2015

The Support Vector Machine and Mixed Integer Linear Programming: Ramp Loss SVM with L1-Norm Regularization

Eric J. Hess
Virginia Commonwealth University, hessej@vcu.edu

J. Paul Brooks
Virginia Commonwealth University, jpbrooks@vcu.edu

http://dx.doi.org/10.1287/ics.2015.0017

14th INFORMS Computing Society Conference
Richmond, Virginia, January 11–13, 2015
pp. 226–235

The Support Vector Machine and Mixed Integer Linear Programming: Ramp Loss SVM with L_1-Norm Regularization

Eric J. Hess and J. Paul Brooks
Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, VA 23284, hessej@vcu.edu, jpbrooks@vcu.edu

Abstract The support vector machine (SVM) is a flexible classification method that accommodates a kernel trick to learn nonlinear decision rules. The traditional formulation as an optimization problem is a quadratic program. In efforts to reduce computational complexity, some have proposed using an L_1-norm regularization to create a linear program (LP). In other efforts aimed at increasing the robustness to outliers, investigators have proposed using the ramp loss which results in what may be expressed as a quadratic integer programming problem (QIP). In this paper, we consider combining these ideas for ramp loss SVM with L_1-norm regularization. The result is four formulations for SVM that each may be expressed as a mixed integer linear program (MILP). We observe that ramp loss SVM with L_1-norm regularization provides robustness to outliers with the linear kernel. We investigate the time required to find good solutions to the various formulations using a branch and bound solver.

Keywords support vector machine, ramp loss, L1-norm regularization, mixed integer programming

1. Introduction

Classification, the task of assigning objects to one of several predefined categories, is a pervasive problem that encompasses many diverse applications [20]. The support vector machine (SVM) has become increasingly popular due to the ease of implementation and ability to apply of nonlinear discriminant functions. SVM is a binary classification method that seeks to find a hyperplane that balances the sometimes-competing objectives of minimizing error while maximizing the distance between correctly classified observations [4]. Traditional SVM with the linear kernel is often used but performance is diminished with the presence of outliers.

Carrizosa and Romero Morales [7] provide an extensive review of the relationship between mathematical optimization and classification, including many approaches for SVM. Below we review some results relating to ramp loss SVM and L_1-norm regularization.

Ramp loss SVM [1, 18, 19, 22, 6] presents a solution to the outlier problem with formulations of SVM that give less weight to problematic training points and therefore better generalizability as compared to traditional SVM. Although performance is greatly improved, these formulations are computationally complex and therefore many classification tasks may be computationally intractable with this approach. To date, computational studies of the robustness of ramp loss SVM are limited. Error measurement for SVM with ramp loss differs from traditional SVM in that all misclassified observations falling outside of the margin add a loss of 2 to the objective function, while the error for observations falling in the margin is between 0 and 2, and depends on the distance to the margin boundary [6]. Shen et al.

[19], Wang et al. [22], and Collobert et al. [9] suggest algorithms for ramp loss SVM that converge to locally optimal solutions. Wu and Liu [23] use difference of convex functions (DC) programming to solve a nonconvex formulation of SVM with the ramp loss and linear kernel. Brooks [6] presents an MIQP formulation that accommodates the kernel trick, describes some facets for ramp loss SVM with the linear kernel, and introduces heuristics for deriving feasible solutions from fractional ones at nodes in the branch and bound tree. Carrizosa et al. [8] describe additional heuristic approaches for solving MIQP formulations of ramp loss SVM. Wang and Slobodan [22] describe a fast online algorithm OnlineSVM$^{\mathrm{R}}$ for estimating ramp loss SVM. Online learning involves learning as data is presented to the algorithm, rather than having a single static training set.

Related efforts to increase robustness to outlier observations include discrete SVM (DSVM) [15, 17, 16] that implements SVM with the hard margin loss and the linear kernel. The hard margin loss assigns an error of one to observations between the margin boundaries and those outside the margin boundaries and misclassified, and zero otherwise. DSVM is formulated as a mixed integer linear program (MILP). Ustun et al. [21] formulate an optimization-based linear classification method that penalizes the number of misclassifications and both the L_0 and L_1 norm of the coefficients of the discriminant via MILP.

