Published in final edited form as:

Artif Intell Med. 2019 May ; 96: 134–141. doi:10.1016/j.artmed.2019.04.004.

Sparse Support Vector Machines with L_0 Approximation for Ultra-high Dimensional Omics Data

Zhenqiu Liu^{a,*}, David Elashoff^b, and Steven Piantadosi^c

^aDepartment of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA

^bDepartment of Medicine, University of California at Los Angeles, CA 90024, USA

^cSamuel Oschin Cancer Center, Cedars-Sinai Medical Center, Los Angeles, CA, 90048, USA

Abstract

Omics data usually have ultra-high dimension (p) and small sample size (n). Standard support vector machines (SVMs), which minimize the L_2 norm for the primal variables, only lead to sparse solutions for the dual variables. L_1 based SVMs, directly minimizing the L_1 norm, have been used for feature selection with omics data. However, most current methods directly solve the primal formulations of the problem, which are not computationally scalable. The computational complexity increases with the number of features. In addition, L_1 norm is known to be asymptotically biased and not consistent for feature selection.

In this paper, we develop an efficient method for sparse support vector machines with L_0 norm approximation. The proposed method approximates the L_0 minimization through solving a series of L_2 optimization problems, which can be formulated with dual variables. It finds the optimal solution for p primal variables through estimating n dual variables, which is more efficient as long as the sample size is small. L_0 approximation leads to sparsity in both dual and primal variables, and can be used for both feature and sample selections. The proposed method identifies much less number of features and achieves similar performances in simulations. We apply the proposed method to feature selections with metagenomic sequencing and gene expression data. It can identify biologically important genes and taxa efficiently.

Keywords

SVM; L_0 Approximation; Ultra-high Dimensional Data; Feature Selection; Classification; Metagenomics Sequencing

*Corresponding author: zliu@phs.psu.edu (Zhenqiu Liu).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declaration of interest:
None

Introduction

With the advances of biotechnologies, massive omics data are available in the public domain and are continuing to grow at a fast pace. The ultimate goal with such big omics data is to gain better insights into the biological, environmental, and behavioral influences on a disease, and disease treatment and prevention will take into account individual variability in genes, environment, and lifestyle for each person. This is the main theme of the precision medicine initiatives [1, 2]. However, even though there are important patterns and rich biological information in the data, we still do not fully know how to discover them. Mining such unprecedented big omics data requires efficient computational tools.

Feature (or variable) selection is a core concept in machine learning and important for overcoming overfitting and improving model performance. Generally speaking, feature selection falls into 3 categories: filter, wrapper and embedded methods. Filter approaches including statistical tests identify feature subsets based on a proxy measure of the data, independent of statistical models [3, 4]. One weakness of such methods is that it examines each feature in isolation, ignoring the combined effects among the selected features. Wrapper methods, on the other hand, wrap around a particular learning algorithm to assess the selected feature subsets in terms of the estimated classification errors and to build the final classifier [5, 6]. Wrapper methods can notably reduce the number of features and significantly improve the classification accuracy. However, one drawback with wrapper approaches is their computational intensity. Different from the other two, embedded methods such as L_1 sparse modeling perform variable selection as part of the statistical learning procedure. They are more computationally efficient than wrapper with similar predictive performance, and have drawn much of the attention recently in the literature [7, 8, 9].

Support vector machines (SVMs) have been an efficient tool for classification and pattern recognition with many applications in different fields [10, 11, 12, 13, 14, 15, 16]. SVM is a constrained margin maximization problem based on the risk minimization principle [17, 18, 19]. It is formulated into a convex quadratic programming and can be solved efficiently. However, standard SVM, minimizing the hinge loss function and L_2 norm, only leads to sparsity for the dual variables, but not the primal variables [20, 21, 22]. To handle the big omics data problem with a lot of features, support vector machines with other penalties including L_1 and elastic net have been proposed for feature selection and prediction [23, 24, 25, 26, 27, 28]. However, these methods are directly dealing with the primal variables. They are not computationally scalable when the number of features (genes) is large, which is common in big omics data. In recent years, more efficient algorithms for SVM with alternating direction method of multipliers (ADMM) have been proposed [29, 30]. However, ADMM for SVM still optimizes the primal variables directly, which is computationally costly. It may take hours for feature selection with more than 5,000 variables on a regular PC. More recently, deep metric learning (DML) algorithms have been developed for learning the similarity of the samples [31]. DML methods aim to learn a set of nonlinear transformations and projections explicitly through exploiting the architecture of neural networks. They are more computationally efficient than kernel-based methods and have found applications in imaging classifications and visual understanding. However, such methods usually perform dimension reduction by nonlinear projection, and are not designed

for identifying features in the original feature space. While in medical researches, one important goal is to identify a small number of genes (features) for potential therapeutic targets.

In this paper, we propose a novel SVM method for feature selection and prediction with L_0 approximation. L_0 is an essential sparsity measure and the estimators from L_0 regulation is asymptotic unbiased [32]. The L_0 norm counts the number of nonzero parameters. When used in feature selection, minimizing L_0 forces the small parameters below certain threshold to zero and keeps the large parameters unchanged, leading to sparse solutions. L_0 optimization is NP-hard. We proposed an approximation method for L_0 optimization through a series of standard SVMs, and each SVM is optimized with a dual formulation. Dual optimization is a much smaller problem with n variables, which can be solved efficiently. The primal variables can also be updated explicitly at each iteration. The sparse solution is achieved when both primal and dual variables converge. Computational results with simulations indicate that L_0 SVM outperforms L_1 SVM in that L_0 SVM selects less number of features and achieves a similar predictive accuracy. When applied to high-dimensional big omics data, biologically significant taxa and genes with strong associations can be identified. The proposed method is easy to implement and software L_0 SVM in MATLAB can be downloaded from our web site.

