Optimal Margin Distribution Additive Machine

CHANGYING GUO, HAO DENG, AND HONG CHEN
College of Science, Huazhong Agricultural University, Wuhan 430070, China
Corresponding author: Hong Chen (chenh@mail.hzau.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities of China under Grant 2662020LXQD001.

ABSTRACT In recent years, sparse additive machines have attracted increasing attention in high dimensional classification due to their flexibility and representation interpretability. However, most of the existing methods are formulated under Tikhonov regularization schemes associated with the hinge loss, where the distribution information of observations is neglected usually. To circumvent this problem, we propose an optimal margin distribution additive machine (called ODAM) by incorporating the optimal margin distribution strategy into sparse additive models. The proposed approach can be implemented by a dual coordinate descent algorithm and its empirical effectiveness is confirmed on simulated and benchmark datasets.

INDEX TERMS Sparse additive machine, margin distribution, dual coordinate descent, classification.

I. INTRODUCTION Sparse additive machines (SAMs) have shown promising performance for classification [5], [8], [33], which are rooted in generalized additive models [14], [25]. Following the popular sparse additive model (SpAM) [20] for regression, the sparse additive machine is proposed in [33] for binary classification. Moreover, the group sparse additive machine (GroupSAM) is formulated in [5] based on kernel-based hypothesis spaces and grouped variables, where it can be implemented via the proximal gradient descent algorithm [19], [33]. Notice that the above additive classification models are formulated under the Tikhonov regularization schemes associated with the hinge loss, which is related closely with the large margin strategy for binary classification [10], [23], [28].

In machine learning literature, there are some classification models from the perspective of margin distribution. [22] suggests margin theory may explain the phenomenon that AdaBoost may resistant to overfitting, and [3] points out the importance of the minimum margin and proposes the corresponding algorithm by maximizing the minimum margin. Later, [21] conjectures that the margin distribution has a more important influence on the generalization performance, which has been verified empirically in [11]. As shown in [11], [35], both margin mean and variance are crucial to characterize the margin distribution. With respect to the support vector machine (SVM), the optimal margin distribution machine (ODM) is proposed in [26], [36] and has shown the competitive performance. Since the previous works mainly focus on linear classification models or kernel-based classifiers, it is natural and important to further incorporate the marginal distribution strategy into additive models.

Inspired by recent works in [26] and [33], we propose a new classification algorithm, called optimal margin distribution additive machine (ODAM), which integrating the optimal marginal distribution strategy, data dependent hypothesis spaces, and sparse additive models together. Firstly, the sparse additive machine is used to construct the frame of model. From the represent theorem of kernel methods, the model can be transformed into a concise expression. And this guarantees the interpretability and flexibility of the model. Then ODAM learns the strategy of [26] and explores the optimal model from the perspective of margin distribution associated with SAM, which assures a strong generalization of the method. Finally, the proposed model can be optimized by a dual coordinate descent method. To support the motivation of algorithmic design, we evaluate the proposed ODAM via data experiment on simulated and benchmark data. Our main contributions can be summarized as follows.

• A new sparse additive machine, called ODAM, is proposed in this paper. The proposed ODAM seeks the decision rule according to marginal distribution, rather than from the perspective of the loss function. To the best of our knowledge, the optimal margin distribution strategy has not been investigated in additive models.
The proposed ODAM shows the empirical effectiveness of classification on both simulated and real-world data sets. Besides, it enjoys the stability and interpretability of classification results.

We organized the rest of this paper as follows. After recalling the preliminaries of binary classification and sparse additive machine in Section II, we formulate the optimal margin distribution additive machine and its optimization algorithm in Section III and IV, respectively. Then an empirical evaluation is presented in Section V. Finally, we conclude this paper in Section VI.

II. PRELIMINARIES

In the section, we recall the preliminaries of additive models for binary classification.