Traditional SVM uses regularization to avoid overfitting noise by introducing a quadratic term in the objective function of the corresponding optimization formulation. Finding an optimal SVM hyperplane for a given training dataset requires the solution of a quadratic program. By linearizing the regularization term, the training problem becomes a linear program (LP). Robust LP formulations for finding optimal hyperplanes for discrimination of linearly inseparable sets were studied in the early 1990s prior to papers on SVM being published in English. Bennett and Mangasarian [3] proposed a single LP formulation which generates a plane that minimizes the average errors of misclassified points belonging to two disjoint sets of observations in n-dimensional real space. Mangasarian [14] introduced a generalized SVM formulation that could be used for both quadratic and LP formulations. Hadzic [10] and Kecman [12] provide a formulation that utilizes the L_1 norm of the separating hyperplane for maximizing the margin that performed well in several empirical tests. Zhou et al. [24] introduce an LP SVM formulation in which the margin is defined as the right hand side of the constraint, $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$, with an unknown in place of 1. This is unique because the right hand side of these constraints is representative of the margin width of the separating hyperplanes.

This paper proposes several formulations for SVM with ramp loss trained via MILP. The new formulations discussed herein minimize the L_1-norm of some measure of the primal variables, $\mathbf{w}$, or dual variables, α by modifying the SVM formulations with linear regularization due to Mangasarian[13], Hadzic and Kecman [10], and Zhou et al. [24]. Branch and bound can be used to solve these problems with subproblems that are LPs. The benefits of using the L_1-norm regularization with the ramp loss include a reduction in computational complexity from the MIQP models for ramp loss and an increase in robustness to outliers from traditional SVM (with L_1-norm or L_2-norm regularization).

The remainder of the paper is organized as follows. In the next section, we describe the previously-proposed QIP formulations for ramp loss SVM. In Section 3, we propose four new formulations for ramp loss SVM and L_1-norm regularization. In Section 4, we present computational results that explore whether the new formulations are less computationally intensive than previous ramp loss SVM formulations and provide better robustness to outliers than traditional SVM.

2. Ramp Loss SVM as a Quadratic Integer Program

We assume training data are given consisting of observations $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ each having an associated class label $y_i \in \{-1, 1\}$. It is typically assumed that there is a probability

distribution $P(\mathbf{x}, y)$ from which these data are drawn, and we try to estimate the true misclassification rates of rules by observing empirical error. A hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ for discrimination purposes is the result of applying linear SVM.

To produce a ramp loss function the errors are weighted as follows.

$$R\left(\mathbf{x}_i f\left(y_i, f(\mathbf{x}_i)\right)\right) = \begin{cases} 0, & y_i f\left(\mathbf{x}_i\right) > 1, \\ 1 - y_i f\left(\mathbf{x}_i\right), & -1 \le y_i\left(\mathbf{x}_i\right) \le 1, \\ 2, & y_i f\left(\mathbf{x}_i\right) \le 1. \end{cases} \tag{1}$$

Brooks [6] presents MIQP formulations for SVM with ramp loss error terms. The formulation that we will call SVMIP1-RL can be used to find a hyperplane that maximizes the margin while minimizing the ramp loss.

$$\begin{aligned} \text{(SVMIP1-RL)} \quad & \min_{\mathbf{w},b,\xi,\mathbf{z}} \left\{ \frac{1}{2}\left|\left|\mathbf{w}\right|\right|^2 + C\sum_{i=1}^{n}(\xi_i + 2z_i) \right\}, \\ \text{subject to} \quad & y_i\left(\mathbf{w}\cdot\mathbf{x}_i + b\right) = 1 - \xi_i \text{ if } z_i = 0, \quad i = 1, 2, ..., n, \\ & 0 \le \xi_i \le 2, \quad i = 1, 2, ..., n, \\ & z_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{2}$$