L_0 Regularized Support Vector Machines

SVM basics and related works

Given a training data with n samples $\mathbf{x}_i \in R^p$, with corresponding labels $y_i \in \{-1, 1\}$, for $i = 1, \dots, n$, the 1-norm soft margin is maximized for a non-separable data. This leads to the following primal optimization problem:

$$\min_{\theta_0, \theta_i} \sum_{i=1}^n \max \{0, [1 - y_i(\theta_0 + \theta^t x_i)]\} + \frac{\lambda}{2} \|\theta\|_2^2 \quad (1)$$

The dual lagrangian formulation can be presented as the following quadratic programming:

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t \mathbf{x}_j \\ \text{subject to:} \quad & \sum_{i=1}^n y_i \alpha_i = 0; \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n, \quad (2) \end{aligned}$$

where C is a predefined regularized parameter that controls the number of support vectors with $\alpha_i = 0$. The out-of-sample performance is related to the VC-dimension [17]. Assuming that the data lie in a ball of radius B , and the hypothesis set consists of hyperplanes of width γ (the margin), the VC-dimension is $O(B^2/\lambda^2)$, the out-of-sample

error is bounded by the in-sample error and the VC-dimension. Unfortunately, the proposed method can only be used to choose the number of support vectors, and can not be used to select the number of features.

In recent years, there are various versions of L_1 related SVMs proposed for feature selection.

$$\min_{\theta_0, \theta_i} \sum_{i=1}^n \max \{0, [1 - y_i(\theta_0 + \theta^t \mathbf{x}_i)]\} + \lambda \|\theta\|_1 \quad (3)$$

One simple approach for L_1 SVM is to solve the following linear programming directly [24]:

$$\min_{\theta_0, \theta, \xi, t_i} \sum_{i=1}^n \xi_i + \lambda \sum_{j=1}^p z_j$$

$$\text{subject to: } y(\theta_0 + \theta^t \mathbf{x}) \geq 1 - \xi_i,$$

$$-z_j \leq \theta_j \leq z_j,$$

$$\xi_i \geq 0 \text{ and } z_j \geq 0,$$

$$\forall i = 1, \dots, n, \text{ and } j = 1, \dots, p.$$

We will compare the L_1 SVM with our proposed L_0 SVM.

The proposed method with L_0 approximation

Given a non-trivial training data $S \in (R^p \times \{-1, 1\})^n$, where $\mathbf{x}_i \in R^p$, $i = 1, \dots, n$, and $y_i \in \{-1, 1\}$, let θ be the primal variable, L_0 SVM can be formulated as

$$\min_{\theta, \xi, \theta_0} \frac{1}{2} \|\theta\|_0 + C \sum_{i=1}^n \xi_i$$

$$\text{subject to: } y_i(\theta^t \mathbf{x}_i + \theta_0) \geq 1 - \xi_i, \quad i = 1, \dots, n$$

$$\xi_i \geq 0, \quad i = 1, \dots, n, \quad (4)$$

where $\|\theta\|_0 = \sum_{j=1}^p I(\theta_j \neq 0)$ is the number of nonzero parameters, and for a fixed θ , the optimal slack variables are $\xi_i = \max(0, 1 - y_i(\theta^T \mathbf{x}_i + \theta_0))$. Defining $\frac{0}{0} = 0$, equation (1) can be approximated by the following two-equation iterative optimization problem:

$$\min_{\theta, \xi, \theta_0} \frac{1}{2} \left\| \frac{\theta}{\eta} \right\|_2^2 + C \sum_{i=1}^n \xi_i$$

$$\text{subject to: } y_i(\theta^T \mathbf{x}_i + \theta_0) \geq 1 - \xi_i, \quad i = 1, \dots, n$$

$$\xi_i \geq 0, \quad i = 1, \dots, n, \quad (5)$$

and

$$\eta = \theta, \quad (6)$$

where

$$\left\| \frac{\theta}{\eta} \right\|_2^2 = \sum_{j=1}^p \frac{\theta_j^2}{\eta_j^2}.$$

For fixed η , the Lagrangian for Equation (5) is given by

$$L(\theta, \theta_0, \xi, \alpha, \beta) = \frac{1}{2} \left\| \frac{\theta}{\eta} \right\|_2^2 + C \sum_{i=1}^n \xi_i - \dots - \sum_{i=1}^n \alpha_i [y_i(\theta^T \mathbf{x}_i + \theta_0) - 1 + \xi_i] - \sum_{i=1}^n \beta_i \xi_i, \quad (7)$$

under the constraints $\alpha_i, \beta_i \geq 0, i = 1, \dots, n$. The KKT conditions for Equation (4) are

$$\frac{\partial L(\theta, \theta_0, \xi, \alpha, \beta)}{\partial \theta} = \theta \odot \eta^{-2} - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i = 0,$$

$$\Rightarrow \theta - \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i \odot \eta^2 = 0,$$

$$\frac{\partial L(\theta, \theta_0, \xi, \alpha, \beta)}{\partial \xi_i} = C - \alpha_i - \beta_i = 0, i = 1, \dots, n,$$

$$\frac{\partial L(\theta, \theta_0, \xi, \alpha, \beta)}{\partial \theta_0} = \sum_{i=1}^n \alpha_i y_i = 0,$$

where $\oslash$ represents elementary division, and $\odot$ indicates elementary multiplication. In addition, the KKT complementary conditions:

$$\alpha_i [\theta^T \mathbf{x}_i + \theta_0 - 1 + \xi_i] = 0,$$

$$\beta_i \xi_i = 0,$$

$$\xi_i (\alpha_i - C) = 0 \quad (\text{since } \beta_i = C - \alpha_i \text{ and } \beta_i \xi_i = 0),$$

$\forall i = 1, \dots, n$. From the KKT condition: $\beta_i = C - \alpha_i$ we get $0 \leq \alpha_i \leq C \forall i$, and $\xi_i (\alpha_i - C) = 0$ indicates that $\mathbf{x}_i$ violates the margin constraint when $\alpha_i = C$

Let

$$D = \begin{bmatrix} \eta_1^2 & 0 & \dots & 0 \\ 0 & \eta_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_p^2 \end{bmatrix}, \text{ and } \Lambda = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \alpha_n \end{bmatrix},$$

$X_\eta = [\mathbf{x}_1 \odot \eta^2, \dots, \mathbf{x}_n \odot \eta^2]^T = XD$ then $\theta = X_\eta^T \Lambda y$. Substituting back into Equation (4), we have the maximization problem with dual variables α .