A. BINARY CLASSIFICATION PROBLEM

We denote by $\mathcal{X} \in \mathbb{R}^p$ a compact input space and $\mathcal{Y} = \{-1, 1\}$ the corresponding output set. Let $\rho$ be an unknown distribution over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ to generate the input-output pair $(x, y)$. Define $\text{sgn}(f(x)) = 1$ if $f(x) \geq 0$ and $\text{sgn}(f(x)) = -1$ otherwise. Given observations $z = \{(x_i, y_i)\}_{i=1}^n$ drawn independently from unknown $\rho$ on $\mathcal{Z}$, the purpose is to search a classifier $\text{sgn}(f(x))$ induced by a decision function $f : \mathcal{X} \rightarrow \mathbb{R}$ to minimize the misclassification risk $f \in \mathcal{Z}$. Let $I(\mathcal{A})$ denote the corresponding output set. Let $I(\mathcal{A})$ be an unknown $\rho(x, y)$, where $I(A) = 1$ if $A$ is true and 0 otherwise.

The most popular surrogate loss of $I(\text{sgn}(f(x)) \neq y)$ is the hinge loss

$$\ell(\text{yf}(x)) = (1 - yf(x))_+ = \max[0, 1 - yf(x)]$$

due to its excellent property in classification margin and support vectors [10], [23], [28]. Associated with the hinge loss, the expected risk and the empirical risk can be denoted as

$$E(f) = \int (1 - yf(x))_+ d\rho(x, y)$$

and

$$E_d(f) = \frac{1}{n} \sum_{i=1}^{n} (1 - yf(x_i))_+.$$ 

For support vector machines [29], the margin of instance $(x_i, y_i)$ is defined as

$$y_i = yf(x_i), \quad \forall i = 1, \ldots, n.$$  (1)

B. SPARSE ADDITIVE MACHINE

As nonlinear extensions of linear models (e.g., Lasso [27], $\ell_1$-norm SVM [2], [34]), sparse additive models have shown much flexibility and adaptivity on prediction and variable selection. In theory, the additive requirement of hypothesis space in additive models is crucial to cure “the curse of dimensionality” of nonparametric learning [13], [17], [24]. For the binary classification problem, additive models under Tikhonov regularization scheme also have shown promising performance., see, e.g., [5], [8], [33].

Now, we recall the kernel additive models in [5], [8]. Let $\mathcal{X} = (X_1, \ldots, X_p)^T$ with $X_j \subset \mathbb{R}$, $1 \leq j \leq p$. Denote $K_j : X_j \times X_j \rightarrow \mathbb{R}$ as a Mercer kernel and denote $\mathcal{H}_{K_j}$ as the corresponding reproducing kernel Hilbert space (RKHS) [1] with norm $\| \cdot \|_{K_j}$. The additive hypothesis space

$$\mathcal{H} = \left\{ \sum_{j=1}^{p} f_j : f_j \in \mathcal{H}_{K_j}, \quad 1 \leq j \leq p \right\}$$

with

$$\| f \|_K = \inf \left\{ \sum_{j=1}^{p} \| f_j \|_{K_j}^2 : f = \sum_{j=1}^{p} f_j \right\}$$

is also an RKHS with additive kernel $K = \sum_{j=1}^{p} K_j$ [9], [18], [32].

The support vector machine with additive kernel $K = \sum_{j=1}^{p} K_j$ in [8] is defined as below:

$$\tilde{f}_z = \arg \min_{f \in \mathcal{H}} \{ E_z(f) + \eta \| f \|_K \}$$

$$= \arg \min_{f \in \mathcal{H}_{K_{j_1}, \ldots, j_p}} \{ E_z(\sum_{j=1}^{p} f_j) + \eta \sum_{j=1}^{p} \| f_j \|_{K_j} \}$$  (2)

where $\eta$ is a trade-off parameter. According to the representation properties of (2), $\tilde{f}_z$ always belongs to the following data dependent hypothesis space

$$\mathcal{H}_z = \left\{ f = \sum_{j=1}^{p} f_j : f_j = \sum_{i=1}^{n} \alpha_{ji} K_j(x_{ji}, \cdot), \alpha_{ji} \in \mathbb{R} \right\},$$  (3)

where $x_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p, i = 1, \ldots, n$. It is natural to consider the coefficient-based penalty (see e.g., [6], [7], [12]) defined as

$$\Omega(f) = \inf \left\{ \sum_{j=1}^{p} \| \alpha_j \|_1 : f = \sum_{j=1}^{n} \alpha_{ji} K_j(x_{ji}, \cdot), \right.$$  

$$\left. \quad \alpha_j = (\alpha_{j1}, \ldots, \alpha_{jn})^T \in \mathbb{R}^p \right\}.$$ 

Then, the sparse additive model in [5] is formulated as follows:

$$f_z = \arg \min_{f \in \mathcal{H}_z} \{ E_d(f) + \lambda \Omega(f) \},$$

where $\lambda > 0$ is a regularization parameter.

