Neural Comput & Applic (2017) 28:969–978
DOI 10.1007/s00521-015-2113-7

REVIEW

An overview on semi-supervised support vector machine

Shifei Ding[1,2] · Zhibin Zhu[1] · Xiekai Zhang[1]

Received: 24 May 2015 / Accepted: 4 November 2015 / Published online: 18 November 2015

Abstract Support vector machine (SVM) is a machine learning method based on statistical learning theory. It has a lot of advantages, such as solid theoretical foundation, global optimization, the sparsity of the solution, nonlinear and generalization. The standard form of SVM only applies to supervised learning. Large amount of data generated in real life is unlabeled, and the standard form of SVM cannot make good use of these data to improve its learning ability. However, semi-supervised support vector machine (S3VM) is a good solution to this problem. This paper reviews the recent progress in semi-supervised support vector machine. First, the basic theory of S3VM is expounded and discussed in detail; then, the mainstream model of S3VM is presented, including transductive support vector machine, Laplacian support vector machine, S3VM training via the label mean, S3VM based on cluster kernel; finally, we give the conclusions and look ahead to the research on S3VM.

Keywords Semi-supervised · Support vector machine · Semi-supervised support vector machine

✉ Shifei Ding
dingsf@cumt.edu.cn

[1] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China

[2] Jiangsu Key Laboratory of Mine Mechanical and Electrical Equipment, China University of Mining and Technology, Xuzhou 221116, China

1 Introduction

In recent decades, the ways of collecting data are more diverse, and the amount of data is also growing. With the rapid development of various technologies, it becomes easy to collect large amounts of data, but most of them have no label. Usually, a huge amount of work is required to label the data. However, the data that people obtain in every day are massive, and this means that it is unfeasible to invest a lot of resources in the work. The generalization ability of a learning system often depends on the number of labeled training samples. If there are small amounts of labeled data, combining labeled and unlabeled data to change or improve the learning behavior would be an interesting work, and the semi-supervised learning [1–3] is supposed to do that.

Support vector machine is a machine learning method based on statistical learning theory. It has a lot of advantages, such as solid theoretical foundation, global optimization, the sparsity of the solution, nonlinear and generalization. And it also shows a good prospect in practical engineering fields [4–8]. The standard form of SVM only applies to supervised learning, which training data must be labeled. However, in many practical problems, unlabeled data are always the majority. It is difficult to get a learner with strong generalization ability by using limited labeled data in supervised learning. When the labeled data are scarce and the unlabeled data are sufficient, semi-supervised learning method can help improve the classification performance of SVM in this situation. Thus, introducing the idea of semi-supervised learning into SVM and combining them will overcome the shortcomings of SVM and get better classification results. Obviously, this will be an important research field. The development of semi-supervised support vector machine is mainly to make support vector machines use small amounts of labeled data and large amounts of unlabeled data for training and classification. Furthermore, those data can be fully utilized to improve the classification accuracy of SVM and increase its generalization performance.

In 1999, Bennett and Demiriz [9] proposed semi-supervised support vector machine. In the next more than 10 years, scholars tried using various semi-supervised learning method to improve S3VM and presented many effective semi-supervised support vector machine algorithm [10, 11]. This paper reviews the latest research progress on the semi-supervised support vector machine algorithm and expatiates on the idea of several existent mainstream semi-supervised support vector machine algorithms and its improved method.

This paper is organized as follows: Sect. 2 discusses the basic theory of SVM. In Sect. 3, we analyze the current research on semi-supervised support vector machine algorithm in detail. Section 4 contains some useful concluding comments.

2 Support vector machine

Support vector machine, proposed by Vapnik, is a machine learning method based on VC dimension and structural risk minimization and is a specific realization for statistical learning theory. Considering from the learning mode, SVM is a supervised learning method.

The basic theory of SVM is to find an optimal classification hyperplane which meets the requirements of classification. As shown below, the solid black dots represent one class of sample, and the white dots represent another one. H is the classification hyperplane. H_1 and H_2 are the plane which contains sample points and have the closest distance with the classification hyperplane. H_1 and H_2 are parallel to H. The distance between H_1 and H_2 is called maximum margin. The optimal classification hyperplane of SVM requires not only separating the samples correctly, but also maximizing the margin. Those sample points contained in H_1 and H_2 are support vector.

Let (x_i, y_i), $i = 1, 2, \ldots, l$, $x \in R^n$, $y \in \{\pm 1\}$ denote the sample dataset, where y_i is the label. Also, let the hyperplane be denoted as $(w \cdot x) + b = 0$. When the two classes are linearly separable, the optimal classification hyperplane can be summarized to solve quadratic programming problems as follows:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}||w||^2 \\ &s.t \quad y_i(w \cdot x_i + b) \geq 1, \quad i = 1, 2, \ldots, l \end{aligned} \tag{1}$$

The formula above is a convex programming problem. A solution to this problem is to use Lagrange function:

$$\begin{aligned} &L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} a_i(y_i(w \cdot x_i + b) - 1) \\ &s.t. \quad a_i \geq 0 \end{aligned} \tag{2}$$

Here, a_i are the corresponding Lagrange multipliers to each sample. The solution of formula (2) is determined by the saddle point. At saddle point, the partial derivative of w and b is zero.

$$\begin{aligned} &\frac{\partial}{\partial w} L(w, b, a) = w - \sum_{i=1}^{l} a_i y_i x_i = 0 \\ &\frac{\partial}{\partial b} L(w, b, a) = \sum_{i=1}^{l} a_i y_i = 0 \end{aligned} \tag{3}$$

Combining (2) with (3), the quadratic programming problem will be transformed into the corresponding dual problem and then maximize the objective function:

$$\begin{aligned} &Max(a) = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i,j=1}^{l} a_i a_j y_i y_j (x_i \cdot x_j) \\ &s.t. \quad a_i \geq 0, \quad i = 1, 2, \ldots l. \quad \sum_{i=1}^{l} a_i y_i = 0 \end{aligned} \tag{4}$$