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T D \mathbf{x}_j$$

$$\text{subject to: } \sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n. \quad (8)$$

Let $k_D(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^t D \mathbf{x}_j$, we have a kernel matrix $K_D = [k_D(\mathbf{x}_i, \mathbf{x}_j)]_{n \times n} = X D X^t$ with parameter matrix D .

After we find α , we then update $\theta = X_\eta^t \Lambda y$ and $\eta = \theta$, the algorithm is as follows:

The L_0 SVM Algorithm:

Given a $\lambda > 0$, $\varepsilon = 1e-6$,

and training data $\{X, y\}$,

Initializing $\theta = [1, \dots, 1]^t$,

While 1,

$$\eta = \theta$$

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t D \mathbf{x}_j$$

$$\text{subject to } \sum_{i=1}^n \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \dots, n.$$

$$\theta = X_\eta^t \Lambda y$$

if $\|\theta - \eta\| < \varepsilon$, Break; End

End

$$sv = \{i | 0 < \alpha_i < C\}.$$

$$\theta_0 = \frac{1}{n_{sv}} \sum_{i \in sv} (y_i - \theta^t \mathbf{x}_i).$$

OUTPUT θ_0, θ, α .

Lemma 1: The proposed algorithm is an approximation of L_0 optimization, the solution of proposed algorithm will reach the L_0 solution when the number of iterations reaches ∞ .

Proof: Given a matrix $X_{n \times p}$ ($n \ll p$) and a threshold γ for the coefficient estimates, L_0 rejects all the coefficient estimates below γ to 0 and keeps the large coefficients unchanged. This is the same as defining a binary vector $s = [\dots, 1, 0, \dots, 1]^t$, with the value of 0 or 1 for each feature, where $s_j = 1$ if the coefficient estimate for that feature is above the threshold γ , and 0 otherwise. Let $S = \text{diag}(s)$ be a matrix with s on its diagonal, we have the selected feature matrix $X_S = XS$. We can build the standard SVM models with the matrix X_S , if we know s in advance. For instance, we can estimate the coefficients of a SVM with L_2 regulation given X_S and Y with

$$\begin{aligned} & \max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \mathbf{x}_i^t S \mathbf{x}_j \\ & \text{subject to } \sum_{i=1}^n \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \dots, n, \quad (9) \end{aligned}$$

Defining kernel $K_S = X_S X_S^t = X S S X^t = X S X^t = X_S X_S^t$, and then, $\theta = X_S^t \Lambda y$. It is guaranteed that the estimate is 0 for feature j with $s_j = 0$. However, in reality we do not know s . Estimating both s and θ is a NP-hard problem, since we need to solve a mixed-integer optimization problem. Comparing equation (9) with equation (8), it is clear that S is replaced by D and a binary s_j is approximated by a continuous η_j^2 in proposed algorithm. Therefore the proposed methods is a L_0 approximation.

Recalling the iterative system in equations (5) and (6), note that each feature is penalized by a different penalty, which is inversely proportional to the squared magnitude of that parameter estimator η_j . i.e.,

$$\lambda_j = \frac{1}{2C\eta_j^2}, \text{ and } \eta_j = \theta_j.$$

Smaller θ_j will lead to larger λ_j . A tiny θ_j will become smaller and θ_j will be getting larger in each iteration of L_0 SVM algorithm. $\theta_j \rightarrow 0$ and $\lambda_j \rightarrow \infty$, when the number of iteration goes to ∞ . On the other hand, a larger θ_j will lead a finite λ_j and nonzero θ_j . The solution of L_0 SVM will converge to that of equation (9), because the effect of nonzero η_j will be canceled out in equation (8).

Although the L_0 optimization problem is not convex, the proposed algorithm approximates this non-convex optimization problem with sequential convex optimization. We can easily demonstrate that it is guaranteed to converge to a local optima. Given the solution of first iteration solved by standard linear SVM: $\theta^{(1)} = [\theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_p^{(1)}]^t$, the second iteration leads to the update of the parameters $\theta^{(2)} = [\theta_1^{(2)}, \theta_2^{(2)}, \dots, \theta_p^{(2)}]^t$. There are two different conditions for each parameter θ_j : $|\theta_j^{(1)}| > |\theta_j^{(2)}|$ or $|\theta_j^{(1)}| \leq |\theta_j^{(2)}|$.

(I). $\forall j$, with $|\theta_j^{(1)}| > |\theta_j^{(2)}|$, we have

$$|\theta_j^{(1)}| > |\theta_j^{(2)}| > |\theta_j^{(3)}| > \dots > |\theta_j^{(r)}| > \dots,$$

because

$$\lambda_j^{(1)} = \frac{1}{C|\theta_j^{(1)}|^2} < \lambda_j^{(2)} = \frac{1}{C|\theta_j^{(2)}|^2} < \dots < \lambda_j^{(r)} < \dots,$$

with the proposed algorithm. Note that,

$$\frac{|\theta_j^{(r)}|}{|\theta_j^{(1)}|} = \frac{|\theta_j^{(2)}|}{|\theta_j^{(1)}|} \frac{|\theta_j^{(3)}|}{|\theta_j^{(2)}|} \dots \frac{|\theta_j^{(r)}|}{|\theta_j^{(r-1)}|}.$$

Let $q_j = \max_k \frac{|\theta_j^{(k)}|}{|\theta_j^{(k-1)}|} < 1, \forall k = 1, \dots, r$, we have

$$\lim_{r \rightarrow \infty} |\theta_j^{(r)}| < \lim_{r \rightarrow \infty} |\theta_j^{(1)}| q_j^{r-1} = 0.$$

(II). $\forall j$, with $|\theta_j^{(1)}| \leq |\theta_j^{(2)}|$ we would like to show that the algorithm converges to a finite solution, i.e., $\exists$ a finite $k, \theta_j^{(k)} = \theta_j^{(k-1)}$.