Just like traditional SVM, this ramp loss model can generate nonlinear discriminants by subsituting primal variables with dual variables; i.e., $\mathbf{w} = \sum_{i=1}^{n} y_i \mathbf{x}_i \alpha_i$, and replacing $\mathbf{x}_i$ with $\Phi\left(\mathbf{x}_i\right)$. Replacement of these variables, as well as application of the kernel trick, i.e. $\mathbf{x}_i \cdot \mathbf{x}_j \mapsto \Phi\left(\mathbf{x}_i\right) \cdot \Phi\left(\mathbf{x}_j\right) = K\left(\mathbf{x}_i, \mathbf{x}_j\right)$ gives us the following.

$$\begin{aligned} \text{(SVMIP2-RL)} \quad & \min_{\alpha,b,\xi,\mathbf{z}} \left\{ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j K\left(\mathbf{x}_i, \mathbf{x}_j\right)\alpha_i\alpha_j + C\sum_{i=1}^{n}(\xi_i + 2z_i) \right\}, \\ \text{subject to} \quad & y_i\left(\sum_{j=1}^{n}\alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_i) + b\right) = 1 - \xi_i \text{ if } z_i = 0, \quad i = 1, 2, ..., n, \\ & 0 \le \alpha_i \le C, \quad i = 1, 2, ..., n, \\ & 0 \le \xi_i \le 2, \quad i = 1, 2, ..., n, \\ & z_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n. \end{aligned} \tag{3}$$

The conditional constraints can be used in a branch and bound solver directly via "indicator constraints" [11]. Alternatively, one can linearize the conditional constraints by replacing the right hand side with $1 - \xi_i - Mz_i$ for all training vectors. The choice of the parameter M is important because making it too small may bias the resulting discriminant towards outliers. On the other hand making M too large may lead to numerical instability. An option for controlling the size of M may be to solve the linear problem for the magnitudes of the ξs, and adjusting M as necessary.

3. New Formulations for Ramp Loss SVM with L_1-Norm Regularization

In this section, we combine ideas for L_1-norm regularization with ramp loss SVM to produce four new formulations that may be solved as MILPs. For each formulation for L_1-norm regularization, incorporating the ramp loss amounts to either using conditional "indicator" constraints [11] or adding a binary integer variable to each soft margin constraint.

Mangasarian [13] introduced a generalized SVM formulation that considers functions for regularization that lead to quadratic programming and LP formulations for SVM and also allows for weighting the misclassification of observations. Two different options for an LP

SVM are given. The first is the L_1 norm of the dual variables, α. This formulation can incorporate the ramp loss as follows.

$$\begin{aligned}
\text{(GSVM1-RL)} \quad & \min_{\alpha,\mathbf{s},b,\xi,\mu} \left\{ \sum_{i=1}^{n} s_i + C \sum_{i=1}^{n} (\xi_i + 2z_i) \right\}, \\
& \text{subject to} \quad y_i \left(\sum_{j=1}^{n} y_j K(\mathbf{x}_i, \mathbf{x}_j) \alpha_j + b \right) \geq 1 - \xi_i + M z_i, \quad i = 1, 2, \ldots, n, \\
& \qquad\qquad s_i \geq \alpha_i \geq -s_i, \quad i = 1, 2, \ldots, n, \\
& \qquad\qquad \xi_i \geq 0, \quad i = 1, 2, \ldots, n, \\
& \qquad\qquad z_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n.
\end{aligned} \tag{4}$$

The variable s_i is the absolute value of α_i.