We can get the optimal solution $a^* = (a_1^*, a_2^*, \ldots, a_l^*)^T$ by solving Eq. (4), so the optimal weight vector w^* and optimal bias b^* can be expressed as:

$$w^* = \sum_{i=1}^{l} a_i^* y_i x_i \tag{5}$$

$$b^* = y_i - w^* \cdot x_i \tag{6}$$

Finally, the optimal classification decision function can be defined as:

$$f(x) = \text{sgn}(w^* \cdot x_i + b^*) \tag{7}$$

The optimal classification hyperplane discussed and solved above is in the linear case. For the nonlinear case, kernel functions which map input vector to high-dimensional feature vector space are used to help SVM construct the optimal classification hyperplane. The optimal classification decision function can be written as:

$$f(x) = \text{sgn}(w \cdot \Phi(x) + b) \tag{8}$$

Here, $\Phi(x)$ is the mapping of x from input space R^n to feature space. By using kernel function, it avoids to compute in high-dimensional space, because it only needs to care the inner product operation between input vectors.

From the above formula, we can find that the label information of samples is used in computing, which is the key of supervised learning. In supervised learning, we always need to get the mapping relationship between the feature and label by training from labeled sample. However, the labeled samples, which are in a tiny minority in reality, cannot represent the detailed characteristics of a class. If only trained by a few labeled samples, the generalization of SVM would be not strong. And it would also result in lower classification ability. S3VM is a good solution to this problem. S3VM takes full use of the unlabeled samples by combining SVM with semi-supervised learning algorithms.

3 Semi-supervised support vector machine

3.1 Semi-supervised learning

Machine learning is a core research area in artificial intelligence. According to containing labeled data or unlabeled data in training sets, the type of machine learning can be divided into unsupervised learning, supervised learning and semi-supervised learning.

In unsupervised learning, there are only some given input data, and we should find the hidden structure or law of the input data through using a certain learning method. Clustering algorithm is an unsupervised learning method. In supervised learning, those data contain the label which indicates the classification of data. Its core idea is to get a mapping relationship between the feature and label by training from labeled sample, and this mapping relationship should be consistent with the labeled sample. Most of the classification algorithm belongs to supervised learning [12, 13].

Semi-supervised learning is the research focus on machine learning in recent years. It is a learning method combining unsupervised learning and supervised learning. The basic idea is using a large number of unlabeled data to help the supervised learning method improve effect [14]. In semi-supervised learning, there are labeled dataset $L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$ and unlabeled dataset $U = \{x'_1, x'_2, \ldots, x'_u\}$. Now, we expect to get a function $f{:}X \to Y$, which could accurately predict the label y for the given x. Here, $x_i, x'_j \in X$ is a d-dimensional vector, and $y_i \in Y$ is the label of x_i. In addition, l and u are the number of samples that L and U contain.

The reason for proposing semi-supervised learning method is that the unlabeled data can indeed help improve the performance of an algorithm in theory. Miller and Uyar [15] analyzed from data distribution estimations and draw a conclusion that the generalization ability of classifier can be improved with a significant increase in unlabeled data. Zhang and Oles [16] also pointed out that if a parametric model satisfies $P(x, y|\theta) = P(y|x, \theta)P(x|\theta)$, the unlabeled data would help model get a better ability of estimating the model parameters and improve its performance. From another perspective, supervised learning assumes there are enough labeled data for training, so that it could get a learner with strong generalization. However, it is difficult to obtain labeled data in practical applications, and it would also cost a great resource to label the data. In order to take a good advantage of the large amount of unlabeled data, scholars proposed the semi-supervised learning.

Currently, a variety of semi-supervised learning algorithms is based on two common assumptions, that is, cluster assumption and manifold assumption [17]. Cluster assumption means that samples in the same cluster have a greater probability in having the same label. Manifold assumption considers that close points along the manifold flat have a similar nature or similar label. In these two assumptions, cluster assumption reflects the global feature of models. However, the manifold assumption reflects the local features.

There are several classical methods in semi-supervised learning, such as self-training method [18–20], expectation–maximization method [21], multiview method [22–24], graph-based method [25].

3.2 S3VM

Semi-supervised support vector machine was proposed by Bennett and Demiriz [9], which is a semi-supervised learning method based on cluster assumption. The optimal goal of S3VM is to build classifier by using labeled data and unlabeled data. Similar to the idea of SVM, S3VM requires the maximum margin to separate the labeled data and unlabeled data. And the new optimal classification boundary must satisfy that the classification on original unlabeled data has the smallest generalization error.

Before providing the formula of S3VM proposed by Bennett and Demiriz, we can explore how to use the unlabeled data in SVM. Now, we start from the basic theory of SVM and find some way to use these data.

It is known that SVM is based on VC dimension and structural risk minimization. And the regularization is exactly a way to achieve structural risk minimization, which is the empirical risk plus a regularization term or penalty term.

After adding a penalty term in formula (1), the objective function may be rewritten as:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & y_i(w \cdot x_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, 2, \ldots, l \end{aligned} \tag{9}$$

where $\xi_i (i = 1, 2, \ldots, n)$ are the slack variables. It means the acceptable deviation between function margin and the corresponding data x_i. C is a parameter which controls the weight of penalty term in objective function.

Formula (9) can be rewritten in a form of regularized risk as follows:

$$\min \Phi(w) = \min_{w,b} \left\{ \frac{1}{2}||w||^2 + C \sum_{i}^{l} \max(1 - y_i[\mathbf{w} \cdot x_i + b], 0) \right\} \tag{10}$$

Here, $\frac{1}{2}||\mathbf{w}||^2$ can be seen as a regularization term, and $\max(1 - y_i[\mathbf{w} \cdot x_i + b], 0)$ is the loss function of labeled data.