Based on the proposed algorithm, we have,

$$|\theta_j^{(1)}| \leq |\theta_j^{(2)}| \leq |\theta_j^{(3)}| \leq \dots \leq |\theta_j^{(r)}| \leq \dots,$$

because

$$\lambda_j^{(1)} = \frac{1}{C|\theta_j^{(1)}|^2} \geq \lambda_j^{(2)} = \frac{1}{C|\theta_j^{(2)}|^2} \geq \dots \geq \lambda_j^{(r)} \geq \dots$$

Let $\theta_j^{(k)} \neq \theta_j^{(k-1)}, \forall k$ and j with $\theta_j^{(k-1)} \leq \theta_j^{(k)}$, so $q_j = \min_k \frac{|\theta_j^{(k)}|}{|\theta_j^{(k-1)}|} > 1$, similarly, we have

$$\lim_{r \rightarrow \infty} |\theta_j^{(r)}| > \lim_{r \rightarrow \infty} |\theta_j^{(1)}| q_j^{r-1} = \infty.$$

However, all estimated parameters from Equation (5) and (6) are bounded above from the solution of the following linear programming with all $\lambda_j = 0$:

$$\min_{\theta, \xi, \theta_0} \sum_{i=1}^n \xi_i$$

$$\text{subject to } y_i(\theta^T \mathbf{x}_i + \theta_0) \geq 1 - \xi_i, \quad i = 1, \dots, n$$

$$\xi_i \geq 0, \quad i = 1, \dots, n, \quad (10)$$

The problem in Equation (10) is convex and has a finite solution with very mild conditions. Hence, we conclude that there must exist a finite $k, \theta_j^{(k)} = \theta_j^{(k-1)}$. Therefore, the proposed L_0 SVM forces the small parameters goes to zero, and keeps the large parameters unchanged, leading to sparse solutions. Q.E.D.

Nonlinear Extension: Although the proposed algorithm is proposed for linear SVM, it is possible to be extended to a nonlinear SVM, as long as the kernel can be represented explicitly as $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^t \phi(\mathbf{x}_j)$. A nonlinear kernels can be generated through mapping the input into high-order feature space $\mathbf{x}_i \mapsto \phi(\mathbf{x}_i)$, and then compute the dot product $\phi(\mathbf{x}_i)^t \phi(\mathbf{x}_j)$. For instance, given $\phi(\mathbf{x}_i) = [x_{i1}, \dots, x_{ip}, x_{i1}x_{i2}, \dots, x_{ij}x_{ik}, \dots, x_{i(p-1)}x_{ip}]^t$, we can construct a non-linear polynomial kernel, and apply L_0SVM to detect not only the associated features but also the interactions among them. However, mapping the features and computing the inner product explicitly is usually not a trivial task and needs a lot of computations.

The complexity of the Algorithm and C Determination: The complexity of the L_0SVM algorithm is similar to the standard linear SVM, since it finds the optimal solution through solving a series of standard linear SVM. The number of iterations is usually less than 100 in practice. Therefore, the complexity of the L_0SVM is $O(n^2)$ with a state-of-the-art linear SVM solver, where n is the size of the training data.

The free parameter C controls the sparsity of the model. The smaller the C , the less features are selected. The free parameter C will be determined by 5-fold cross-validation with the training data only. We split the data into 5 approximately equal parts, build the model with 4

parts, and determine the optimal C with the smallest test error $E = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{2n}$ of the

remaining part. The performance of the model is evaluated with the test error or the area under the ROC curves (AUCs) of an independent test data in all the simulations.

Computational Results

Simulations

The simulation data was generated from two normal distributions with the first 5 features. We generate $n = 100$ samples for each class and different number of features ($p = 100$ and 1000). The two classes have different means with the mean difference of 0.4 for the first 5 features, and the same mean (0) for the rest of the features. The following common correlation structure was used for data generation.

$$\Sigma = \begin{bmatrix} 1 & r & r^2 & \dots & r^{|p-1|} \\ r & 1 & r & \dots & r^{|p-2|} \\ r^2 & r & 1 & \dots & r^{|p-3|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r^{|p-1|} & r^{|p-2|} & r^{|p-3|} & \dots & 1 \end{bmatrix},$$

where different r s indicate different correlation structures. Our simulations were performed with $r = 0, 0.3$, and 0.6 , respectively. The free-parameter C for L_0SVM , and λ for L_1SVM are estimated with 5-fold cross-validation, and the performance of the models is evaluated with an independent test data. The performance of T-test for feature selection is assessed with the Benjamini & Hochberg (FDR) correction with adjusted P-value of 0.05. The computational results of L_0SVM , L_1SVM and T-test are reported in Table 1.