III. OPTIMAL MARGIN DISTRIBUTION ADDITIVE MACHINE

This section formulates the optimal margin distribution sparse additive machine (ODAM). We know the first-order and second-order statistics are the most straightforward metrics for characterizing the margin distribution, e.g., the mean and the variance of the margin. Denote $y := [y_1, \ldots, y_n]^T$, and $Y$ is a $n \times n$ diagonal matrix with $y_1, \ldots, y_n$ as the
diagonal elements. According to the definition in (1) and (3),
the margin of sparse additive machine is

\[ t_k = y_k f(x_k) = y_k \left( \sum_{j=1}^{p} \sum_{i=1}^{n} \alpha_{ij} K(x_{ij}, x_{ij}) \right) \]

\[ = y_k \left( \sum_{j=1}^{p} \alpha_j K(x_{ij}) \right) \]

\[ = y_k \alpha^T K(x_k) \]

where \( \alpha = (\alpha_1, \ldots, \alpha_p)^T \in \mathbb{R}^{np \times 1}, K(x_{ij}) = (K_1(x_{ij}, x_{ik}), \ldots, K_p(x_{ij}, x_{ik}))^T \in \mathbb{R}^{n \times 1}, j = 1, \ldots, p, K(x) = (K_1(x), \ldots, K_p(x_{ik}))^T \in \mathbb{R}^{np \times 1}, k = 1, \ldots, n. \)

For simplicity, denote \( X = (K(x_i))_{i=1}^{n} \in \mathbb{R}^{np \times n} \). Then, the margin mean is

\[ \bar{\iota} = \frac{1}{n} \sum_{i=1}^{n} y_i \alpha K(x_i) = \frac{1}{n} (X y)^T \alpha \] (4)

and the margin variance is

\[ \tilde{\iota} = \frac{1}{n} \sum_{i=1}^{n} (y_i \alpha^T K(x_i) - \bar{\iota})^2 \]

\[ = \frac{1}{n} \alpha^T \left( \frac{1}{n} \sum_{i=1}^{n} K(x_i) K(x_i)^T \right) \alpha - \frac{2}{n} \sum_{i=1}^{n} y_i \alpha^T K(x_i) \alpha + \tilde{\iota}^2 \]

\[ = \frac{1}{n} \alpha^T X X^T \alpha - \frac{1}{n^2} \alpha^T X y y^T X^T \alpha \]

\[ = \alpha^T X n I_n - y y^T X^T \alpha \] (5)

where \( I_n \) is the \( n \)-order identity matrix. Following the large margin strategy of SVM in [4], we can get that the maximization of the minimum distance in additive model as below

\[ \max_{\alpha} \bar{\iota}, \]

\[ \text{s.t. } y_i \alpha^T K(x_i) \geq \bar{\iota}, \quad i = 1, \ldots, n \]

where \( \bar{\iota} = \frac{\iota}{\|\alpha\|} \) denotes the margin of additive model. Since \( \iota \) does not have influence on the optimization, we can set \( \iota = 1 \) simply. Besides, maximizing \( 1/\|\alpha\| \) is equivalent to minimizing \( \|\alpha\|^2 / 2 \), we get the following optimization:

\[ \min_{\alpha} \frac{1}{2} \alpha^T \alpha, \]

\[ \text{s.t. } y_i \alpha^T K(x_i) \geq 1, \quad i = 1, \ldots, n. \] (6)

If the training examples cannot be separated with the zero error, (6) can be reformulated as below:

\[ \min_{\alpha, \xi} \frac{1}{2} \alpha^T \alpha + \frac{C}{n} \sum_{i=1}^{n} \xi_i, \]

\[ \text{s.t. } y_i \alpha^T K(x_i) \geq 1 - \xi_i, \]

\[ \xi_i \geq 0, \quad i = 1, \ldots, n, \] (7)

where \( \xi_i \) is slack factor and \( C \) is the slack parameter.