In order to get semi-supervised support vector machine, we must use the unlabeled data. Now assume that the unlabeled data are labeled, and let the label be $\hat{y} = sign(\mathbf{w} \cdot x + b)$. So, the loss function of unlabeled data can be expressed as:

$$\max(1 - \hat{y}[\mathbf{w} \cdot x + b], 0) = max(1 - |\mathbf{w} \cdot x + b|, 0) \tag{11}$$

After using the unlabeled data, adding (9)–(11) we can get the basic form of semi-supervised support vector machine:

$$\min \Phi(w) = \min_{w,b} \left\{ \frac{1}{2}||w||^2 + C_1 \sum_{i}^{l} \max(1 - y_i[\mathbf{w} \cdot x_i + b], 0) + C_2 \sum_{i=l+1}^{l+u} \max(1 - |\mathbf{w} \cdot x_i + b|, 0) \right\} \tag{12}$$

Here, C_1 and C_2 are the weight of two loss function. $i = l + 1, l + 2, \ldots, l + u$ is the unlabeled data. In the above formula, it needs to add constraints to avoid that those unlabeled data would be divided into a same class:

$$\frac{1}{u} \sum_{j=l+1}^{l+u} f(x_j) = \frac{1}{l} \sum_{i+1}^{l} y_i \tag{13}$$

Until now, we get a reasonable semi-supervised support vector machine. Support vector machine is a convex optimization problem, and more detailed information about convex optimization can refer to supervised tensor learning [26]. However, this semi-supervised support vector machine is a non-convex quadratic programming problem and has a complex calculation in solving. With the new data adding, it needs to resolve the programming problem. In order to solve this problem, scholars have proposed a lot of optimization techniques for non-convex optimization problem [27]. Some typical optimization methods include gradient descent [28, 29], concave convex procedure [30], deterministic annealing [31], continuation method [32], semi-definite programming [33], DC programming and DCA [34], branch-and-bound algorithms [35] and genetic algorithm [36].

Bennett and Demiriz proposed semi-supervised support vector machine on the basis of the robust linear programming (RLP) approach to SVM. The corresponding robust linear program for SVM was defined as:

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,s,\xi} & C \sum\limits_{i=1}^{l} \xi_i + \sum\limits_{j=1}^{n} s_j \\ s.t. & y_i|\mathbf{w} \cdot x_i - b| + \xi_i \geq 1 \\ & \xi_i \geq 0, \quad i = 1, 2, \ldots, l \\ & -s_j \leq w_j \leq s_j, \quad j = 1, 2, \ldots, n. \end{array} \tag{14}$$

In fact, the above formula changes the 2-norm $||w||_2$ in (9) to 1-norm $||w||_1$. One of the advantages is to make the dimension decreases, and another benefit is using linear programming instead of quadratic programming to solve. Moreover, if it is a nonlinear case, it only needs to use the kernel function and can get a good extension.

Based on formula (14), they add two restrictions on the unlabeled data. One restriction assumes that the unlabeled data belong to one of the class, and calculate its error rate; another restriction assumes that the unlabeled data belong to another class, and also calculate its error rate. Then, the objective function calculates the minimum of the two possible misclassifications errors. So, the S3VM could be defined as:

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\eta,\xi,z} & C \left[\sum\limits_{i=1}^{l} \xi_i + \sum\limits_{j=l+1}^{l+k} \min(\eta_i, z_i) \right] + ||w|| \\ s.t. & y_i(\mathbf{w} \cdot x_i + b) + \xi_i \geq 1 \ \xi_i \geq 0 \quad i = 1, 2, \ldots, l \\ & \mathbf{w} \cdot x_j + b + \eta_j \geq 1 \ \eta_j \geq 0 \quad j = l + 1, \ l + 2, \ \ldots, \ l + k \\ & -(\mathbf{w} \cdot x_j + b) + z_j \geq 1 \ z_j \geq 0 \end{array} \tag{15}$$

where $C > 0$ is a fixed misclassification penalty.

They use integer programming to solve this problem. By adding a zero or one decision variable d_j for each x_i, they formulate the semi-supervised support vector machine as a mixed integer program:

$$\begin{array}{ll} \min\limits_{\mathbf{w},b,\eta,\xi,z,d} & C \left[\sum\limits_{i=1}^{l} \eta_i + \sum\limits_{j=l+1}^{l+k} (\xi_j + z_j) \right] + ||\mathbf{w}|| \\ s.t. & y_i[\mathbf{w} \cdot x - b] + \eta_i \geq 1, \ \eta_i \geq 0, \quad i = 1, 2, \ldots, l \\ & \mathbf{w} \cdot \mathrm{x}_j - b + \xi_j + M(1 - d_j) \geq 1, \ \xi_j \geq 0, \quad j = l + 1, \ \ldots, \ l + k \\ & -(\mathbf{w} \cdot x_j - b) + z_j + Md_j \geq 1, \ z_j \geq 0, \ d_j = \{0, 1\} \end{array} \tag{16}$$

However, with the increase in unlabeled data, the complexity of computation will rapidly increase. It is difficult to solve for large unlabeled data. To overcome this difficulty, in 2001, Glenn et al. [37] proposed concave semi-supervised support vector machine (VS3VM). Its main idea is to formulate the problem as a concave minimization problem which is solved by a successive linear approximation algorithm. That reduces some complexity in calculation and makes it able to handle large unlabeled datasets.

Although Bennett and Demiriz proposed S3VM initially, in most cases, the S3VM does not refer to the formula (15). We prefer treating the original form [i.e., formula (12)] as the S3VM. Except mentioning Bennett and Demiriz, the S3VM in this paper refers to formula (12). Now, we will present some good improvements in S3VM.