Table 1 shows that T-test with Benjamini & Hochberg correction is the most conservative method. With different correlation structures, only around 4 features are selected when $p = 100$, and less than 2 out of 5 true features are identified on average, since the number of false negative (FN) is greater than 3. T-test performed the worst when the dimension is large ($p = 1000$). On average, less than 1 feature is selected, and the chance of selected one of the 5 true features is very low, since the number of FN is at least 4.26, indicating that T-test fails to detect true features when the dimension is high. When comparing L_0 SVM and L_1 SVM, while the predictive errors are similar, the number of features selected are quite different. The number of features selected with L_0 SVM is less than 9 features and at least 3 out of 5 true features are selected when $p = 100$, while the number of features identified with L_1 SVM is more than 19 with slightly better number of FN. Particularly, when $p = 1000$, both the total number of selected features and the number of false positive (FP) features for L_0 SVM are much smaller than L_1 SVM with different correlation structures. While less than 11 features are selected with L_0 SVM, more than 66 features are identified with L_1 SVM, and the number of FP is greater than 62 with slightly smaller number of FN, indicating that L_1 SVM chooses more features than necessary. Reducing the number of false positive features is crucial, since it is costly to validate a potential gene signature biologically and clinically. One interesting observation is that the number of features identified with L_1 SVM increases with the number of dimensions (p), while similar number of features are selected with different p s with L_0 SVM. Therefore, L_0 SVM outperforms L_1 SVM in that it selects less number of features with the similar predictive (test) error. It also performs T-test in that its number of FN is smaller than T-test with Benjamini & Hochberg correction, particularly when the data dimension p is higher.

Simulation with Ultra-high Dimensional Data and Different Sample Sizes

With the sample sizes of $n = 500$ and 1000 , and the dimensions of $p = 10,000, 20,000$, and $30,000$, the number of true features of 5, and the same correlation structure with $r = 0.3$ and 0.6 . Again, the mean difference between classes for the first 5 features is 0.4, and 0 for the remaining features. L_1 SVM fails to solve the problem in days, because of the unscalable linear programming, while the proposed L_0 SVM solves the problems with ultra-high dimension and different sample sizes efficiently. The free parameter C is determined by 5-fold cross-validation, and the test area under ROC curves (AUCs) is evaluated with an independent test data. The computational results are reported in Figure 1.

Figure 1 demonstrates the performance of L_0 SVM with different correlation structures, sample size and dimensions. First, the test AUCs improve as the sample sizes get larger. For instance, given $r = 0.3$, when increasing the sample sizes from 500 to 1000, the test AUCs increase from 0.74, 0.73 and 0.68 to 0.76, 0.74 and 0.76, respectively, with the corresponding dimensions of $p = 10000, 20000$ and 30000 . Similar results can also be achieved with $r = 0.6$. Moreover, even though similar test AUCs are achieved with different correlation structures, the predictive AUCs seem to get better with higher correlations among the features. As demonstrated with the right panels of Figure 1, with $n = 1000$, the test AUCs improve from 0.76, 0.74 and 0.76 to 0.84, 0.88 and 0.84, respectively, with the corresponding data dimensions of $p = 10000, 20000$ and 30000 . One interesting observation is that the test AUCs are getting worse as the data dimension increases from 10000 to 30000

with a smaller sample size $n = 500$, but there is no statistically significant difference among the test AUCs with different number of dimensions with a large sample size $n = 10000$. Finally, more than 3 out of 5 true features are correctly identified with different sample sizes, correlation structures and data dimensions, indicating the high efficiency of L_0 SVM with large sample size and ultra-high dimensional data.

Real Metagenomic Sequencing Data

Inflammatory bowel disease (IBD), including both Crohn's disease (CD) and ulcerative colitis (UC), is a heterogeneous disease that affects over 1.4 million people in the U.S. Changes in the intestinal bacterial composition play a major role in the pathogenesis and the clinical course of IBD. The real metagenomic data set was first published by Tong et al. 2013 [33]. There are total 299 samples with 76 inflammatory bowel disease (IBD) subjects and 223 non-IBD subjects. Two hundred eighty five (285) out of 299 samples (72 IBDs and 213 controls) have the 16S metagenomic data available. After sequencing alignment, we have total 5648 OTUs available. We merged OTUs at the genus and species level, and then genera and species with high abundances were selected for further study after discarding those taxa with less than two reads on average. We further normalized the data with proportion and arcsine square-root transformation. Taxa selected with L_0 SVM and L_1 SVM at genus and species levels are reported on Table 2.

Table 2 demonstrates that both L_0 SVM and L_1 SVM achieve the similar test AUCs. While L_1 SVM selects 9 genera and 17 species, L_0 SVM identifies 4 IBD associated genera and 6 species with 5-fold cross-validation, which indicates that L_0 SVM is more sparse and selects less features with cross-validation. The normalized relative abundances of the 4 selected genera on the left of Table 2 with L_0 SVM are also shown on Figure 2.

Figure 2 shows that the relative abundances of *Faecalibacterium* and *Ruminococcus* are reduced in IBD comparing to control, while the relative abundance of *Klebsiella* and *Acidaminococcus* are increased in IBD. Both *Faecalibacterium* and *Ruminococcus* have been well studied and are known to be associated with IBD [34, 35, 36]. Both *Faecalibacterium* and *Ruminococcus* have less richness and diversity in IBD, indicating that both are protective microbiota. Particularly, it has been shown both in vitro and vivo that *Faecalibacterium* has an anti-inflammatory effect and Crohn's disease (CD)-associated dysbiosis is characterised by a loss of *Faecalibacterium* [37]. On the other hand, both *Acidaminococcus* and *Klebsiella* are positively associated with IBD, but their associations with IBD have been less studied. More investigations are needed to evaluate the effect of those two 'bad' microbiota. The results on the species level are reported on the right hand side of Table 2. Similar conclusions can also be reached. While L_0 SVM identifies 6 features at the species level, L_1 SVM discovers 17 IBD associated features with similar prediction AUC (0.82), indicating the better performance of L_0 SVM.

