The Sparse Learning of The Support Vector Machine

Lipeng Cui, Jie Shen and Song Yao
Tianjin Light Industry Vocational Technical College, Tianjin, 300350, China
*Corresponding author’s e-mail: cuilipeng@tjlivtc.edu.cn

Abstract. The sparse model plays an important role in many areas, such as in the machine learning, image processing and signal processing. The sparse model has the ability of variable selection, so they can solve the over-fitting problem. The sparse model can be introduced into the field of support vector machine in order to get classification of the labels and sparsity of the variables simultaneously. This paper summarizes various sparse support vector machines. Finally, we revealed the research directions of the sparse support vector machines in the future.

1. Introduction
In the era of big data, people are faced with all kinds of high-dimensional data. How to dig out useful information from high-dimensional data is an important issue. Moreover, in the fields of modern machine learning, data mining and bioinformatics, the explanatory variable space of many classification and regression problems is often very dimensional, even super-highly dimensional. However, high-dimensional data can lead to over-fitting in machine learning, which makes the generalization performance of statistical models worse. Therefore, the problem of dimensionality reduction and variable selection in variable space needs to be solved urgently. What is more, the purpose of variable selection lies in two aspects. One is to achieve accurate prediction and classification, and the other is to make the model more interpretable and reduce the complexity of the statistical model. The so-called interpretability refers to the simplicity of the model. Obviously, the lower the dimensionality of the variable space is, the better the interpretability of the statistical model will be. In short, people always expect to use as few variables as possible to achieve higher prediction accuracy.

The sparse model has been extensively studied in recent years, which can generate sparse solutions to automatically realize feature selection. Since Lasso [1] (Lasso) was proposed, a large number of sparse models [2-4] have been proposed one after another. Therefore, Lasso has an extremely important position in the field of variable selection in statistics. However, through experiments and theoretical analysis, scholars have found that Lasso also has various shortcomings, and many scholars have conducted more in-depth researches on these shortcomings of Lasso, one of which is that Lasso also compresses the coefficients of important variables. Sparse models often have the form of "loss function term + penalty function term". For example, the form of Lasso is "least square loss function + L1 norm penalty function". On the basis of Lasso, there are two ways to promote the regularized sparse model: (1) Change the penalty function term; (2) Change the loss function term. However, in the development of regularized sparse models, changing the penalty function term is more popular than changing the loss function term. Moreover, in addition to the problem of variable selection in statistics, sparse models like Lasso have also been applied to many fields such as compressed sensing, signal reconstruction and image reconstruction, which has more and more extensive applications in the fields of bio-statistics, machine learning, data mining, image processing and signal processing.
Support vector machine (SVM) has good classification performance, which has been studied extensively in the past ten years. What is more, the standard $L_2$-norm SVM classifier uses a ridge penalty function, so feature selection cannot be implemented. However, feature selection is very important in many situations. For example, selecting important genes is a very important issue in bioinformatics. In many cases, eliminating noise variables can improve the classification accuracy of the classifier. Therefore, how to achieve class label classification and feature selection at the same time is crucial, so people naturally think of combining sparse model with support vector machine to achieve classification and feature selection at the same time. Moreover, the sparse model and SVM are similar in form, and both have the form of "loss function + penalty function". Besides, the combination of sparse model and support vector machine produces a series of sparse SVMs that can automatically realize feature selection. For example, the $L_1$-norm SVM in the form of "hinge loss function + $L_1$ norm penalty function" and the SCAD SVM in the form of "hinge loss function + SCAD penalty function", etc. Therefore, it can be seen that by introducing the penalty functions of different regularized sparse models into the SVM, SVMs with different functions of feature selection can be obtained.