In 2011, Li et al. [38] proposed safe semi-supervised support vector machine (S4VM). S3VM is based on low-density assumption and tries to find a low-density separator which favors the decision boundary going across low-density regions in the feature space. However, S4VM tries to exploit multiple candidate low-density separators and then select the representative separators. In 2014, Hu et al. [39] combined traditional S3VM with multikernel learning and proposed a multikernel S3VM optimization model based on L_p norm constraint ($L_p - MKL - S3VM$). In this multikernel framework, both basic kernels and manifold kernels are used to achieve a combination of two assumptions (i.e., cluster assumption and manifold assumption). By adding L_p norm constraint to kernel combination coefficients θ, they get the non-sparse solution for θ and improve the interpretability and generalization of S3VM.

Semi-supervised support vector machine is non-smooth and non-convex. To tackle this problem, scholars use the quasi-Newton approach [40] in S3VM. In 2011, Reddy et al. [41] applied a modified quasi-Newton method for S3VM. In 2012, Gieseke et al. [42] proposed a sparse quasi-Newton optimization for S3VM. Both of them made an effective improvement. In 2015, by using a general three-moment method, Jiang et al. [43] constructed a quantic spline symmetric hinge approximation function with three-time differentiability at origin. They applied to smooth S3VM model and proved that it has a good convergence.

When should people use semi-supervised SVM? In fact, it can be used on all two-class problems in semi-supervised learning. Now, assuming that there is a dataset D with 1000 sets of data and D is a two-class problem, we will use these data to get a classifier. In dataset D, 100 sets of data have been labeled, and the others have no label. If we use SVM, we can only use 100 sets of data. However, semi-supervised SVM can make full use of 1000 sets of data. Finally, the performance of resulting classifier would be better by choosing semi-supervised SVM. All in all, if there are small amounts of labeled data and large amounts of unlabeled data, semi-supervised SVM is the great choice.

With the increase in unlabeled data, semi-supervised support vector machines also present a rising trend in practical application. Semi-supervised support vector machine used initially in text classification [44] and made a good result. Semi-supervised support vector machines have been also applied in other areas, such as image classification [19, 45–47], iris image recognition [48], face recognition [49], physical education effect evaluation [50], cancer classification [51] and rapid identification between edible oil and swill-cooked dirty oil [52].

On the application side, semi-supervised support vector machines can be widely used in some classification problems. At present, we can see there are a lot of applications which semi-supervised support vector machines have been applied in, and the potential applications are image classification and text classification. On these two issues, semi-supervised support vector machines have shown good results.

3.3 Transductive support vector machines

In 2009, Joachims [44] proposed transductive support vector machine (TSVM). TSVM is an important method in semi-supervised support vector machines. It is based on cluster assumption. It is interesting that TSVM and S3VM are not only proposed in the same year, but also similar in their main idea and optimization problem. The objective of TSVM is to focus on a particular working set, making it possible to achieve the optimal classification in this working set. And this leads to a poor ability in dealing with new data, i.e., poor generalization.

There are labeled data $(x_1, y_1), \ldots, (x_n, y_n)$, $x_i \in R^m$, $y_i \in \{-1, +1\}$ and unlabeled data $\overset{*}{x_1}, \overset{*}{x_2}, \ldots, \overset{*}{x_k}$. In linearly non-separable cases, the optimization problem of TSVM is shown as follows:

$$\begin{array}{ll} \text{Minimize over} & (y_1^*, \ldots, y_n^*, w, b, \xi_1, \ldots, \xi_n, \xi_1^*, \ldots, \xi_k^*): \\ & \frac{1}{2}||\mathrm{w}||^2 + C\sum_{i=0}^{n} \xi_i + C^* \sum_{j=0}^{k} \xi_j^* \\ s.t. & y_i[(w \cdot x_i) + b] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n \\ & y_j[(w \cdot x_j) + b] \geq 1 - \xi_j^*, \quad j = 1, 2, \ldots, k \\ & \xi_i \geq 0, \quad i = 1, 2, \ldots, n \\ & \xi_j^* \geq 0, \quad j = 1, 2, \ldots, k \end{array} \tag{17}$$

Here, ξ_i and ξ_j^* stand for the relaxation factor of labeled data and unlabeled data, respectively. C and C^* are weight, which represent the effect of labeled data and unlabeled data, respectively, in objective function.

The most important thing in TSVM is the transductive approach. It will have a good classification ability in a particular working set. As for unknown working sets, its performance is just ordinary. In addition, TSVM has a high time complexity and needs to preset the number of positive samples. To solve these problems, scholars have done a lot of work to improve the algorithm. In 2003, Chen et al. [53] proposed a progressive transductive support vector machine (PTSVM). In this algorithm, it is no need to preset the number of positive samples. Instead, according to certain principles, PTSVM gives a possible label for the unlabeled data in the training process and retrains the new dataset of labeled data. However, this algorithm requires labeling the unlabeled data frequently. And with the increase in unlabeled data, the time complexity of this algorithm will increase rapidly, which makes the training speed slow.

The primary deficiency of TSVM is that the number of positive samples N_p in unlabeled data must be appointed before training, and the main cause of this problem is the pairwise exchange criterion. For this problem, Wang et al. [54] proposed individually judging and changing criterion and then put forward an algorithm to label the unlabeled data progressively and adjust the N_p dynamically. Unlike TSVM, the algorithm is not sensitive to the unreasonable number of positive samples. When N_p is given randomly, this method will also show a better performance than TSVM.

TSVM and PTSVM both require solving the quadratic programming problems. Zhang et al. [55] replaced the ε insensitive function by using quadratic loss function, which converts quadratic programming problems to solving linear equations and then proposed least square transduction support vector machine.