Two Other Real Datasets

One dataset is another metagenomic data for obese study downloaded from dbGaP under study ID phs000258. The data and analytical results were first reported by [38]. There were a total of 310 Amish adult samples with 112 males and 198 females. After aligning the 16S

rRNA sequences of gut microbiota to reference sequences and taxonomy databases, there were a total of 240 taxa at the genus level. The body mass indices (BMI) is the clinical phenotype. Among the 310 subjects, we only interested in the two groups with 73 normal (N: BMI < 25) and 138 obese (BMI ≥ 30) samples. We choose 62 out of 240 features that average more than 2 reads for further studies.

The other data is a gene expression profiling for ovarian cancer downloaded from NCBI website (GSE26712) [39]. There are total 185 patient samples with 90 optimal and 95 suboptimal debulking samples. Suboptimal debulking is defined as tumor nodules large then 1cm when operating surgery, and is associated with reduced chemosensitivity and poor survival. There are total 13238 genes. L_1 SVM takes too long to find a solution with such a large number of features. For comparison purpose, we choose 961 genes with $P < 0.05$ using t-test for further study. Our goal is to identify a small number of disease associated features with the proposed methods. We randomly split both data 50 times with 5-fold cross validation, the average number of selected features and test errors are reported in Table 3.

Table 3 demonstrates that L_0 SVM selects a much smaller number of features with different omics datasets. On average, 8 and 6 features are identified with L_0 SVM for obese and ovarian cancer data, respectively, while L_1 SVM chooses too many features, 38 and 118 features are selected with L_1 SVM. It also shows that the test errors of L_0 SVM and L_1 SVM are quite similar with the average test errors around 0.3 for both data, although the test errors of L_1 SVM are slightly smaller. In biological or medical research, validating a potential drug target experimentally is usually very expensive. Correctly identifying a small number of genes is crucial for downstream studies. L_0 SVM outperforms L_1 SVM in selecting a much smaller number of disease associated features. Finally, we also apply L_0 SVM to the original ovarian cancer data with 13238 genes. Four features (genes) are selected with a slightly higher test errors (0.35) for such a high-dimensional data, indicating the efficiency of L_0 SVM.

Conclusions and Discussions

Feature or variable selection is an important task in machine learning and statistics. Sparse SVM models are a popular tool for feature selection and class prediction in big data mining. However, most sparse SVM methods have been developed based on the primal problem, which is unsuitable for problem with small sample size and ultra-high dimensions. In this paper, we propose a novel algorithm for approximating L_0 SVM. Our L_0 SVM method approximates L_0 SVM through iteratively solving multiple standard SVMs. L_0 SVM seeks sparse solutions through optimizing the dual problem with kernel. It is particularly computationally efficient for a big data problem encountered in bioinformatics and computational biology. We demonstrate the proposed methods with simulation and real sequencing data. L_0 SVM outperforms L_1 SVM in that it selects much less features with similar performance in predictive accuracy. When applied to the real taxa and gene identification problem, it can identify disease associated taxa and biological meaningful genes efficiently.

When comparing with filter and wrapper methods, L_1 SVM for feature selection is computationally more efficient than wrapper methods, and more accurate than filter methods such as T-test. It also chooses a group of features simultaneously, and take the combined effect of multiple features into account. However, similar to other embedded methods, L_0 SVM is model-specific and linear. Although it is possible to extend to nonlinear features selection with explicitly high-dimensional feature projection or multiple kernel learning, it is not trivial and requires intensive computations.

Acknowledgment

This work was partially supported by the DMS-1343506 grant from the National Science Foundation (ZL), P01 CA098912–10, P01 DK046763, and UL1 TR0001881–01 from NIH, and Four Diamonds fund. The funders had no role in the preparation of the article.

References

- [1]. Jaffe S, Planning for us precision medicine initiative underway, *Lancet* 385(9986) (2015) 2448–9. [PubMed: 26122056]
- [2]. Ashley E, The precision medicine initiative: a new national effort, *JAMA* 313(21) (2015) 2119–20. [PubMed: 25928209]
- [3]. Tohka J, Moradi E, Huttunen H, Comparison of feature selection techniques in machine learning for anatomical brain mri in dementia, *Neuroinformatics* 14(3) (2016) 279–96. [PubMed: 26803769]
- [4]. Lyu H, Wan M, Han J, Liu R, Wang C, A filter feature selection method based on the maximal information coefficient and gram-schmidt orthogonalization for biomedical data mining, *Comput Biol Med.* 89 (2017) 264–274. [PubMed: 28850898]
- [5]. Mustaqeem A, Anwar S, Majid M, Khan A, Wrapper method for feature selection to classify cardiac arrhythmia, *Conf Proc IEEE Eng Med Biol Soc.* 2017 (2017) 3656–3659. [PubMed: 29060691]
- [6]. Hui K, Ooi C, Lim M, Leong M, Al-Obaidi S, An improved wrapper-based feature selection method for machinery fault diagnosis, *PLoS One* 12(12) (2017) e0189143. [PubMed: 29261689]
- [7]. Weston J, Elisseeff A, Schoelkopf B, Tipping M, Use of the zero-norm with linear models and kernel methods, *Journal of Machine Learning Research* 3 (2003) 14391461.
- [8]. W. Z Liu Q, Gu Q, Feature selection method based on support vector machine and shape analysis for high-throughput medical data, *Comput Biol Med* 91 (2017) 103–111. [PubMed: 29049908]
- [9]. Shen Y, Liu Z, Ott J, Support vector machines with l1 penalty for detecting gene-gene interactions, *Int J Data Min Bioinform.* 6(5) (2012) 463–70. [PubMed: 23155775]
- [10]. Gaonkar B, Shinohara R, Davatzikos C, Interpreting support vector machine models for multivariate group wise analysis in neuroimaging, *Med Image Anal.* 24(1) (2015) 190–204. [PubMed: 26210913]
- [11]. Tran Vdu T, Tempel S, Zerath B, Zehraoui F, Tahi F, mirboost: boosting support vector machines for microrna precursor classification, *RNA* 21(5) (2015) 775–85. [PubMed: 25795417]
- [12]. Brayet J, Zehraoui F, Jeanson-Leh L, Israeli D, Tahi F, Towards a pirna prediction using multiple kernel fusion and support vector machine, *Bioinformatics* 30(17) (2014) i364–70. [PubMed: 25161221]
- [13]. Heikamp K, Bajorath J, Support vector machines for drug discovery, *Expert Opin Drug Discov.* 9(1) (2014) 93–104. [PubMed: 24304044]
- [14]. Chen H, Chen L, Support vector machine classification of drunk driving behaviour, *Int J Environ Res Public Health* 14(1) (2017) pii: E108.
- [15]. Polat H, Danaei Mehr H, Cetin A, Diagnosis of chronic kidney disease based on support vector machine by feature selection methods, *J Med Syst.* 41(4) (2017) 55. [PubMed: 28243816]