Firstly, we consider the separable case where the training examples can be separated with zero error. The minimum of the margin variance and the maximization of the margin mean induce to the following hard-margin ODAM,

\[ \min_{\alpha} \frac{1}{2} \alpha^T \alpha + \lambda_1 \bar{\iota} - \lambda_2 \tilde{\iota} \]

\[ \text{s.t. } y_i \alpha^T K(x_i) \geq 1, \quad i = 1, \ldots, n. \] (8)

where \( \lambda_1 \) and \( \lambda_2 \) are the trade-off parameters to balance the margin mean, the margin variance and the model complexity. When \( \lambda_1 = \lambda_2 = 0 \), (8) is reduced to the origin optimization (6).

In the non-separable cases, similar to soft-margin SVM, the soft-margin ODAM leads to

\[ \min_{\alpha, \xi} \frac{1}{2} \alpha^T \alpha + \lambda_1 \bar{\iota} - \lambda_2 \tilde{\iota} + \frac{C}{n} \sum_{i=1}^{n} \xi_i, \]

\[ \text{s.t. } y_i \alpha^T K(x_i) \geq 1 - \xi_i, \]

\[ \xi_i \geq 0, \quad i = 1, \ldots, n. \] (9)

Because (8) is a special case of (9), we will focus on soft-margin ODAM. If without clarification, ODAM is referred to the soft-margin ODAM.

IV. COMPUTING ALGORITHM

In this section, we present the optimization algorithm for ODAM. The dual of ODAM is convex quadratic optimization with simple constraints, and can be computed via a dual coordinate descent method [26], [30]. By substituting (4) and (5) into (9), we get the following quadratic programming problem:

\[ \min_{\alpha, \xi} \frac{1}{2} \alpha^T \alpha + \alpha^T X \left( \frac{\lambda_1(n I_n - y y^T)}{n^2} X^T \alpha \right) \]

\[ \text{s.t. } Y X^T \alpha \geq e - \xi, \]

\[ \xi \geq 0, \] (10)

where \( e \in \mathbb{R}^n \) stands for the all-one vector. We introduce the lagrange multipliers \( \eta > 0 \) and \( \beta > 0 \) for first and second constraints respectively to obtain the Lagrangian function of (10)

\[ L(\alpha, \xi, \eta, \beta) = \frac{1}{2} \alpha^T \alpha + \alpha^T X \left( \frac{\lambda_1(n I_n - y y^T)}{n^2} X^T \alpha \right) \]

\[ - \frac{\lambda_2}{n} \eta^T Y X^T \alpha - \frac{C}{n} \xi^T \xi \]

\[ - \eta^T (Y X^T \alpha - e + \xi) - \beta^T \xi \]

\[ = \frac{1}{2} \alpha^T \omega \alpha - \alpha^T Y \left( \frac{\lambda_2}{n} e + \eta \right) \]

\[ + \eta^T e + \xi^T \left( \frac{C}{n} e - \eta - \beta \right) \] (11)

where \( \omega = I_{n \times p} + X^T \left( \frac{2 \lambda_2(n I_n - y y^T)}{n^2} X \right). \) In order to solve this problem, we set the partial derivations of \{\alpha, \xi\} to 0, then get

\[ \frac{\partial L}{\partial \alpha} = \omega \alpha - Y \left( \frac{\lambda_2}{n} e + \eta \right) \Rightarrow \alpha = \omega^{-1} Y \left( \frac{\lambda_2}{n} e + \eta \right), \] (12)

\[ \frac{\partial L}{\partial \xi} = \frac{C}{n} e - \xi - \beta \Rightarrow 0 \leq \xi \leq \frac{C}{n} e. \] (13)
Substituting (12) and (13) into (11), we get the dual of (10)
\[
\min_{\eta} f(\eta) = \frac{1}{2} \left( \frac{\lambda_2}{n} e + \eta \right)^T X Y \left( \frac{\lambda_2}{n} e + \eta \right) - e^T \eta,
\]
s.t. \( 0 \leq \eta \leq \frac{C}{n} e. \) \hspace{1cm} (14)