There are also other good improved algorithms for TSVM. In 2012, Yu et al. [56] proposed a transductive support vector machine algorithm based on spectral clustering. This algorithm uses spectral clustering to cluster unlabeled data and labels them and then makes transductive inference on the mixed dataset composed by both labeled and unlabeled data. It improves the stability greatly. At the same year, Tian et al. [57] proposed a multiple kernel version of TSVM which combines both cluster assumption and manifold assumption. And in 2014, Zhou et al. [58] applied the quasi-linear kernel to TSVM and proposed a transductive support vector machine with adjustable quasi-linear kernel.

3.4 Laplacian support vector machines

In Sect. 3.1, we mentioned that there are two assumptions in semi-supervised learning: cluster assumption and manifold assumption. And S3VM is exactly based on cluster assumption. The manifold assumption reflects mostly a graph-based semi-supervised learning method. Currently, there are a large number of graph-based semi-supervised learning methods. Their idea is to build a graph and regard the labeled data and unlabeled data as the node in the graph. The similarity between these data can be represented by the edge weights in the graph. By using such a graph, the label information of these data can be passed to the other node, and then, we can label the unlabeled data. In fact, the Laplacian support vector machine (LapSVM) [59] which will be presented in this section is exactly a graph-based semi-supervised learning method. This method is applied in image classification [60] and achieves a good result.

It should be noted that the support vector machine and most of supervised learning methods all use the regularization theory. Their optimization goals can be written as:

$$\min_{f\in H_k} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma||f||_K^2 \tag{18}$$

where $\gamma \geq 0$ and V is a loss function. K is an effective kernel function. H_k corresponds to the reproducing kernel Hilbert space. $||\cdot||_K$ is the inner product on Hilbert space.

Basing on this framework, Belkin proposed a manifold regularization framework [59]. Its main idea is to add a regularization term in formula (18), so formula (18) becomes:

$$\min_{f\in H_k} \frac{1}{l}\sum_{i=1}^{l} V(x_i, y_i, f) + \gamma||f||_K^2 + \eta||f||_K^2 \tag{19}$$

The regularization term contains the labeled data and unlabeled data, which reflects the internal structure of sample data distribution. It should be noted that the manifold regularization is a very important semi-supervised learning method. The manifold regularization can fuse supervised learning and unsupervised learning into semi-supervised learning. Its core idea is to mine the geometry of the distribution of data and add it in formula as a regularization term. The data consist of supervised and unsupervised data. Manifold regularization is a hot topic in semi-supervised learning, and many scholars have done a lot work in this field. For some deficiencies of manifold regularization, scholars also tried to study and improve it. In 2012, Bo et al. [61] proposed ensemble manifold regularization framework to automatically and implicitly estimate the hyperparameters of the manifold regularization. In 2013, Yong et al. [62] proposed multiview vector-valued manifold regularization and manifold regularized multitask learning algorithm [63] to achieve multilabel classification, and both of them were successfully applied to multilabel image classification.

The Laplacian support vector machine that was proposed by Belkin is exactly based on this manifold regularization framework. In LapSVM, the regularization term is a Laplace operator, which is given by

$$||f||_L^2 = \frac{1}{(l+u)^2} f^T L f \tag{20}$$

So the Laplacian support vector machine can be defined as follows:

$$\begin{aligned} \min_{\alpha\in R^{l+u}, \xi\in R^l} \quad & \frac{1}{l}\sum_{i=1}^{l}\xi_i + \gamma_A \alpha^T K \alpha + \frac{\gamma_1}{(u+l)^2}\alpha^T KLK\alpha \\ s.t. \quad & y_i\left(\sum_{j=1}^{l+u}\alpha_j K(x_i, x_j) + b\right) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l \\ & \xi_i \geq 0, \quad i = 1, 2, \ldots, l \end{aligned} \tag{21}$$

For solving the formula (21), we need to use the Lagrange function to transfer it into a dual problem, and then, it can be solved easily.

Compared to S3VM, LapSVM is more outstanding in solving and time complexity. Certainly, it also has some deficiencies. LapSVM is based on the manifold assumption and reflects the local structure information of dataset. In fact, this assumption in some classification problems may be too broad. When these samples are near to the classification boundary, they may exactly belong to different class. In recent years, scholars made a great progress in LapSVM. In order to improve the training speed, in 2011, Melacci et al. [64] trained the Laplacian support vector machine in the primal programming problem instead of transferring it into dual problem. This method reduces the time complexity effectively, which reduces the $O(n^3)$ to $O(k^2)$. In 2013, Qi et al. [65] proposed a cost-sensitive Laplacian support vector machine (Cos-LapSVM), which can deal with the sensitive problem in semi-supervised learning. To control the sparsity and feature selection of LapSVM, in 2014, Tan et al. [66] introduced an adjustable p-norm and proposed Laplacian p-norm proximal support vector machine (Lap-PPSVM). At the same year, Yang et al. [67] proposed a spatio-spectral Laplacian support vector machine (SS-LapSVM) and applied it to hyperspectral image classification. In 2015, Qi et al. [68] further improved LapSVM and proposed a fast Laplacian SVM (FlapSVM). This method does not need to deal with extra matrix and can be effectively solved by SOR technology, which makes it more suitable for large-scale problems.

3.5 meanS3VM

We have presented the S3VM, TSVM and LapSVM above. S3VM proposed by Bennett and Demiriz needs to solve the mixed integer program, whose solving procedure has a high time complexity. TSVM will lead to a large number of iterations. And the LapSVM needs to calculate the inverse of a $n \times n$ matrix. All of them have a deficiency in time complexity.