- [16]. Seyres D, Darbo E, Perrin L, Herrmann C, Gonzalez A, Ledpred: an r/bioconductor package to predict regulatory sequences using support vector machines, *Bioinformatics* 32(7) (2016) 1091–3. [PubMed: 26628586]
- [17]. Cortes C, Vapnik V, Support-vector networks, *Machine Learning* 20(3) (1995) 273–297.
- [18]. Nedaie A, Najafi A, Support vector machine with dirichlet feature mapping, *Neural Netw.* 98 (2018) 87–101. [PubMed: 29223012]
- [19]. Khemchandani R, Goyal K, Chandra S, Twsvr: Regression via twin support vector machine, *Neural Netw.* 74 (2016) 14–21. [PubMed: 26624223]
- [20]. Jaggi M, An equivalence between the lasso and support vector machines, in: *International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: Theory and Applications(ROKS)*, Leuven, Belgium, 2013.
- [21]. Burden F, Winkler D, Relevance vector machines: Sparse classification methods for qsar, *J Chem Inf Model* 55(8) (2015) 1529–34. [PubMed: 26158341]
- [22]. Xia X, Training sparse least squares support vector machines by the qr decomposition, *Neural Netw.* 106 (2018) 175–184. [PubMed: 30075354]
- [23]. Liu Z, Lin S, Tan M, Sparse support vector machine with lp penalty for biomarker discovery, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 7(1) (2010) 100–7.
- [24]. Liu Z, Chen D, Sheng L, Liu Y, Class prediction and feature selection with linear optimization for metagenomic count data, *PLoS One* 8(3) (2013) e53253. [PubMed: 23555553]
- [25]. Tan M, Wang L, Tsang I, Learning sparse svm for feature selection on very high dimensional datasets, in: *ICML*, Haifa, Israel, 2010.
- [26]. Nguyen M, de la Torre F, ptimal feature selection for support vector machines, *Pattern Recognition* 43 (2010) 584–591.
- [27]. Zhou Q, Chen W, Song S, Gardner J, Weinberger K, Chen Y, A reduction of the elastic net to support vector machines with an application to gpu computing, in: *Association for the Advancement of Artificial Intelligence (AAAI)*, Austin Texas, USA, 2015.
- [28]. Li N, Ainsworth R, Wu M, Ding B, Wang W, Miec-svm: automated pipeline for protein peptide/ligand interaction prediction, *Bioinformatics* 32(6) (2016) 940–942. [PubMed: 26568623]
- [29]. Ye G, Chen Y, Xie X, Efficient variable selection in support vector machines via the alternating direction method of multipliers, *JMLR W&CP* 15 (2011) 832–840.
- [30]. Xu Y, Akrotirianakis I, Chakraborty A, Proximal gradient method for huberized support vector machine, *Pattern Analysis and Applications* 19(4) (2016) 989–1005.
- [31]. Lu J, Liong V, Zhou J, Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(59) (2017) 1–1.
- [32]. Liu Z, Li G, Efficient regularized regression with l0 penalty for variable selection and network construction, *Comput Math Methods Med.* 2016 (2016) 3456153. [PubMed: 27843486]
- [33]. Tong M, Li X, Wegener Parfrey L, Roth B, Ippoliti A, Wei B, Borneman J, McGovern D, Frank D, Li E, Horvath S, Knight R, Braun J, A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease, *PLoS One* 8 (2013) e80702. [PubMed: 24260458]
- [34]. Shaw K, Bertha M, Hofmekler T, Chopra P, Vatanen T, Srivatsa A, Prince J, Kumar A, Sauer C, Zwick M, Satten G, Kostic A, Mulle J, Xavier R, Kugathasan S, Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease, *Genome Med.* 8(1) (2016) 75. [PubMed: 27412252]
- [35]. Takahashi K, Nishida A, Fujimoto T, Fujii M, Shioya M, Imaeda H, Inatomi O, Bamba S, Sugimoto M, Andoh A, Reduced abundance of butyrate-producing bacteria species in the fecal microbial community in crohn's disease, *Digestion* 93(1) (2016) 59–65. [PubMed: 26789999]
- [36]. Eppinga H, Sperna Weiland C, Thio H, van der Woude C, Nijsten T, Peppelenbosch M, Konstantinov S, Similar depletion of protective faecalibacterium prausnitzii in psoriasis and inflammatory bowel disease, but not in hidradenitis suppurativa, *J Crohns Colitis* 10(9) (2016) 1067–75. [PubMed: 26971052]

- [37]. Quevrain E, Maubert M, Michon C, Chain F, Marquant R, Tailhades J, Miquel S, Carlier L, Bermdez-Humaran L, Pigneur B, Lequin O, Kharrat P, Thomas G, Rainteau D, Aubry C, Breyner N, Afonso C, Lavielle S, Grill J, Chassaing G, Chatel J, Trugnan G, Xavier R, Langella P, Sokol H, Seksik P, Identification of an anti-inflammatory protein from faecalibacterium prausnitzii, a commensal bacterium deficient in crohn's disease, Gut 65(3) (2016) 415–25. [PubMed: 26045134]
- [38]. Zupancic M, Cantarel B, Liu Z, Drabek E, Ryan K, et al., Analysis of the gut microbiota in the old order amish and its relation to the metabolic syndrome, PLoS One 7(8) (2012) e43052. [PubMed: 22905200]
- [39]. Vathipadiekal V, Wang V, Wei W, Waldron L, et al., Creation of a human secretome: A novel composite library of human secreted proteins: Validation using ovarian cancer gene expression data and a virtual secretome array, Clin Cancer Res 21(21) (2015) 4960–9. [PubMed: 25944803]

Research Highlights

The proposed method is novel.