2. Sparse Support Vector Machine

2.1. $L1$ SVM

The optimization problem required to be solved by the $L1$ norm SVM\[5\] based on the hinge loss function is as follows:

$$
\min_{w,h} \frac{1}{n} \sum_{i=1}^{n} \left[ 1 - y_i \left( b + w^T \cdot x_i \right) \right] + \lambda \|w\|_1
$$

(1)

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD}) \in R^D$ is a sample, $y_i \in \{+1, -1\}$ is the class label, $\lambda$ refers to tuning parameter, and $\lambda > 0$. $\|w\|_1 = \sum_{d=1}^{D} w_d$ is the $L1$ norm penalty, and $[1 - y_i \left( b + w \cdot x_i \right)]_+$ is the hinge loss function. In addition, the $L1$ norm SVM adopts the $L1$ norm penalty, so it has a sparsity effect, which can greatly reduce the dimensionality of the feature space. In many cases, it can achieve high classification accuracy by selecting only a few features. However, the $L1$ norm SVM also has shortcomings, one of which is a biased estimation. The shortcoming stems from the fact that the $L1$ norm penalty imposes the same size penalty on the model vector components of all features. The $L1$ norm SVM is the first sparse SVM in the form of "loss function + penalty function". In addition, replacing the hinge loss function in (1) with the least square loss function will get the $L1$ norm least square support vector machine, and replacing the hinge loss function in (1) with the likelihood function of logistic regression will result in an $L1$ norm logistic classifier, which has the performance of both the classification and feature selection. Therefore, $L1$ SVM is the first sparse support vector machine that combines a sparse model with a support vector machine.

2.2. SCAD SVM

The $L1$ norm SVM with $L1$ norm penalty has the disadvantage of biased estimation. SCAD SVM\[6\] is a sparse SVM proposed for this problem. What is more, SCAD penalty is a non-convex penalty function with unbiased estimation properties and sparse effect proposed by Fan et al., which is introduced into the field of support vector machine classification by Zhang et al. The SCAD penalty is used to replace the $L1$ norm penalty in the objective function of the $L1$ norm SVM optimization problem, and the SCAD SVM based on the hinge loss function is proposed. The optimization problem required to be solved is as follows:

$$
\min_{w,h} \frac{1}{n} \sum_{i=1}^{n} \left[ 1 - y_i \left( b + w^T \cdot x_i \right) \right]_+ + \sum_{d=1}^{D} \phi_{\lambda,d} (|w_d|)
$$

(2)
where $\lambda > 0$, and $\varphi_{\lambda,a} (\cdot)$ is SCAD penalty. The reason why SCAD penalty is of unbiased estimates is that the penalty for large model vector components (i.e. the model vector components of important features) is small, while the penalty for small model vector components (i.e. the model vector components of noise features) is large, which overcomes the shortcomings of biased estimation of L1 norm penalty, making SCAD SVM have Oracle properties. Besides, the penalty function similar to the SCAD penalty is the MC penalty proposed by Zhang et al. [7] (minimax concave penalty). The penalty function also has Oracle properties and the principle is similar to the SCAD penalty. Additionally, SCAD SVM also has a shortcoming that the non-convex SCAD penalty in the objective function of the optimization problem will lead to the difficulty in solving the optimization problem.

2.3. EN SVM

The penalty function of the elastic net support vector machine [8] (EN SVM) is derived from the elastic net, and its penalty function is a combination of the L1 norm penalty with the L2 norm penalty. The optimization problem required to be solved is:

$$
\min_{\mathbf{w}, b} \frac{1}{n} \sum_{i=1}^{n} \left[ 1 - y_i \left( b + \mathbf{w}^T \cdot \mathbf{x}_i \right) \right] + \lambda_1 \sum_{i=1}^{n} \left| \mathbf{w}_i \right| + \lambda_2 \sum_{i=1}^{n} \left( \mathbf{w}_i \right)^2
$$