The meanS3VM was proposed by Li et al. [69], and they found that the label mean, being a simple statistic, can be useful in building a semi-supervised learner. The meanS3VM estimates the label means of the whole unlabeled data firstly. In fact, after we get the label means of unlabeled data, we will find that the semi-supervised support vector machine is similar to the SVM which knows the all label of data. And this will make meanS3VM be more efficient. The goal of meanS3VM is to maximize the margin between the label means of the unlabeled data. The expression can be defined as follows:

$$\begin{aligned}
\min_{d\in\Delta}\ \min_{\mathbf{w},b,\rho,\xi}\quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C_1\sum_{i=1}^{l}\xi_i - C_2\rho \\
s.t.\quad & y_i(\mathbf{w}'\phi(x_i)+b) \geq 1-\xi_i, \quad i=1,2,\ \ldots,\ l \\
& \frac{1}{u_+}\left(\mathbf{w}'\sum_{j=l+1}^{l+u} d_{j-l}\phi(x_j)\right) + b \geq \rho \\
& \frac{1}{u_-}\left(\mathbf{w}'\sum_{j=l+1}^{l+u} (1-d_{j-l})\phi(x_j)\right) + b \leq -\rho
\end{aligned} \tag{22}$$

where $\Delta = \{\mathbf{d} | d_i \in \{0,1\}, \sum_{i=1}^{u} d_i = u_+\}$. Balance constraints are contained in Δ. In addition, using the label mean does not mean to add a constraint to each unlabeled data. The main characteristic of the meanS3VM is highly efficient with a short time in training. In particular, on relatively large data, meanS3VM is 100 times faster than TSVM and 10 times faster than LapSVM.

After that, in 2010, Li et al. [70] proposed cost-sensitive semi-supervised support vector machine (CS4VM), which combines the cost-sensitive learning with meanS3VM. CS4VM uses the different cost for different misclassification based on meanS3VM. In fact, CS4VM is the cost-sensitive version of meanS3VM.

3.6 S3VM based on cluster kernels

In semi-supervised learning, kernels have been given a certain attention. Labeled data and unlabeled data are used to construct the kernel functions, which make kernel function better reflect the similarity between samples. In 2002, relying on the cluster assumption, Chapelle [71] changed the spectrum of the kernel matrix and proposed a framework for constructing kernels. They applied it to SVM and got a good result compared to TSVM. By using this framework, we can get a semi-supervised support vector machine based on cluster kernels. The key is to use the new kernel function constructed by labeled data and unlabeled data to train the SVM.

This S3VM based on cluster kernels has an advantage, which it is no need to modify the objective function of SVM. If the objective function was modified, we would face with the difficult in solving and low efficiency. Those semi-supervised support vector machines that we mentioned above have a same characteristic; that is, they all need to modify their objective function. These unlabeled data are added in the optimization problem, and then, it will bring a new objective function and optimization problem. The S3VM based on cluster kernels only needs to construct a new kernel function, which will reduce some complexity. Certainly, this method has a high time complexity, and it may have a poor performance in dealing with large datasets.

Using different method, we can get different kernels, and there are some main kernels: spectral clustering kernel [72], random walk kernel [73], bagged clustering kernel [74–76], mixture models of Gaussian kernel [77].

4 Conclusions

This paper presents these mainstream models of semi-supervised support vector machines, including S3VM, TSVM, LapSVM, meanSVM and S3VM based on cluster kernel. In addition, this paper gives a brief understanding on their corresponding improved method.

Without a doubt, lapSVM and meanSVM are the outstanding and potential algorithm in these five mainstream algorithms. We can analyze these five algorithms from the following aspects. Considering from the solving process, lapSVM and S3VM based on cluster kernel are easier than others. LapSVM is based on manifold regularization framework, which can be solved by using Lagrange function and duality theory. The S3VM based on cluster kernel only needs to construct a special kernel function. Considering from the time complexity, lapSVM and meanSVM all have a good performance. In particular, on relatively large data, meanS3VM is 100 times faster than TSVM and 10 times faster than LapSVM. Considering from the accuracy, these algorithms have their prominence in different datasets, and it is not accurate to say which the best is.

The prosperous application of semi-supervised support vector machine has attracted more and more attention. The current research hot spots can be summarized in two aspects: modify the semi-supervised support vector machine and improve its efficiency; the other is to find a more efficient optimal method to solve the optimization function, enhance the robustness of algorithms and improve its generalization. These two aspects are also the future research directions for semi-supervised support vector machines. These works all contribute to making semi-supervised support vector machines better and more convenient in application.

These semi-supervised SVMs in this article are improved and modified from the basic function of SVM. By doing this, SVM can be changed into semi-supervised SVM and applied in semi-supervised learning. Surely, by using semi-supervised learning models like co-training models, SVM also can be applied in semi-supervised learning. This method does not modify the basic function of SVM, and its focus is on the study of semi-supervised learning models. Applications also exist in this area, and there are good results, such as multitraining support vector machine for image retrieval proposed by Jing et al. [78].

It is no doubt that semi-supervised learning is more suitable for the reality. Meanwhile, the SVM is also an important classification method in machine learning, and combining it with semi-supervised learning seems to be more significant. On the one hand, semi-supervised support vector machines inherit the solid theory of SVM, which will enhance its generalization ability. On the other hand, by introducing semi-supervised learning, semi-supervised support vector machines can use the unlabeled data to improve its performance and extend its application.

However, semi-supervised support vector machines also have deficiency. Those methods will all cost a large time in training, and it is the biggest challenge in semi-supervised support vector machines. Some methods improve this situation to a certain degree, but there are still a lot of works to do in the future.

Acknowledgments This work is supported by the National Natural Science Foundation of China (No. 61379101).