According to Lemma 1 in [36], \( X^T \omega^{-1} X = G(I_n + AG)^{-1} \), where \( A = \frac{2\lambda_2 (\sigma^2 - \eta_1^2)}{n^2} \) and \( G = X^T X \). We denote \( H = YG(I_n + AG)^{-1}Y \), then the objective function of (14) can be cast as
\[
f(\eta) = \frac{1}{2} \left( \frac{\lambda_2}{n} e + \eta \right)^T H \left( \frac{\lambda_2}{n} e + \eta \right) - e^T \eta
= \frac{1}{2} \eta^T H \eta + \frac{\lambda_2}{n} e^T H \eta - e^T \eta + \text{const}
= \frac{1}{2} \eta^T H \eta + \left( \frac{\lambda_2}{n} He - e \right)^T \eta + \text{const}.
\]

Because the value of const is not influential on the optimization, we neglect the const term. Then, the final formulation of ODAM is described as below
\[
\min_{\eta} \frac{1}{2} \eta^T H \eta + \left( \frac{\lambda_2}{n} He - e \right)^T \eta,
\]
s.t. \( 0 \leq \eta \leq \frac{C}{n} e. \)

According to (12) and \( \omega^{-1} X = X(I_n + AG)^{-1} \), we can obtain the coefficients \( \alpha \) from the optimal \( \eta^* \) as
\[
\alpha = X(I + AG)^{-1}Y \left( \frac{\lambda_2}{n} e + \eta^* \right).
\]

For each input \( x \), we predict its output as
\[
\text{sgn}(f)(x) = \text{sgn} \left( \sum_{j=1}^{p} \sum_{i=1}^{n} \alpha_i K(x_i, x) \right) \hspace{1cm} (15)
\]

As shown in [31], the convex quadratic objective function with simple decoupled box constraints can be efficiently solved by the dual coordinate descent method. Observe that ODAM can be considered as a special case of convex quadratic objective function (that is, \( v = \infty \)),
\[
\min_{\eta} f(\eta) = \frac{1}{2} \eta^T H \eta + q^T \eta
\]
s.t. \( 0 \leq \eta \leq v \)

According to [15], one of the variables is selected to minimize the target while the other variables are kept as constants at each iteration. The following close-form can be obtained at each iteration:
\[
\min_{i} f(\eta + te_i) \hspace{1cm} \text{s.t.} \ 0 \leq \eta_i + t \leq v_i
\]

where \( e_i \) denotes the vector with 1 in the \( i-th \) coordinate and 0 elsewhere. Let \( H = [h_{ij} \mid i,j=1,\ldots,n] \), we have
\[
f(\eta + te_i) = \frac{1}{2} h_{ii} t^2 + [\nabla f(\eta)]_i t + f(\eta) \hspace{1cm} (16)
\]

where \([\nabla f(\eta)]_i\) is \( i-th \) component of the gradient \( \nabla f(\eta) \). From (16), we know that \( f(\eta + te_i) \) is a simple quadratic function of \( t \), so we get a close-form solution
\[
\eta_i^{\text{new}} = \min(\max(\eta_i - \frac{[\nabla f(\eta)]_i}{h_{ii}}, 0), v_i). \]

We summary the optimization algorithm of ODAM as below.

\textbf{Algorithm 1 Dual Coordinate Descent Method for ODAM}
\begin{itemize}
  \item \textbf{Input:} Data \([x_i, y_i]\)\(_{i=1}^{n}\).
  \item \textbf{Initialization:} \( \eta = 0 \).
  \item \textbf{Calculation:} kernel \( K \) (Gaussian kernel), matrix \( H \) and vector \( q \).
  \item \textbf{While} \( \eta \) not converged \textbf{do}
    \begin{itemize}
      \item for \( i = 1, \ldots, n \)
      \begin{itemize}
        \item \( [\nabla f(\eta)]_i \leftarrow [H \eta + q]_i \)
        \item \( \eta_i \leftarrow \min(\max(\eta_i - \frac{[\nabla f(\eta)]_i}{h_{ii}}, 0), v_i) \)
      \end{itemize}
    \end{itemize}
  \item \textbf{End While}
  \item \textbf{Calculation:} \( \alpha = X(I + AG)^{-1}Y \left( \frac{\lambda_2}{n} e + \eta^* \right) \)
  \item \textbf{Output:} \( \text{sgn}(f)(x) \) from (15).
\end{itemize}

\section{Experimental Analysis}
In this section, we assess the empirical performance of our approach on simulated and real-world datasets, where \( \ell_1 \)-SVM [34], SAM [33], and ODM [36] are employed as baselines. In Section V-A, we introduce the experimental setting. And we conduct the simulated and the real-world data experiments in Section V-B, V-C respectively, and give related analysis from classification accuracy and interpretation.