The second LP SVM uses absolute value of $\sum_{j=1}^{n} (y_j K(\mathbf{x}_i, \mathbf{x}_j) \alpha_j)$. The extension to incorporate the ramp loss is given in the following formulation.

$$\begin{aligned}
\text{(GSVM2-RL)} \quad & \min_{\alpha,\mathbf{s},b,\xi,\mu} \left\{ \sum_{i=1}^{n} s_i + C \sum_{i=1}^{n} (\xi_i + 2z_i) \right\}, \\
& \text{subject to} \quad y_i \left(\sum_{j=1}^{j=n} y_j K(\mathbf{x}_i, \mathbf{x}_j) \alpha_j + b \right) \geq 1 - \xi_i + M z_i, \quad i = 1, 2, \ldots, n, \\
& \qquad\qquad s_i \geq \sum_{j=1}^{n} y_j K(x_i, x_j) \alpha_j \geq -s_i, \quad i = 1, 2, \ldots, n, \\
& \qquad\qquad \xi_i \geq 0, \quad i = 1, 2, \ldots, n, \\
& \qquad\qquad z_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n.
\end{aligned} \tag{5}$$

In examining (5), the first set of constraints can be seen as generalizing the soft margin constraints on each training observation in traditional SVM. The second set of constraints ensures that each s_i is the absolute value of the each regularization term. If the kernel is linear then it is interesting to note that the regularization term we are minimizing is $\sum_{i=1}^{n} |(\mathbf{w} \cdot \mathbf{x}_i)|$. In this way, the function that we are minimizing is perhaps more similar to traditional SVM than (4).

Hadzic and Kecman [10] and Kecman [12] provide a formulation that utilizes the L_1 norm of the separating hyperplane for maximizing the margin. The first formulation is derived from classical SVM with the only exception being that the L_1 norm is used as a regularization term instead of the L_2 norm of the discriminant coefficients. Kecman and Hadzic's LP classifier can be modified to accommodate the ramp loss as follows.

$$\begin{aligned}
\text{(MILPSVM-RL)} \quad & \min_{\mathbf{w}^+,\mathbf{w}^-,b,\xi} \left\{ \sum_{i=1}^{n} \left(w_i^+ + w_i^- \right) + \sum_{i=1}^{n} (\xi_i + 2z_i) \right\}, \\
& \text{subject to} \quad y_i (\mathbf{x}_i \cdot \mathbf{w}^+ - \mathbf{x}_i \cdot \mathbf{w}^- + b) \geq 1 - \xi_i - M z_i, \quad i = 1, 2, \ldots, n, \\
& \qquad\qquad \mathbf{w}^+, \mathbf{w}^- \geq 0, \\
& \qquad\qquad \mathbf{z} \in \{0, 1\}.
\end{aligned} \tag{6}$$

In this SVM formulation, the attribute vector $\mathbf{w}$ is split into positive $\mathbf{w}^+$ and negative $\mathbf{w}^-$ parts so that the generalization terms in the objective function are linear instead of quadratic. This new measure is the L_1 norm because the sum of $\mathbf{w}^+$ and $\mathbf{w}^-$ equates to the sum of the absolute value of all parts of the vector $\mathbf{w}$. This formulation is the closest literal transition from the traditional L_2-norm regularization of SVM to L_1-norm with ramp loss error terms. However, it is not clear how the formulation can be used with the kernel trick, and so only linear discriminants can be learned.

The fourth and last formulation for ramp loss SVM with L_1 regularization is an extension of an LP formulation due to Zhou et al.[24]. They introduce an LP SVM formulation by considering the L_∞ norm for regularization. Their formulation replaces the metric representing the magnitude of the classifier margin in the objective function with a variable r in the equation for the margin, $d = 2r/||\mathbf{w}||_\infty$. By replacing the right hand side of the bounding constraints in traditional SVM with r (in place of 1), we allow for the substitution of terms in the objective function.

$$\begin{aligned}
\text{(rSVM-RL)} \quad & \min_{\alpha,b,r,\xi} \left\{ -r + C \sum_{i=1}^{n} (\xi_i + 2z_i) \right\}, \\
\text{subject to} \quad & y_i \left(\sum_{j=1}^{n} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i) + b \right) \geq r - \xi_i - M z_i, \quad i = 1, 2, \ldots, n, \\
& -1 \leq \alpha_i \leq 1, \quad i = 1, 2, \ldots, n, \\
& r \geq 0, \\
& 0 \leq \xi_i \leq 2, \quad i = 1, 2, \ldots, n, \\
& z_i \in \{0, 1\}, \quad i = 1, 2, \ldots, n.
\end{aligned} \tag{7}$$