References

1. Altun Y, Belkin M, Mcallester DA (2005) Maximum margin semi-supervised learning for structured variables. In: Proceedings of the 2005 annual conference on neural information processing systems, pp 33–40
2. Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 3(1):1–130
3. Hady MFA, Schwenker F (2013) Semi-supervised learning. Handbook on neural information processing. Springer, Berlin, pp 215–239
4. Ding SF, Qi BJ, Tan YH (2011) An overview on theory and algorithm of support vector machines. J Univ Electron Sci Technol China 40(1):2–10
5. Gu YX, Ding SF (2011) Advances of support vector machines. J Comput Sci Technol 38(2):14–17
6. Schölkopf B, Smola AJ (2001) Learning with kernels: SUPPORT vector machines, regularization, optimization, and beyond. MIT Press, Cambridge
7. Ding S (2011) Incremental learning algorithm for support vector data description. J Softw 6(7):1166–1173
8. Liu XL, Ding SF (2010) Appropriateness in applying SVMs to text classification. Comput Eng Sci 32(6):106–108
9. Bennett K, Demiriz A (1999) Semi-supervised support vector machines. In: Advances in neural information processing systems, vol 11, pp 368–374
10. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
11. Steinwart I, Christmann A (2008) Support vector machines. Springer, New York
12. Jordan MI, Jacobs RI (2014) Supervised learning and divide-and-conquer: a statistical approach. In: Proceedings of the tenth international conference on machine learning, pp 159–166
13. Shipp MA, Ross KN, Tamayo P et al (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8(1):68–74
14. Zhang CG, Zhang Y (2013) Semi-supervised learning. China Agricultural Science and Technology Press, Beijing
15. Miller DJ, Uyar HS (1997) A mixture of experts classifier with learning based on both labelled and unlabelled data. In: Advances in neural information processing systems. MIT Press, Cambridge, pp 571–577

16. Zhang T, Oles F (2000) The value of unlabeled data for classification problems. In: Proceedings of 17th international conference on machine learning, pp 1191–1198
17. Chapelle O, Schölkopf B, Zien A (2006) Semi-supervised learning. MIT Press, Cambridge
18. Triguero I, Sáez JA, Luengo J et al (2014) On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. Neurocomputing 132:30–41
19. Dópido I, Li J, Marpu PR et al (2013) Semi-supervised self-learning for hyperspectral image classification. IEEE Trans Geosci Remote Sens 51(7):4032–4044
20. Li Y, Li H, Guan C, et al (2007) A self-training semi-supervised support vector machine algorithm and its applications in brain computer interface. In: International conference on acoustics, speech, and signal processing, pp 385–387
21. McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley-Interscience, New York
22. Dong C, Yin Y, Guo X et al (2008) On co-training style algorithms. Int Conf Nat Comput 18(20):196–201
23. Feger F, Koprinska I (2006) Co-training using RBF nets and different feature splits. In: International joint conference on neural networks, pp 1878–1885
24. Yu S, Krishnapuram B, Rosales R et al (2011) Bayesian co-training. J Mach Learn Res 12:2649–2680
25. Zhu X (2005) Semi-supervised learning with graphs, Carnegie Mellon University, language technologies institute, school of computer science
26. Tao D, Li X, Wu X et al (2007) Supervised tensor learning. Knowl Inf Syst 13(1):1–42
27. Chapelle O, Sindhwani V, Keerthi SS (2008) Optimization techniques for semi-supervised support vector machines. J Mach Learn Res 9:203–233
28. Chapelle O, Zien A (2005) Semi-supervised classification by low density separation. In: Proceedings of the 10th international workshop on artificial intelligence and statistics, pp 57–64
29. Gieseke F, Airola A, Pahikkala T et al (2014) Fast and simple gradient-based optimization for semi-supervised support vector machines. Neurocomputing 123:23–32
30. Collobert R, Sinz F, Weston J et al (2006) Large scale transductive SVMs. J Mach Learn Res 7:1687–1712
31. Sindhwani V, Keerthi SS, Chapelle O (2006) Deterministic annealing for semi-supervised kernel machines. In: Proceedings of the 23rd international conference on machine learning, pp 841–848
32. Chapelle O, Chi M, Zien A (2006) A continuation method for semi-supervised SVMs. In: Proceedings of the 23rd international conference on machine learning, pp 185–192
33. De Bie T, Cristianini N (2006) Semi-supervised learning using semi-definite programming. MIT press, Cambridge
34. Le HM, Le Thi HA, Nguyen MC (2015) Sparse semi-supervised support vector machines by DC programming and DCA. Neurocomputing 153:62–76
35. Chapelle O, Sindhwani V, Keerthi S (2007) Branch and bound for semi-supervised support vector machines. In: 20th Annual conference on neural information processing systems, pp 217–224
36. Adankon MM, Cheriet M (2010) Genetic algorithm–based training for semi-supervised SVM. Neural Comput Appl 19(8):1197–1206
37. Fung G, Mangasarian OL (2002) Semi-supervised support vector machines for unlabeled data classification. Optim Methods Softw 15:29–44
38. Li Y, Zhou Z (2015) Towards making unlabeled data never hurt. IEEE Trans Pattern Anal Mach Intell 37(1):175–188
39. Hu QH, Ding LX, He JR (2013) L_p norm constraint multi-kernel learning method for semi-supervised support vector machine. J Softw 24(11):2522–2534
40. Yu J, Vishwanathan SVN, Günter S et al (2010) A quasi-Newton approach to nonsmooth convex optimization problems in machine learning. J Mach Learn Res 11(5):1145–1200
41. Reddy IS, Shevade S, Murty MN (2011) A fast quasi-Newton method for semi-supervised SVM. Pattern Recogn 44(10): 2305–2313
42. Gieseke F, Airola A, Pahikkala T, et al (2012) Sparse quasi-newton optimization for semi-supervised support vector machines. In: Proceedings of the 1st international conference on pattern recognition applications and methods, pp 45–54
43. Jiang W, Yao L, Jiang X et al (2015) A new classification method based on semi-supervised support vector machine. In: Human-centred computing. First international conference, pp 633–645
44. Joachims T (1999) Transductive inference for text classification using support vector machines. In: Proceedings of the sixteenth international conference, vol 99, pp 200–209
45. Guillaumin M, Verbeek J, Schmid C (2010) Multimodal semi-supervised learning for image classification. In: 2010 IEEE Computer society conference on computer vision and pattern recognition, pp 902–909
46. Li M, Wang R, Tang K (2013) Combining Semi-Supervised and active learning for hyperspectral image classification. In: Proceedings of the 2013 IEEE symposium on computational intelligence and data mining, pp 89–94
47. Xie L, Pan P, Lu Y (2014) Markov random field based fusion for supervised and semi-supervised multi-modal image classification. Multimed Tools Appl 74(2):613–634
48. Yang L, Su Q, Yang B et al (2014) A new semi-supervised support vector machine classifier based on wavelet transform and its application in the iris image recognition. Int J Appl Math Stat 52(5):86–93
49. Lu K, He X, Zhao J (2006) Semi-supervised support vector learning for face recognition. Advances in neural networks-ISNN 2006. Springer, Berlin, pp 104–109
50. Liang P, Xueming Y (2012) Explore semi-supervised support vector machine algorithm for the application of physical education effect evaluation. Int J Adv Comput Technol 4(9):266–271
51. Jun CA, Habibollah H, Haza NAH (2015) Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. Springer International Publishing, Switzerland, pp 468–477
52. Zhou Y, Liu T, Li J (2015) Rapid identification between edible oil and swill-cooked dirty oil by using a semi-supervised support vector machine based on graph and near-infrared spectroscopy. Chemometr Intell Lab Syst 143:1–6
53. Chen YS, Wang GP, Dong SH (2003) A progressive transductive inference algorithm based on support vector machine. J Softw 14(3):451–460
54. Wang Y, Huang S (2005) Training TSVM with the proper number of positive samples. Pattern Recogn Lett 26(14): 2187–2194
55. Zhang R, Wang W, Ma Y et al (2009) Least square transduction support vector machine. Neural Process Lett 29(2):133–142
56. Yu X, Yang J, Zhang J (2012) A transductive support vector machine algorithm based on spectral clustering. AASRI Procedia 1:384–388
57. Tian X, Gasso G, Canu S (2012) A multiple kernel framework for inductive semi-supervised SVM learning. Neurocomputing 90:46–58
58. Zhou B, Hu C, Chen B, et al. (2014) A Transductive Support Vector Machine with adjustable quasi-linear kernel for semi-supervised data classification. In: Proceedings of the 2014 international joint conference on neural networks, pp 1409–1415
59. Belkin M, Niyogi P, Sindhwani V (2006) Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. J Mach Learn Res 7:2399–2434

