coefficient vector, which can be intuitively illustrated in Figure 3. It can be seen that the OFSSVM yields more sparse and smooth coefficient due to it averages the adjacent features, whereas the fused SVM, EN-SVM and HHSVM comtain more noise. Indeed, it gives insight into how the OFSSVM results more true helpful features.

C. MULTICLASS CLASSIFICATION

Since the OFSSVM has a good performance in binary classification, we extend it to multiclass classification by using one-versus-all (OVA) approach to investigate its effectiveness in multiclass cancer diagnosis.

Table 4 shows the prediction results from nine different classification methods, of which the results of method 1 to 8 come from the work of Hastie *et al.* [50]. The method 1 is a regularized version of the diagonal-covariance form of linear discriminant analysis (LDA), which can automatically drops out features that are not contributing to the class predictions. The term 'multinomial' in method 6, 7 and 8 refers to multinomial logistic regression model. In each case, the regularization parameter has been chosen to minimize the cross-validation error, and the test error at that value of the parameter is shown. Setting a threshold 10^{-3}, it can be seen that although the OFSSVM yields a bit higher CV errors than L_1-penalized multinomial, it achieves better results in the

TABLE 4. Prediction results for GCM dataset.

Methods	CV errors out of 144	Test errors out of 54	Number of genes used
1. Nearest shrunken centroids	35	17	6520
2. L_2-penalized discriminant analysis	25	12	16063
3. Support vector classifier	26	14	16063
4. Lasso regression (one vs all)	30.7	12.5	1429
5. k-nearest neighbors	41	26	16063
6. L_2-penalized multinomial	26	15	16063
7. L_1-penalized multinomial	17	13	269
8. Elastic-net penalized multinomial	22	11.8	384
9. OFSSVM (one vs all)	19	10	254

test set when uses less genes, which is more important in the classification problems.

IV. CONCLUSIONS

The emergence of abundant gene expression data has made precision medicine practical, and the flourishing development of artificial intelligence, such as machine learning, has made it possible to discover the link between genetic data and cancer. The use of classification models for cancer predictions, such as the diagnosis of subtypes of cancer or cancer, is a pioneering application of precision medicine. At the same time, simple and effective feature selection methods are helpful for understanding the origin of cancer, and provide a theoretical basis for pathologists to research on cancer and make the choice of targeted therapies.

In this paper, we come up with a new classification model named oriented feature selection SVM (OFSSVM) for cancer prediction using gene expression data. Taking the natural order of genes into account, the OFSSVM uses the fused lasso and elastic net as regularization for linear SVM which uses huberized hinge loss as the loss function, resulting in sparse and smooth solution. We demonstrate its effectiveness in binary classification (i.e. determining whether have a cancer or the subtypes of cancer) and multiclass classification (i.e. determining multiple cancer types together) in the sense of comprehensive evaluation that not only the classification accuracy but also the interpretability is considered. The experiments show that the OFSSVM is an appealing compromise between interpretability and classification accuracy, and is superior to other traditional methods in the sense of comprehensive evaluation, which will promote the development and application of precision medicine.

Due to the nature of the elastic network, OFSSVM can provide automatic feature selection while fused lasso controls the smoothness of the coefficient vector. This saves a lot of time compared to classifiers that take feature selection as an independent step, but it also brings some problems — although features have been selected, they do not provide the same reliability as forward- and backward-stepwise selection. How to keep the reliability and make the model simple is the next work.

APPENDIX

$I_{p-1}+DD^T$ is a real symmetric tridiagonal matrix with those elements located at main diagonal are all equal to 3 and elements lie on the first diagonal below or above the main diagonal are all equal to -1, while the rest elements are all zero. Provide that $p = 6$, then

$$I_{p-1}+DD^T=\begin{pmatrix}3&-1&0&0&0\\-1&3&-1&0&0\\0&-1&3&-1&0\\0&0&-1&3&-1\\0&0&0&-1&3\end{pmatrix}.$$