Bounds are placed on α so that the absolute value of each variable is at most 1. Even with these constraints, however, the original LP formulation is unbounded when $C < \max\{1/n^+, 1/n^-\}$, where n^+ and n^- represent the number of observations in each class. Also, setting all variables to zero always provides a feasible solution. Therefore, non-trivial solutions occur only when a feasible solution exists where the objective function value is negative. Therefore, C must be large enough to avoid an unbounded problem but small enough to avoid a rule that places all observations in one class. With the incorporation of the ramp loss, (7) is no longer unbounded. However, setting all variables to zero remains feasible with a zero objective function value.

4. Computational Results

We hypothesized that incorporating the ramp loss in each of the methods for LP SVM would provide an increased robustness to outliers when compared to traditional SVM and would not result in formulations requiring as much computation as the QIP formulations SVMIP1-RL and SVMIP2-RL. We used simulated data sampled from normal distributions with the covariance matrix set to the identity matrix, which is taken from [6]. The origin is the mean for group 1 and $(2/\sqrt{d}, 2/\sqrt{d}, \ldots, 2/\sqrt{d})$ is the mean for group 2, such that the Mahalanobis distance between the two groups is 2, a classic "twonorm" benchmark model due to Breiman [5]. The *Bayes rule*, the classification rule that minimizes the probability of misclassification, assigns group labels based on which mean is the nearest. Therefore, the minimum misclassification error/Bayes error is $P(z > 1) \approx 15.9\%$, where z follows a standard normal distribution. Outliers are intentionally added to create a problem set that could alter the performance of the decision linear boundary found using a traditional SVM model. Outlier observations created by sampling a Gaussian distribution with a mean of $(10/\sqrt{d}, 10/\sqrt{d}, \ldots, 10/\sqrt{d})$ and covariance matrix that is 0.001 times the identity matrix and, so that 10 is the Mahalanobis distance between the outlier and non-outlier distributions. Ten percent of the observations in each training set are outliers. Outliers are not present in the test data set. The test sets are comprised of combinations of sizes in terms of observations and variables. For each data set type we have made training and validation data set for every combination of $n = 60$, 100, 200, 500 observations and $d = 2$, 5, 10 variables.

The corresponding LP SVM formulations due to Mangasarian [13] are hereafter called GSVM1 and GSVM2. The LP SVM formulation due to Hadzic and Kecman [10] is called LPSVM. The LP SVM formulation due to Zhou et al. [24] is called rSVM. The corresponding ramp loss versions are named as labelled in this paper.

All instances were created and solved using Gurobi 5.0 C API on a machine with an Intel Core 2 Duo CPU E8400 3.00 GHz with 4 GB RAM. Nested cross validation was used to train models. Five-fold cross validation was employed to tune the tradeoff parameter C. The values tested were 0.01, 0.1, 1, 10, 100, 1000. Training instances were each allowed 20 seconds to find the best solution within the branch and bound tree with no cuts allowed. In the validation phase, the best parameter settings were used with a time limit of 600 seconds (10 minutes) on the entire training dataset. The test set used for accuracy consists of 50,000 samples from each group distribution and no outliers. The validation instances are available at `http://scholarscompass.vcu.edu/ssor_data/1/`.

The user parameter M was set to 10 for each instance to create an environment in which most or all possible solutions are feasible. A larger value for M is usually recommended but under the Gurobi parameter settings utilized, a large M can produce small non-integer solutions for some of these discrete variables z_i. Any time ξ_i is nonzero then z_i should be zero and visa versa. These small fractional solutions for z_i break this rule therefore making the model invalid. It is recommended to test for these types of constraint errors during the test phase of classification.

Table 1 contains the gap remaining after 600 seconds for the validation models for the SVM formulations with binary integer variables. Let z be the incumbent objective value and let z_{LB} be the best lower bound after 600 seconds. The gap remaining is defined as $(z - z_{LB})/z$. The gap remaining is not precisely comparable across formulations, but we take it as a proxy for computational demands for the formulations. For five instances of rSVM-RL, the solution setting all variables to zero is provably optimal. For the other seven instances, the solver found provably optimal solutions with strictly negative objective function values.