(3)

where $\lambda_1 > 0$, and $\lambda_2 > 0$. $\left| \mathbf{w}_i \right|$ is the L1 norm penalty, and $\left( \mathbf{w}_i \right)^2$ is the L2 norm penalty. Meanwhile, ENSVM is divided into two types. One is the above formula (3), which is also called DrSVM (doubly regularized support vector machine). EN SVM uses both the L1 norm penalty and the L2 norm penalty. Therefore, the highly correlated features have a grouping effect. The group effect is essentially due to the L2 norm penalty, and the group effect means that the model vector components of the highly correlated features in the solution model vector obtained by DrSVM are almost equal, so DrSVM always tries to select or discard highly correlated features at the same time, which has a group selection effect on highly correlated features. Moreover, it can be seen that DrSVM has the capabilities of feature selection, feature group selection regarding highly correlated features, and classification. In addition, another form of EN SVM is HHSVM[9]. HHSVM also has a group effect. It differs from DrSVM in that the loss function of HHSVM is replaced from the hinge loss function to the Huber hinge loss function.

2.4. AEN SVM

It is assumed that $\tilde{w}_d$ is marginal regression estimator

$$
\tilde{w}_d = \sum_{d=1}^{m} \mathbf{x}_{id} \cdot y_i / \sum_{d=1}^{m} \mathbf{x}_{id}^2
$$

where $d \in \{1, 2, \cdots, D\}$. If $|\tilde{w}_1| \geq |\tilde{w}_2| \geq \cdots \geq |\tilde{w}_d| \geq 0$ and $m$ represents the largest index $d \in \{1, 2, \cdots, D\}$ that makes inequality $|\tilde{w}_d| \geq \delta$ true, and $\delta$ is the artificial threshold, then the optimization problem that the adaptive elastic net support vector machine [10] (AEN SVM) requires to be solved is:

$$
\min_{\mathbf{w}, b} \frac{1}{n} \sum_{i=1}^{n} \left[ 1 - y_i \left( b + \mathbf{w}^T \cdot \mathbf{x}_i \right) \right] + \lambda_1 \left\| \mathbf{\Gamma}_1 \mathbf{w} \right\|_1 + \lambda_2 \left\| \mathbf{\Gamma}_2 \mathbf{w} \right\|_2
$$

(4)

where $\lambda_1 > 0$, $\lambda_2 > 0$, and

$$
\left\| \mathbf{\Gamma}_1 \mathbf{w} \right\|_1 = \sum_{d=1}^{D} \omega_d |\tilde{w}_d|
$$

$$
\left\| \mathbf{\Gamma}_2 \mathbf{w} \right\|_2 = \sum_{d=1}^{D} \omega_d \tilde{w}_d^2
$$

(5)

(6)

Similar to the adaptive L1 SVM, the adaptive elastic network SVM is actually divided into two steps: 1) Calculate the marginal regression estimation $\tilde{w}_d$; 2) Substitute the marginal regression estimation into the objective function of the optimization problem of the adaptive elastic network SVM to solve the problem. Furthermore, the principle of adaptive elastic net SVM is also similar to that of adaptive L1 SVM, whose essence is to use marginal regression estimation to modify the
penalty degree of the L1 norm penalty on the model vector components corresponding to each feature, thereby reducing the compression of larger model vector components and increasing the compression of smaller model vector components. Therefore, the adaptive elastic network SVM also has the properties of oracle. In addition, the loss function is replaced by the huberized hinge loss function, and an adaptive elastic net SVM based on the huberized hinge loss function is proposed in [11].

2.5. PEN SVM

Pairwised Elastic Net support vector machine [12-14] (PEN SVM) is required to solve the optimization problem:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \left[ 1 - y_i (b + w^T \cdot x_i) \right] + \lambda \cdot w^T P w$$

(7)

where $\lambda > 0$, and $P$ is a symmetric positive semi-definite matrix composed of non-negative elements and its elements have the following characteristics: the greater the high correlation between feature $l$ and feature $k$ is, the smaller the corresponding element in matrix $P$ will be; meanwhile, the smaller the high correlation between feature $l$ and feature $k$ is, the larger the element corresponding to $P$ in the matrix will be. The selection of matrix $P$ is completely determined by users and has many forms, one of which is the pairwise elastic net penalty $P = I + 11^T - R$, where the correlation matrix corresponds to the $R$ feature, at this time

$$|w^T P|w| = \|w\|^2 + \|w\|^2 - |w^T R|w|$$

(8)