\subsection{Experimental Setup}
In all experiments, the RKHS \( \mathcal{H}_K \) (with Gaussian kernel \( K_\sigma (u, v) = \exp(-\frac{|u-v|^2}{2\sigma^2}) \)) is employed as the hypothesis function space. The learning performance is measured by the average classification accuracy with standard deviation. All parameter selections are performed on the training set and the specific selection method is as follows. For ODM and ODAM, the regularization parameter \( C \) is selected by 5-fold cross validation from \([0.1, 0.5, 1, 5, 10, 50]\), and \( \lambda_1 \) and \( \lambda_2 \) are selected from the set of \([10^{-6}, 10^{-5}, \ldots, 10^{-1}] \). In addition, the bandwidth of kernel \( \sigma \) is obtained from \([1 + 10^{-1}i, i = 0, 1, \ldots, 100]\).

\subsection{Simulated Data}
To illustrated the performance of our method, we consider two synthetic data following the experimental design in [5], [33].

\textbf{Example 1:} We generate each \( p \) dimension input \( x_i = (x_{i1}, \ldots, x_{ip})^T \) by \( x_i \leftarrow \frac{W_0 + V_i}{W_0 + V_i} \), where both \( W_0 \) and \( V_i \) are extracted from the uniform distribution \( U(0, 1) \). The function \( f(x_i) = x_{i1}^2 + 3 \cdot \sin(\pi x_{i3} + x_{i4}) - \log(x_{i5} \cdot x_{i6} + 2) \) is
selected as the discriminant function in first example and the label is assigned by \( y_t = \text{sgn}(f(x_t)) \).

**Example 2:** The second example follows the way of Example 1 to generate data. The different is that the discriminant function is \( f(x_t) = x_{t1} + 2 \cdot x_{t2} + 2 \cdot x_{t3} + \sin(\pi x_{t4}) - \log(x_{t5} + 3) - |x_{t6}| \).

For each evaluation, we consider training set with different size \( n = 100, 200, 400 \) and dimension \( p = 100, 200, 400 \). And the training and test sets are generated with identical sample size. To ensure reliable results, each evaluation is repeated 50 times. The results of experiments are displayed in Table 1. To illustrate the stability of the proposed method, we applied the method in [16] to conduct another experiment on Example 2 with sample size \( n = 200 \), dimension \( p = 100 \) and different flipping rate \( r = 0.0, 0.025, 0.05, \ldots, 0.3 \). The experimental results are shown in Figures 1 and 2. Finally, we give the coefficient-based norm of ODAM associated with each input variable \( ||\alpha_j|| = \sum_{i=1}^{n} |\alpha_{ji}| \), and the corresponding results are shown in Figure 3.

### TABLE 1. The averaged performance on simulated data in Example 1 and Example 2.

| n  | p   | SVM | OAM | ODM | ODM  |
|----|-----|-----|-----|-----|------|
| 100| 100 | 0.811 ± 0.029 | 0.838 ± 0.034 | 0.818 ± 0.047 | 0.859 ± 0.051 |
| 200| 200 | 0.841 ± 0.018 | 0.820 ± 0.026 | 0.805 ± 0.034 | 0.866 ± 0.025 |
| 400| 400 | 0.853 ± 0.020 | 0.836 ± 0.021 | 0.816 ± 0.028 | 0.892 ± 0.021 |

**Example 2**

| n  | p   | SVM | OAM | ODM | ODM  |
|----|-----|-----|-----|-----|------|
| 100| 100 | 0.781 ± 0.040 | 0.793 ± 0.032 | 0.731 ± 0.048 | 0.841 ± 0.038 |
| 200| 200 | 0.836 ± 0.021 | 0.832 ± 0.020 | 0.795 ± 0.035 | 0.852 ± 0.031 |
| 400| 400 | 0.847 ± 0.016 | 0.827 ± 0.022 | 0.811 ± 0.021 | 0.871 ± 0.016 |

| n  | p   | SVM | OAM | ODM | ODM  |
|----|-----|-----|-----|-----|------|
| 100| 100 | 0.724 ± 0.039 | 0.729 ± 0.040 | 0.714 ± 0.045 | 0.837 ± 0.038 |
| 200| 200 | 0.769 ± 0.041 | 0.818 ± 0.047 | 0.761 ± 0.037 | 0.844 ± 0.028 |
| 400| 400 | 0.849 ± 0.017 | 0.827 ± 0.016 | 0.776 ± 0.024 | 0.855 ± 0.015 |