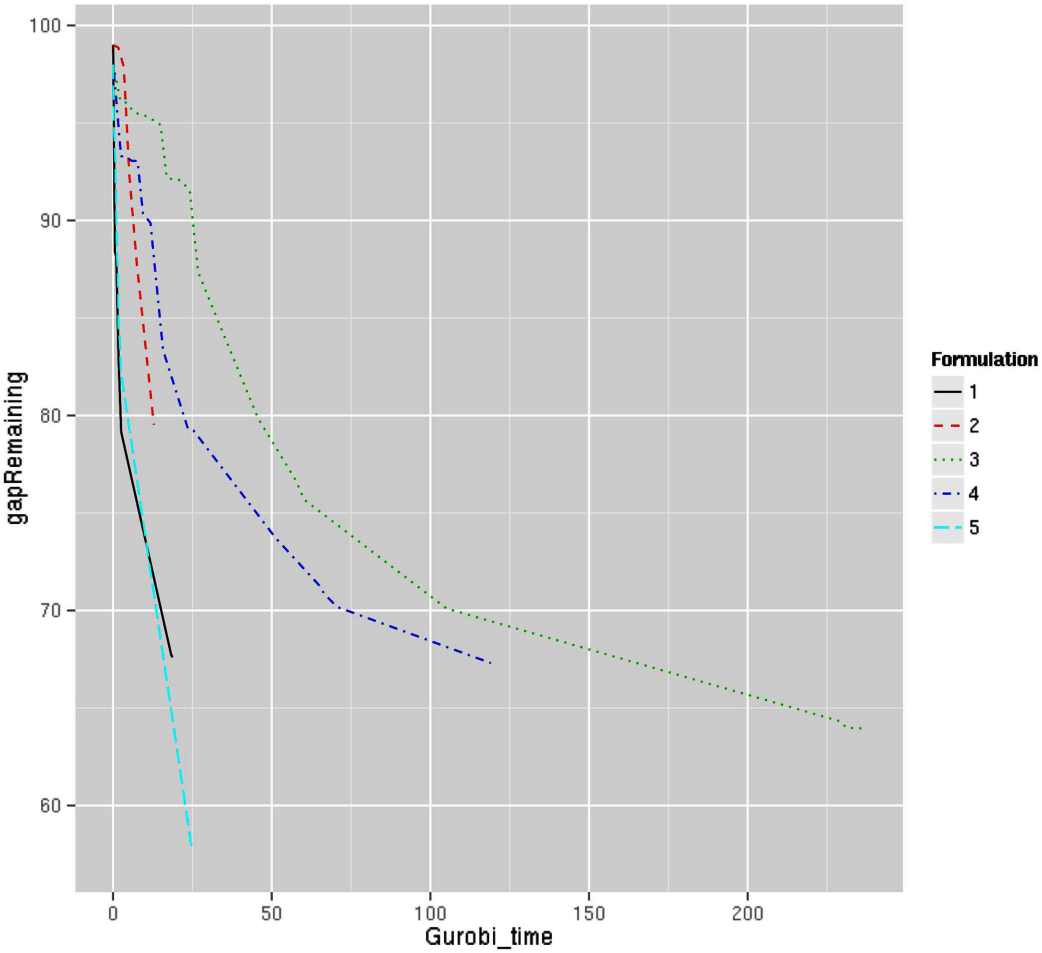
Table 1. Gap remaining (%) after 600 seconds for validation models for SVM formulations with binary integer variables for different numbers of observations (n) and attributes (d)