Big omics data usually have ultra-high dimension (p) and small sample size (n). We proposed a novel L_0 SVM algorithm for handling such problems efficiently. Our contributions include:

- L_0 SVM for simultaneous feature selection and classification
- L_0 SVM outperformed L_1 SVM in that it identifies true features more accurately with less false positive rate
- Implemented with dual optimization and is highly computationally efficient with ultra-high dimensional data ($n \ll p$)
- Identifying biological important features with metagenomic and expression data
- More sparse estimators with less bias

Figure 1:
Performance of L_0 SVM with different dimensions ($p = 10000, 20000$ and 30000 , correlations $r = 0.3$, and 0.7 , and sample sizes of $n = 500$ and 1000).

Figure 2:
The normalized relative abundances of L_0 SVM identified genera in control and IBD.

Table 1:

Performance of L_0 SVM and L_1 SVM, and T-test over 100 simulations, where values in the parenthesis are the standard deviations. Feature selection for T-test is based on Benjamini & Hochberg (FDR)adjusted P-value, where T: T-test; # SF: average number of selected features; #SP: average number of false positive features; #FN: average number of false negative features.

p	Methods	r	#SF	#FP	#FN	Test Error
100	L_0	0.0	9.0(± 5.7)	5.62(± 3.02)	1.62(± 1.14)	0.39(± 0.04)
		0.3	8.2(± 5.6)	4.92(± 5.30)	1.72(± 1.03)	0.37(± 0.04)
		0.6	6.7(± 4.7)	3.06(± 4.65)	1.40(± 1.05)	0.31(± 0.07)
	L_1	0.0	22.1(± 20.5)	17.9(± 20.0)	0.80(± 1.01)	0.39(± 0.04)
		0.3	19.9(± 15.8)	15.9(± 15.4)	1.00(± 0.92)	0.37(± 0.05)
		0.6	27.8(± 18.9)	23.6(± 18.3)	0.78(± 1.04)	0.32(± 0.06)
	T	0.0	3.94(± 13.94)	2.1(± 13.4)	3.16(± 1.31)	NA
		0.3	4.1(± 14.0)	2.38(± 13.4)	3.30(± 1.47)	NA
		0.6	3.84(± 12.0)	1.62(± 11.12)	3.28(± 1.34)	NA
1000	L_0	0.0	10.5(± 3.85)	8.0(± 3.83)	2.42(± 0.88)	0.42(± 0.04)
		0.3	7.9(± 3.98)	5.4(± 3.6)	2.5(± 1.05)	0.40(± 0.05)
		0.6	9.02(± 2.99)	6.80(± 3.01)	2.78(± 1.11)	0.40(± 0.06)
	L_1	0.0	91.7(± 69.9)	88.3(± 69.1)	1.58(± 1.28)	0.44(± 0.04)
		0.3	74.4(± 62.6)	71.1(± 62.2)	1.74(± 1.16)	0.43(± 0.04)
		0.6	66.2(± 58.6)	62.9(± 58.0)	1.68(± 1.36)	0.40(± 0.05)
	T	0.0	0.60(± 0.73)	0.04(± 0.20)	4.4(± 0.73)	NA
		0.3	0.86(± 0.99)	0.12(± 0.33)	4.26(± 0.83)	NA
		0.6	0.66(± 0.94)	0.12(± 0.32)	4.46(± 0.86)	NA

Table 2:Taxa selected with L_0 SVM and L_1 SVM at the genus and species level.

Genus Level		Species Level	
L_0 SVM	L_1 SVM	L_0 SVM	L_1 SVM
Acidaminococcus	Akkermansia	Bacteroidesacidifaciens	Aggregatibactersegnis
Klebsiella	Bacteroides	Clostridiuminnocuum	Bacteroidesacidifaciens
Faecali-bacterium	Faecali-bacterium	Pseudomona-snitroreducens	Bifidobacterium-pseudolongum
Ruminococcus	Haemophilus	Dialisterinvisus	Clostridiuminnocuum
	Klebsiella	Coprococcuseutactus	Clostridiumperfringens
	Others	Serratiamarcescens	Collinsellaerofaciens
	Prevotella		Desulfovibriodesulfurica
	Ruminococcus		Dialisterinvisus
	Sutterella		Enterococcusfaecalis
			Haemophilusparainfluenzae
			Lactobacillusruminis
			Oscillospiraguilliermondii
			Others
			Parabacteroidesdistasonis
			Pseudomonasnitroreducens
			Roseburiafaecis
			Serratiamarcescens
Test AUC = 0.830	Test AUC = 0.823	Test AUC = 0.82	Test AUC = 0.82

Table 3:

Features selection over 50 random partitions with L_0 SVM and L_1 SVM for two omics datasets, where ANF: the average of number of features selected, and Test Error: average test error. The values in the parenthesis are the standard deviation

Datasets	Methods	ANF	Test Error
Obese	L_0 SVM	8.03(± 2.37)	0.290(± 0.007)
	L_1 SVM	37.90(± 4.34)	0.286(± 0.016)
Ovarian Cancer	L_0 SVM	6.0(± 2.74)	0.314(± 0.017)
	L_1 SVM	118.30(± 6.11)	0.298(± 0.018)