Note that it is almost a circulant matrix. Define $A \in \mathbf{R}^{(p-1)\times(p-1)}$ whose elements are all zero except $(1, p-1)$ and $(p-1, 1)$ entries which are equal to -1. Then $I_{p-1}+DD^T+A$ is exactly a circulant matrix which can be diagonalized as

$$I_{p-1}+DD^T+A=U\Lambda U^H=F\Lambda F^{-1},$$

where $U=\frac{1}{\sqrt{p-1}}F$ is a unitary matrix, $F \in \mathbf{C}^{(p-1)\times(p-1)}$ is discrete Fourier Transform (DFT) matrix and $\Lambda = diag(Fx)$ is a diagonal matrix with that x is the first column of $I_{p-1}+DD^T+A$. It seems like that the product Fx would require $O(p^2)$ operations; nevertheless, it is possible to compute the DFT Fx in only $O(p\log p)$ operations by a fast Fourier transform (FFT) algorithm. Then the inverse of $I_{p-1}+DD^T+A$ can be solved efficiently, which has a form as follow

$$(I_{p-1}+DD^T+A)^{-1}=U\Lambda^{-1}U^H=\frac{1}{p-1}\mathbf{fft}(\Lambda^{-1})F^H,$$

where $\mathbf{fft}(\cdot)$ denotes FFT operation.

Labeled the inverse of $I_{p-1}+DD^T+A$ as B and define $I_{p-1}+DD^T=M$, then

$$(M+A)^{-1}(M+A)=B(M+A)=I_{p-1}.$$

Multiply both sides by M^{-1}, and perform some algebraic operations, then

$$(I_{p-1}-BA)M^{-1}=B.$$

It can be shown that $I_{p-1}-BA$ is a "N" shape matrix with those elements lie on main diagonal are all equal to 1 except the $(1, 1)$ and $(p-1, p-1)$ entries. Suppose that $B \in \{b_{i,j}\}^{(p-1)\times(p-1)}$, $b_{i,j} \in \mathbf{R}$, then

$$I_{p-1}-BA=\begin{pmatrix}1+b_{1,n}&0&0&\dots&0&b_{1,1}\\b_{2,n}&1&0&\dots&0&b_{2,1}\\b_{3,n}&0&1&\dots&0&b_{3,1}\\\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\b_{n-1,n}&0&0&\dots&1&b_{n-1,1}\\b_{n,n}&0&0&\dots&0&1+b_{n,1}\end{pmatrix},$$

where $n = p-1$.

Given a same structure "N" shape matrix C as above,

$$C=\begin{pmatrix}a_{1,1}&0&0&\dots&0&a_{1,n}\\a_{2,1}&1&0&\dots&0&a_{2,n}\\a_{3,1}&0&1&\dots&0&a_{3,n}\\\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\a_{n-1,1}&0&0&0&1&a_{n-1,n}\\a_{n,1}&0&0&\dots&0&a_{n,n}\end{pmatrix},$$

suppose C is invertible, then it can be shown that the inverse of C has following form

$$C^{-1}=\begin{pmatrix}d_{1,1}&0&0&\dots&0&d_{1,n}\\d_{2,1}&1&0&\dots&0&d_{2,n}\\d_{3,1}&0&1&\dots&0&d_{3,n}\\\vdots&\vdots&\vdots&\ddots&\vdots&\vdots\\d_{n-1,1}&0&0&0&1&d_{n-1,n}\\d_{n,1}&0&0&\dots&0&d_{n,n}\end{pmatrix},$$

where

$$d_{i,1} = \frac{1}{\det C} \times \begin{cases} a_{n,n}, & \text{if } i = 1 \\ -a_{n,1}, & \text{if } i = n \\ a_{i,n}a_{n,1} - a_{i,1}a_{n,n}, & \text{otherwise.} \end{cases}$$

$$d_{i,n} = \frac{1}{\det C} \times \begin{cases} -a_{1,n}, & \text{if } i = 1 \\ a_{1,1}, & \text{if } i = n \\ a_{i,1}a_{1,n} - a_{i,n}a_{1,1}, & \text{otherwise.} \end{cases}$$

$$\det C = a_{1,1}a_{n,n} - a_{1,n}a_{n,1}.$$

Now, the inverse of $I_{p-1} + DD^T$ can be given as

$$M^{-1} = (I_{p-1} - BA)^{-1}B,$$

where both $(I_{p-1} - BA)^{-1}$ and B can be solved efficiently; result in an efficient method for solving $(D^T D + I_p)^{-1}$.