60. Gómez-Chova L, Camps-Valls G, Munoz-Mari J et al (2008) Semisupervised image classification with Laplacian support vector machines. IEEE Geosci Remote Sens Lett 5(3):336–340
61. Geng B, Tao D, Xu C et al (2012) Ensemble manifold regularization. IEEE Trans Pattern Anal Mach Intell 34(6):1227–1233
62. Luo Y, Tao D, Xu C et al (2013) Multiview vector-valued manifold regularization for multilabel image classification. IEEE Trans Neural Netw Learn Syst 24(5):709–722
63. Luo Y, Tao D, Geng B et al (2013) Manifold regularized multitask learning for semi-supervised multilabel image classification. IEEE Trans Image Process 22(2):523–536
64. Melacci S, Belkin M (2011) Laplacian support vector machines trained in the primal. J Mach Learn Res 12:1149–1184
65. Qi Z, Tian Y, Shi Y et al (2013) Cost-sensitive support vector machine for semi-supervised learning. Procedia Comput Sci 18:1684–1689
66. Tan J, Zhen L, Deng N et al (2014) Laplacian p-norm proximal support vector machine for semi-supervised classification. Neurocomputing 144:151–158
67. Yang L, Yang S, Jin P et al (2014) Semi-supervised hyperspectral image classification using spatio-spectral Laplacian support vector machine. IEEE Geosci Remote Sens Lett 11(3):651–655
68. Qi Z, Tian Y, Shi Y (2015) Successive overrelaxation for Laplacian support vector machine. IEEE Trans Neural Netw Learn Syst 26(4):674–683
69. Li Y, Kwok JT, Zhou Z (2009) Semi-supervised learning using label mean. In: Proceedings of the 26th international conference on machine learning, pp 633–640
70. Li Y, Kwok J, Zhou Z (2010) Cost-sensitive semi-supervised support vector machine. In: Proceedings of the 24th AAAI conference on artificial intelligences, pp 500–505
71. Chapelle O, Weston J, Scholkopf B (2002) Cluster kernels for semi-supervised learning. In: Proceedings of 16th annual neural information processing systems conference, pp 585–592
72. Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 2:849–856
73. Szummer M, Jaakkola T (2002) Partially labeled classification with Markov random walks. Adv Neural Inf Process Syst 14:945–952
74. Tuia D, Camps-Vall G (2009) Semi-supervised remote sensing image classification with cluster kernels. IEEE Geosci Remote Sens Lett 6(1):224–228
75. Li T, Wang XL (2013) Semi-supervised SVM classification method based on cluster kernel. Appl Res Comput 30(1):42–45
76. Gao HZ, Wang JW, Xu K et al (2011) Semisupervised classification of hyperspectral image based on clustering kernel and LS-SVM. J Signal Process 27(2):76–81
77. Tao XM, Cao PD, Song SY et al (2013) The SVM classification algorithm based on semi-supervised gauss mixture model kernel. Inf Control 42(1):18–26
78. Li J, Allinson N, Tao D et al (2006) Multitraining support vector machine for image retrieval. IEEE Trans Image Process 15(11):3597–3601