The regularization term corresponding to element $R_{lk}$ is:

$$R_{lk} (\beta_l^2 + \beta_k^2) + (1 - R_{lk}) (|\beta_l| + |\beta_k|)^2$$

(9)

Obviously, the size of the correlation $R_{lk}$ between feature $l$ and feature $k$ determines the form of the regularization term. When $R_{lk} \to 1$, the regularization term becomes $(\beta_l^2 + \beta_k^2)$, namely the regularization term becomes L2 norm penalty at this time, which has the effect of group selection. When $R_{lk} \to 0$, the regularization term becomes $(|\beta_l| + |\beta_k|)^2$. At this time, the regularization term becomes the square of the L1 norm penalty, which has the property of sparseness. Therefore, the pairwise elastic net SVM has the advantage of adaptively changing the regularization term as the correlation between features changes. The higher the correlation between features is, the greater the proportion of the L2 norm penalty imposed on them in the entire regularization term will be. The lower the correlation between features is, the larger the square of the L1 norm penalty imposed on them in the entire regularization term will be. The idea of this method is similar to that of structured elastic nets. It also introduces the combined Laplacian matrix in the spectrograph theory into the regularization term. If $L$ represents the combined Laplacian matrix corresponding to the graph, and $C = (11^T + L)^\alpha$ is obtained, where $\alpha$ is a positive integer, and the construction method of the element $p_{ij}$ in the $l$th row and $k$ column in $P$ is $p_{lk} = c_{lk} / \sqrt{c_{ll} c_{kk}}$, where $c_{ll}$, $c_{kk}$, and $c_{lk}$ are all elements in matrix $C$. This method of constructing matrix $P$ based on graphs also has the advantage of adaptively changing the regularization terms as the correlation between features changes.

2.6. Fused SVM

The optimization problem to be solved by fused SVM [15] is:

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} \left[ 1 - y_i (b + w^T \cdot x_i) \right] + \lambda_1 \|w\|^2 + \lambda_2 \sum_{d=2}^{D} \|w_d - w_{d-1}\|$$

(10)
where \( \lambda_1 > 0 \) and \( \lambda_2 > 0 \). The penalty function of the fused SVM is a combination of the L1 norm penalty and the fusion penalty, where the L1 norm penalty makes the fused SVM have a sparse solution. Moreover, the fusion penalty penalizes the difference between the model vector components of adjacent features, and makes the model vector components of adjacent features equal, so that the model vector obtained by fused SVM has the characteristic of piece-wise constant.

Fused penalty is proposed because other penalty functions ignore the common characteristics of adjacent variables in ordered variables, and fused SVM superimposes the fusion penalty on the basis of the L1 norm penalty, which not only penalizes the size of the model vector component itself, but also penalizes the absolute value of the difference between the vector components of adjacent models. Therefore, it will not only sparse the solution, but also make the adjacent model vector components change flat (namely, the obtained solution has the characteristics of segmented constant). Moreover, fused SVM is of great significance in practical applications. For example, in the field of bio-medicine, if the number of copies of a gene is significantly greater than that of copies of other genes, then the gene will be likely to be mutated. The identification of these genes with abnormal replication times is of great significance in the identification of the development of human cancer. Besides, in the process of comparative genome hybridization (CGH), the genome is composed of a series of genes arranged in a straight line. Scientists often think that the number of copies of adjacent genes is roughly the same. Therefore, it is hoped to reveal the commonality of adjacent genes in the number of replications, while the solution of L1 SVM only has the effect of sparseness, which obviously cannot reveal the commonality of adjacent genes. The fused SVM is just suitable for dealing with the above situation.