**TABLE 2. Information about the data sets.**

| Dataset | # Instance | # Feature |
|---------|------------|-----------|
| phishing | 1355 | 9 |
| spam | 1151 | 18 |
| balance scale | 765 | 4 |
| nlp | 110 | 18 |
| credit | 691 | 30 |
| bankruptcy | 1372 | 4 |
| blood classification | 748 | 4 |
| wine | 178 | 13 |

As shown in Table 3, the proposed ODAM has better performance on most benchmark datasets, and has only extremely small gaps on other datasets. In addition, the small standard deviations in most datasets illustrate the stability of rates besides promising classification accuracy. Figure 3 illustrates that the coefficient-based norms associated with true variables are larger than the other variables, which implies that we can screen them with a suitable threshold.

### C. REAL-WORLD DATA

We select 12 benchmark datasets from UCI (http://archive.ics.uci.edu/ml) to illustrate the classification performance of ODAM. The partial information of these data sets is counted in Table 2. For each dataset, we select 50% as the test set, and let the rest as the training set. All features are normalized to \([0, 1]\) and each experiment is repeated 25 times. The main metric is classification accuracy with standard deviation and the average results are shown in Table 3.
the new method. All these outcomes verify the effectiveness of our method.

VI. CONCLUSION
This paper proposes a new optimal margin distribution additive machine (ODAM) in RKHS by integrating the additive machine and the optimal margin distribution strategy. With the help of an optimization strategy based on the dual coordinate descent algorithm, we verified the effectiveness of ODAM on simulated and benchmark datasets.

ACKNOWLEDGMENT
(Changing Guo and Hao Deng contributed equally to this work.)