ACKNOWLEDGMENT

The authors would like to appreciate all anonymous reviewers for their insightful comments and constructive suggestions to polish this paper.

REFERENCES

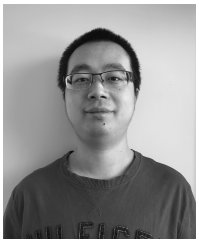
[1] *Precision Medicine Initiative | the White House*. [Online]. Available: https://obamawhitehouse.archives.gov/node/333101
[2] C. Ingolf, "Significance of pharmacogenomics in precision medicine," *Clin. Pharmacol. Therapeutics*, vol. 103, no. 5, pp. 732–735, 2018.
[3] G. L. Semenza, "Targeting HIF-1 for cancer therapy," *Nature Rev. Cancer*, vol. 3, no. 10, pp. 721–732, 2003.
[4] J. Downward, "Targeting RAS signalling pathways in cancer therapy," *Nature Rev. Cancer*, vol. 3, no. 1, pp. 11–22, 2003.
[5] A. H. Bild *et al.*, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, 2006.
[6] T. M. Allen, "Ligand-targeted therapeutics in anticancer therapy," *Nature Rev. Cancer*, vol. 2, no. 10, pp. 750–763, 2002.
[7] J. S. Ross, E. A. Slodkowska, W. F. Symmans, L. Pusztai, P. M. Ravdin, and G. N. Hortobagyi, "The HER-2 receptor and breast cancer: Ten years of targeted anti–HER-2 therapy and personalized medicine," *Oncologist*, vol. 14, no. 4, pp. 320–368, 2009.
[8] L. Brannon-Peppas and J. O. Blanchette, "Nanoparticle and targeted systems for cancer therapy," *Adv. Drug Del. Rev.*, vol. 56, no. 11, pp. 1649–1659, 2004.
[9] J. C. W. Edwards *et al.*, "Efficacy of B-cell–targeted therapy with rituximab in patients with rheumatoid arthritis," *New England J. Med.*, vol. 350, no. 25, pp. 2572–2581, 2004.
[10] M. R. Green, "Targeting targeted therapy," *New England J. Med.*, vol. 350, no. 21, pp. 2191–2193, 2004.
[11] L. H. Hurley, "DNA and its associated processes as targets for cancer therapy," *Nature Rev. Cancer*, vol. 2, no. 3, pp. 188–200, 2002.
[12] R. Bumgarner, "Overview of DNA microarrays: Types, applications, and their future," *Current Protocols Mol. Biol.*, vol. 101, no. 1, pp. 1–11, 2013.
[13] *Next-Generation Sequencing vs. Microarrays | Gen*. [Online]. Available: https://www.genengnews.com/gen-articles/next-generation-sequencing-vs-microarrays/4689
[14] Y. Ji, Y. Si, G. A. McMillin, and E. Lyon, "Clinical pharmacogenomics testing in the era of next generation sequencing: Challenges and opportunities for precision medicine," *Expert Rev. Mol. Diagnostics*, vol. 18, no. 5, pp. 1–11, 2018.
[15] Y. Lu and J. Han, "Cancer classification using gene expression data," *Inf. Syst.*, vol. 28, no. 4, pp. 243–268, 2003.
[16] A. Bhola and A. K. Tiwari, "Machine learning based approaches for cancer classification using gene expression data," *Mach. Learn. Appl., Int. J.*, vol. 2, nos. 3–4, pp. 1–12, 2015.
[17] Z. Jagga and D. Gupta, "Machine learning for biomarker identification in cancer research—Developments toward its clinical application," *Personalized Med.*, vol. 12, no. 4, pp. 371–387, 2015.
[18] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.
[19] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informat.*, vol. 2, pp. 59–77, Jan. 2006.
[20] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.
[21] D. T. Saleh, A. Attia, and O. Shaker, "Studying combined breast cancer biomarkers using machine learning techniques," in *Proc. IEEE 14th Int. Symp. Appl. Mach. Intell. Inform. (SAMI)*, Jan. 2016, pp. 247–251.
[22] J. Jeon *et al.*, "A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening," *Genome Med.*, vol. 6, 2014, Art. no. 57.
[23] M. Q. Ding, L. J. Chen, G. F. Cooper, J. D. Young, and X. Lu, "Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics," *Mol. Cancer Res.*, vol. 16, no. 2, pp. 269–278, 2017.
[24] A. Bashiri, M. Ghazisaeedi, R. Safdari, L. Shahmoradi, and H. Ehtesham, "Improving the prediction of survival in cancer patients by using machine learning techniques: Experience of gene expression data: A narrative review," *Iranian J. Public Health*, vol. 46, no. 2, pp. 165–172, 2017.
[25] C. Chandrasekar and P. Meena, "Microarray gene expression for cancer classification using fast extreme learning machine with ANP," *Int. J. Eng. Res. Appl.*, vol. 2, no. 2, pp. 229–235, 2012.
[26] L. Abdel-Ilah and H. Sahinbegovic, "Using machine learning tool in classification of breast cancer," in *CMBEBIH*, A. Badnjevic, Ed. Singapore: Springer, 2017, pp. 3–8.
[27] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3465–3469, 2009.
[28] S. Begum, D. Chakraborty, and R. Sarkar, "Cancer classification from gene expression based microarray data using SVM ensemble," in *Proc. Int. Conf. Condition Assessment Techn. Elect. Syst. (CATCON)*, 2015, pp. 13–16.
[29] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, 2009.
[30] S. Student and K. Fujarewicz, "Stable feature selection and classification algorithms for multiclass microarray data," *Biol. Direct*, vol. 7, no. 1, p. 33, 2012.
[31] C.-M. Lai, W.-C. Yeh, and C.-Y. Chang, "Gene selection using information gain and improved simplified swarm optimization," *Neurocomputing*, vol. 218, pp. 331–338, Dec. 2016.
[32] Y. Chen, Z. Zhang, J. Zheng, Y. Ma, and Y. Xue, "Gene selection for tumor classification using neighborhood rough sets and entropy measures," *J. Biomed. Inform.*, vol. 67, pp. 59–68, Mar. 2017.
[33] L. Wang, J. Zhu, and H. Zou, "Hybrid huberized support vector machines for microarray classification and gene selection," *Bioinformatics*, vol. 24, no. 3, pp. 412–419, 2008.
[34] F. Rapaport, E. Barillot, and J.-P. Vert, "Classification of arrayCGH data using fused SVM," *Bioinformatics*, vol. 24, no. 13, pp. i375–i382, 2008.
[35] S. Rosset and J. Zhu, "Piecewise linear regularized solution paths," *Ann. Statist.*, vol. 35, no. 3, pp. 1012–1030, 2007.
[36] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 67, no. 1, pp. 91–108, 2005.
[37] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B (Stat. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.
[38] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Statist.*, vol. 1, no. 2, pp. 302–332, Dec. 2007.
[39] G.-B. Ye and X. Xie, "Split bregman method for large scale fused lasso," *Comput. Statist. Data Anal.*, vol. 55, no. 4, pp. 1552–1569, 2011.
[40] J. Liu, L. Yuan, and J. Ye, "An efficient algorithm for a class of fused lasso problems," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2010, pp. 323–332.
[41] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Ann. Statist.*, vol. 39, no. 3, pp. 1335–1371, 2011.
[42] B. Xin, Y. Kawahara, Y. Wang, and W. Gao, "Efficient generalized fused lasso and its application to the diagnosis of alzheimer's disease," in *Proc. 28th AAAI Conf. Artif. Intell. (AAAI)*, CA, USA: AAAI Press, 2014, pp. 2163–2169.