### Table 1. Summary of the form of each sparse SVM

| SVM         | Loss function | Penalty function |
|-------------|---------------|------------------|
| L2 SVM      | Hinge loss    | \( \lambda_1 \| w \|_2^2 \) |
| L1 SVM      | Hinge loss    | \( \lambda_1 \| w \|_1 \) |
| SCAD SVM    | Hinge loss    | \( \sum_{d=1}^{D} \phi_{\lambda,d}(w_d) \) |
| EN SVM      | Hinge loss (DrSVM) or huberized hinge loss (HHSVM) | \( \lambda_1 \| w \|_1 + \lambda_2 \| w \|_2^2 \) |
| AEN SVM     | Hinge loss or huberized hinge loss | \( \lambda_1 \| \Gamma_1 w \|_1 + \lambda_2 \| \Gamma_2 w \|_2^2 \) |
| PEN SVM     | Hinge loss    | \( \lambda \cdot w^\top Pw \) |
| Fused SVM   | Hinge loss    | \( \lambda_1 \| w \|_1 + \lambda_2 \sum_{d=2}^{D} |w_d - w_{d-1}| \) |

### Table 2. Summary of the sparsity of each sparse SVM

| SVM         | Sparsity | Oracle quality |
|-------------|----------|----------------|
| L2 SVM      | Y        | Y              |
| L1 SVM      | Y        | N              |
| SCAD SVM    | Y        | Y              |
| AEN SVM     | Y        | Y              |
| PEN SVM     | Y        | N              |
| Fused SVM   | Y        | N              |
3. Conclusions
Sparse model is a hot research topic in the fields of machine learning and artificial intelligence in recent years. The Lasso based on the L1-norm penalty proposed by Tibshirani can effectively reduce the dimension of variable space but can only achieve distributed variable selection. In many cases, there is a certain structure among variables. For example, in gene microarray analysis, there are often multiple variants in a gene. When identifying the genetic variant that is associated with the disease, it is more reasonable to divide the variants belonging to the same gene into one group. Given this, some scholars conduct variables selection by taking the structure among variables as a priori information. Group Lasso is a sparse model that uses the group structure of variables as a priori information, which can achieve the selection of variable groups. Sparse models such as SCAD, MCP, and adaptive Lasso improve the consistency in variables selection of the Lasso. Elastic Net can select groups of highly correlated variables simultaneously. Models such as Lasso have also been applied to the field of support vector machines (SVMs) to constitute sparse SVMs. In addition to applications in statistical variables selection, Lasso and other sparse models have also been applied in many fields such as compressed sensing, signal reconstruction, and image reconstruction, and have an increasingly wide range of applications in fields such as biostatistics, machine learning, data mining, image processing, and signal processing.

4. Future Work
A wider selection of support vector machines and other loss functions should be developed. The current sparse support vector machines are mostly researched on the basis of a single hinge loss function. In the future, the sparse support vector machines can be expanded to other types of loss functions such as multi-class loss functions and least squares loss functions to achieve classification in different scenarios.

Other penalty functions can be introduced to construct a new sparse support vector machine. In addition to the penalty functions reviewed in this paper, there are many penalty functions that have not been introduced into the support vector machine field to achieve sparseness, such as, $L_{2,1}$ norm penalty with structure sparsity feature and tree structure penalty function, etc.

Statistical properties of sparse support vector machines should be interpreted and proved. Many sparse support vector machines reviewed in this paper do not have rigorous discussion and mathematical proof of statistical properties such as oracle properties.

The Elastic Corr-Net has the feature selection ability and the grouping effect too. Its penalty is the combination of norm penalty and the correlation based penalty

$$
\phi(w) = \sum_{i=1}^{D-1} \sum_{j>i} \left\{ \frac{(w_i - w_j)^2}{1 - \rho_{ij}} + \frac{(w_i + w_j)^2}{1 + \rho_{ij}} \right\}, \text{ where } \rho_{ij} \text{ is the correlation coefficient between two different features, } D \text{ is the number of features, } w_i \text{ and } w_j \text{ are the coefficients of two different features. Obviously, the correlation based penalty is comprised of correlation coefficients between different features. Due to using the values of the correlation coefficients, the Elastic Corr-Net imposes exacter grouping effect on the highly correlated features than the Elastic Net. Thus we suppose that if we introduce the correlation based penalty into the SVM, the SVM with correlation based penalty has exacter grouping effect than the HHSVM, DrSVM and the LOGR SVM.}
$$