REFERENCES
[1] N. Aronszajn, “Theory of reproducing kernels,” Trans. Amer. Math. Soc., vol. 68, no. 3, pp. 337–404, 1950.
[2] P. S. Bradley and O. L. Mangasarian, “Feature selection via concave minimization and support vector machines,” in Proc. Int. Conf. Mach. Learn. (ICML), San Francisco, CA, USA, 1998, pp. 82–90.
[3] L. Breiman, “Prediction games and arcing algorithms,” Neural Comput., vol. 11, no. 7, pp. 1493–1517, Oct. 1999.
[4] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[5] H. Chen, X. Wang, C. Deng, and H. Huang, “Group sparse additive machine,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2017, pp. 198–208.
[6] H. Chen and Y. Wang, “Kernel-based sparse regression with the correntropy-induced loss,” Appl. Comput. Harmon. Anal., vol. 44, no. 1, pp. 144–164, Jan. 2018.
[7] H. Chen, Z. Pan, L. Li, and Y. Tang, “Error analysis of coefficient-based regularized algorithm for density-level detection,” Neural Comput., vol. 25, no. 4, pp. 1107–1121, Apr. 2013.
[8] A. Christmann and R. Hable, “Consistency of support vector machines using additive kernels for additive models,” Comput. Statist. Data Anal., vol. 56, no. 4, pp. 854–873, Apr. 2012.
[9] A. Christmann and D.-X. Zhou, “Learning rates for the kernel-based quantile regression estimators in additive models,” Anal. Appl., vol. 14, no. 03, pp. 449–477, May 2016.
[10] F. Cucker and D. X. Zhou, Learning Theory: An Approximation Theory Viewpoint. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[11] W. Gao, Z. H. Zhou, and “On the doubt about margin explanation of boosting,” Artif. Intell., vol. 203, pp. 1–18, Feb. 2013.
[12] T. Gong, Z. Xu, and H. Chen, “Generalization analysis of fredholm kernel regularized classifiers,” Neural Comput., vol. 29, no. 7, pp. 1879–1901, Jul. 2017.
[13] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, A Distribution-Free Theory Nonparametric Regression. New York, NY, USA: Springer-Verlag, 2002.
[14] J. T. Hastie and R. J. Tibshirani, Generalized Additive Models. London, U.K.: Chapman-Hall, 1990.
[15] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, “A dual coordinate descent method for large-scale linear SVM,” in Proc. 25th Int. Conf. Mach. Learn., 2008, pp. 408–415.
[16] X. Huang, L. Shi, and J. A. K. Suykens, “Ramp loss linear programming support vector machine,” J. Mach. Learn. Res., vol. 15, pp. 2185–2211, Jan. 2014.
[17] K. Kandasamy and Y. Yu, “Additive approximations in high dimensional nonparametric regression via the SALSA,” in Proc. Int. Conf. Mach. Learn. (ICML), New York, USA, 2016, pp. 69–78.
[18] S. Lv, H. Lin, H. Lian, and J. Huang, “Oracle inequalities for sparse additive quantile regression in reproducing kernel Hilbert space,” Ann. Statist., vol. 46, no. 2, pp. 781–813, Apr. 2018.
[19] Y. Nesterov, “Smooth minimization of non-smooth functions,” Math. Program., vol. 103, pp. 127–152, May 2005.
[20] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman, “SPAM: Sparse additive models,” J. Roy. Stat. Soc., Ser. B, vol. 71, pp. 1009–1030, Oct. 2009.
[21] L. Reyzin and R. E. Schapire, “How boosting the margin can also boost classifier complexity,” in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 753–760.
[22] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, “Boosting the margin: A new explanation for the effectiveness of voting methods,” Ann. Statist., vol. 26, no. 5, pp. 1651–1686, Oct. 1998.
[23] I. Steinwart and A. Christmann, Support Vector Machine. Berlin, Germany: Springer, 2008.
[24] C. J. Stone, “Optimal global rates of convergence for nonparametric regression,” Ann. Statist., vol. 10, no. 4, pp. 1040–1053, 1983.
[25] C. J. Stone, “Additive regression and other nonparametric models,” Ann. Statist., vol. 13, no. 2, pp. 689–705, Jun. 1985.
[26] T. Zhang and Z. H. Zhou, “Optimal margin distribution machine,” in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, New York, NY, USA, 2014, pp. 313–322.
[27] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. Roy. Stat. Soc., Ser. B Methodol., vol. 58, no. 1, pp. 267–288, Jan. 1996.
[28] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer-Verlag, 1995.
[29] Y. Xu and W. Yin, “A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion,” SIAM J. Imag. Sci., vol. 6, no. 3, pp. 1758–1789, Jan. 2013.
[30] G.-X. Yuan, C.-H. Ho, and C.-J. Lin, “Recent advances of large-scale linear classification,” Proc. IEEE, vol. 100, no. 9, pp. 2584–2603, Sep. 2012.
[31] M. Yuan and D.-X. Zhou, “Minimax optimal rates of estimation in high dimensional additive models,” Ann. Statist., vol. 44, no. 6, pp. 2564–2593, Dec. 2016.
[32] T. Zhao and H. Liu, “Sparse additive machine,” in Proc. 15th Int. Conf. Artif. Intell. Statist., vol. 22, 2012, pp. 1435–1443.
[33] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, “1-norm support vector machines,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), British Columbia, 2003, pp. 49–56.
[34] Z. H. Zhou, “Large margin distribution learning,” in Proc. 6th IAPR Int. Workshop Artif. Neural Netw. Pattern Recognit., Montreal, QC, Canada, 2014, pp. 1–11.
[35] T. Zhang and Z.-H. Zhou, “Optimal margin distribution machine,” IEEE Trans. Knowl. Data Eng., vol. 32, no. 6, pp. 1143–1156, Jun. 2020.
HAO DENG received the M.Sc. degree from Hubei University, Wuhan, China, in 2016, and the Ph.D. degree from the Macau University of Science and Technology, Macau, China, in 2019. He is currently a Lecturer with the College of Science, Huazhong Agricultural University, Wuhan. His current research interests include statistical learning theory and machine learning.

HONG CHEN received the B.Sc. and Ph.D. degrees from Hubei University, Wuhan, China, in 2003 and 2009, respectively. He is currently a Professor with the Department of Mathematics and Statistics, College of Science, Huazhong Agricultural University, Wuhan. His current research interests include statistical learning theory, approximation theory, and machine learning.