[43] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
[44] T. Watanabe, D. Kessler, C. Scott, M. Angstadt, and C. Sripada, "Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine," *NeuroImage*, vol. 96, pp. 183–202, Aug. 2014.
[45] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statist. Sinica*, vol. 16, no. 2, pp. 589–615, 2006.
[46] D. Singh *et al.*, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.
[47] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.
[48] S. Ramaswamy *et al.*, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 26, pp. 15149–15154, 2001.
[49] T. H. G. James, D. Witten, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, NY, USA: Springer-Verlag, 2013.
[50] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.
[51] G. Wahba, Y. Lin, and H. Zhang, "GACV for support vector machines," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 2000, pp. 297–311.
[52] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 1998, pp. 82–90.
[53] S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu, "ℓ_1 regularization in infinite dimensional feature spaces," in *Learning Theory*, N. H. Bshouty and C. Gentile, Eds. Berlin, Germany: Springer, 2007, pp. 544–558.
[54] D. Gabay, "Chapter ix applications of the method of multipliers to variational inequalities," in *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems* (Studies in Mathematics and Its Applications), vol. 15, M. Fortin and R. Glowinski, Eds. Amsterdam, The Netherlands: Elsevier, 1983, pp. 299–331.
[55] J. Eckstein and D. Bertsekas, "On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 3, pp. 293–318, Jun. 1992.
[56] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, vol. 49. New York, NY, USA: Springer, 2009, pp. 185–212.
[57] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014.
[58] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 389–422, 2002.