The fused penalty in the fused Lasso has been applied to support vector machine classification field and proposes the fused SVM whose solution is piecewise constant. However, the pairwise fused penalty has not been applied to the SVM field. We know that the fused SVM uses the fused penalty and thus it can only make the adjacent coefficients nearly equal. However, the pairwise fused penalty can fuse the coefficients pairwisely and does not just fuse the adjacent coefficients.

The shortcoming of the Elastic Net, HHSVM and DrSVM is that only highly correlated features can be selected as a group simultaneously. In other words, the Elastic Net, HHSVM and DrSVM can not achieve feature group selection about arbitrary features. How to solve this problem is still an open
question. Note that the group Lasso can achieve feature group selection about arbitrary features through dividing the features into different groups manually in advance, so can we apply the norm penalty function of the group Lasso into the support vector machine classification field to achieve classification and feature group selection about arbitrary features simultaneously? In addition, the group SCAD, group bridge and group MCP are nonconvex unbiased feature group selection methods, can we apply them to the support vector machine classification field to achieve the feature group selection about arbitrary features and classification simultaneously? It is worth exploring these questions in future.

The sparse group Lasso uses the norm penalty and norm penalty simultaneously and thus has the ability of feature selection about arbitrary features and feature group selection about arbitrary features simultaneously. If we use the norm penalty and norm penalty simultaneously as the penalty function of SVM, can we achieve feature selection about arbitrary features, feature group selection about arbitrary features and class label classification simultaneously?

References
[1] Tibshirani R. (1996) Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological): 267-288.
[2] Liu Jianwei, Cui Lipeng, Liu Zeyu, Luo Xionglin. (1996) Regularized sparse model. Chinese Journal of Computers, 07: 1307-1325.
[3] Liu Jianwei, Cui Lipeng, Luo Xionglin. (2017) Structure sparse model. Chinese Journal of Computers, 06: 1309-1337.
[4] Liu Jianwei, Cui Lipeng, Luo Xionglin. (2015) Group sparse model and its algorithm review. Acta Electronica Sinica. 04:776-782
[5] Zhu J, Rosset S, Hastie T, et al. (2003) 1-norm Support Vector Machines. In: Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: 49-56.
[6] Zhang H H, Ahn J, Lin X, et al. (2006) Gene selection Using Support Vector Machines with Non-convex Penalty. Bioinformatics, 22(1): 88-95.
[7] Zhang C H. (2010) Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics, 38 (2): 894-942
[8] Wang L, Zhu J, Zou H. (2006) The doubly regularized support vector machine. Statistica Sinica. 16(2): 589-615.
[9] Wang L, Zhu J, Zou H. (2008) Hybrid huberized support vector machines for microarray classification and gene selection. Bioinformatics, 24(3): 412-419.
[10] Jun-Tao L I, Ying-Min J I A. (2010) An Improved Elastic Net for Cancer Classification and Gene Selection. Acta Automatica Sinica, 36(7):976-981.
[11] Li J, Jia Y, Li W. Adaptive huberized support vector machine and its application to microarray classification. Neural Computing and Applications, 2011, 20(1): 123-132.
[12] Lorbert A, Ramadge P J. (2013) The Pairwise Elastic Net support vector machine for automatic fMRI feature selection. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE: 1036-1040.
[13] Lorbert A, Eis D, Kostina V, et al. (2010) Exploiting covariate similarity in sparse regression via the pairwise elastic net. In: International Conference on Artificial Intelligence and Statistics: 477-484.
[14] Slawski M. (2012) The structured elastic net for quantile regression and support vector classification. Statistics and Computing, 22(1): 153-168.
[15] Rapaport F, Barillot E, Vert J P. (2008) Classification of arrayCGH data using fused SVM. Bioinformatics, 24(13): 375-382.