YANG SHEN is currently pursuing the degree in software engineering with the School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include data mining and bioinformatics.

CHUNXUE WU received the Ph.D. degree in control theory and control engineering from the China University of Mining and Technology, Beijing, China, in 2006. He is currently a Professor with the Computer Science and Engineering and Software Engineering Division, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include wireless sensor networks, distributed and embedded systems, wireless and mobile systems, and networked control systems.

CONG LIU received the Ph.D. degree in computer application technology from East China Normal University, Shanghai, China, in 2013. He is currently a Lecturer with the Computer Science and Engineering and Software Engineering Division, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China. His research interests include pattern recognition, artificial intelligence, and machine learning.



YAN WU received the Ph.D. degree in environmental chemistry and ecotoxicology from Southern Illinois University, Carbondale. He is currently a Post-Doctoral Associate with the School of Public and Environmental Affairs, Indiana University, Bloomington. His research interests involve elucidations of environmental fate of contaminants using chemical and computational techniques, predictions of their associated effects on wildlife and public health, and data processing and analysis in environmental related fields.

NAIXUE XIONG received the Ph.D. degree in sensor system engineering from Wuhan University and the Ph.D. degree in dependable sensor networks from the Japan Advanced Institute of Science and Technology. Before he attended Northeastern State University, he was with Georgia State University, Wentworth Technology Institution, and Colorado Technical University, about 10 years. He is currently an Associate Professor (third year) with the Department of Mathematics and Computer Science, Northeastern State University Tahlequah, Tahlequah, OK, USA. His research interests include cloud computing, security and dependability, parallel and distributed computing, networks, and optimization theory.

He has published over 280 international journal papers and over 120 international conference papers. Some of his works were published in the IEEE JSAC, IEEE or ACM Transactions, ACM Sigcomm Workshop, IEEE INFOCOM, ICDCS, and IPDPS. He has received the Best Paper Award from the 10th IEEE International Conference on High Performance Computing and Communications in 2008 and the Best Student Paper Award from the 28th North American Fuzzy Information Processing Society Annual Conference in 2009. He was the general chair, the program chair, the publicity chair, a PC member, and a OC member of over 100 international conferences. He was a Reviewer of about 100 international journals, including the IEEE JSAC, IEEE SMC (Park: A/B/C), the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON MOBILE COMPUTING, and the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. He is serving as an Editor-in-Chief, Associate Editor, or Editor Member for over 10 international journals, including an Associate Editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: SYSTEMS, an Associate Editor for Information Science, an Editor-in-Chief of the *Journal of Internet Technology* and the Journal of *Parallel & Cloud Computing*, and a Guest Editor for over 10 international journals, including *Sensor* journal, WINET, and MONET.

Dr. Xiong is the Chair of Trusted Cloud Computing' Task Force, IEEE Computational Intelligence Society, http://www.cs.gsu.edu/~cscnxx/index-TF.html, and the Industry System Applications Technical Committee, http://ieee-cis.org/technical/isatc/. He is a Senior Member of IEEE Computer Society.