n	d	SVMIP1	SVMIP2	GSVM1-RL	GSVM2-RL	MILPSVM-RL	rSVM-RL
60	2	21.8	0.0	0.0	12.5	0.0	0.0
100	2	69.8	60.7	77.7	51.4	64.5	0.0
200	2	73.9	79.8	86.9	81.5	77.7	0.0
500	2	90.5	92.8	95.3	93.7	89.8	0.0
60	5	21.0	26.8	44.2	38.2	0.0	0.0
100	5	71.1	64.2	78.0	61.1	69.6	0.0
200	5	76.2	80.4	82.9	82.2	79.9	0.0
500	5	88.2	96.7	95.3	95.6	90.1	0.0
60	10	0.0	0.0	61.3	23.9	39.9	0.0
100	10	51.7	56.5	58.0	60.5	38.8	0.0
200	10	85.9	84.2	91.0	85.5	82.9	0.0
500	10	92.8	96.9	97.6	95.0	91.4	0.0

Of the other 60 instances derived from 12 datasets and the other five formulations, only six were solved to provable optimality. For half of the datasets, MILPSVM-RL has the smallest gap remaining and for another four instances SVMIP1-RL has the smallest gap remaining. Comparing the results for SVMIP2-RL and the MILP-based ramp loss formulations, there does not appear to be a dramatic difference in computational requirements.

To further investigate the time required to generate feasible solutions, we plot the gap remaining as a function of time for the instance with $n = 100$ and $d = 10$ in Figure 1. The gap remaining changes each time a new incumbent solution is found. With the exception of GSVM1-RL and GSVM2-RL, the best known solution is found within 25 seconds. For rSVM-RL, an optimal solution is found within 0.04 seconds and so is not included in the plot.

FIGURE 1. Gap remaining as a function of time for ramp loss SVM formulations for $n = 100$ observations and $d = 10$ attributes.



If good feasible solutions are easy to find for a formulation, but good lower bounds are not, then the gap remaining is not a good measure of the quality of solutions found within the time limit. Table 2 contains the accuracy of the validation models produced by the various formulations on test data. Accuracy can reflect the quality of solutions at the time limit.

As expected, traditional SVM performs poorly in the presence of outliers. The accuracy is as bad as a random guess on 10 of 12 datasets, and is never better than 65%. SVMIP1-RL generates rules that have accuracy of at least 82% on all of the datasets. SVMIP2-RL has accuracy of at least 82% on all but one dataset, where the accuracy is 77.7%.

Incorporating the ramp loss into three of the LP SVM formulations provides an improvement in accuracy. Accuracy increased for eight of 12 datasets for GSVM1, 10 of 12 datasets for GSVM2, and 11 of 12 datsets for LPSVM. For rSVM, accuracy increases for only four of 12 datasets. Unbounded instances and solutions where setting variables to zero are optimal plague both rSVM and rSVM-RL. For rSVM, accuracy is below 65% on seven of 12 datasets, and for rSVM-RL, accuracy is below 65% on 11 of 12 datasets. Because the accuracy increased for these ramp loss formulations, we infer that issues with setting the M parameter in the formulations did not cause numerical instability.

Overall, SVMIP1-RL and GSVM2-RL appear to have the highest accuracy on the outlier-contaminated datasets. The rules generated by these methods are near the Bayes optimal accuracy of 84.1%.

TABLE 2. Accuracy (%) of SVM formulations for different numbers of observations (n) and attributes (d)

n	d	SVM	SVMIP1	SVMIP2	GSVM1	GSVM1-RL	GSVM2	GSVM2-RL	LPSVM	MILPSVM-RL	rSVM	rSVM-RL
60	2	50.0	82.1	77.7	80.6	80.0	84.2	82.7	80.6	80.0	36.1	53.0
100	2	50.0	83.3	82.5	73.8	50.0	82.4	84.0	76.1	80.7	82.7	59.5
200	2	50.0	83.8	83.3	64.2	83.0	77.9	83.9	76.9	84.0	72.6	69.9
500	2	50.0	84.2	84.1	51.0	83.5	80.8	84.1	73.9	83.9	44.5	60.8
60	5	48.8	82.6	82.6	47.5	80.1	47.1	84.3	47.5	80.3	32.6	35.5
100	5	48.3	83.6	83.1	46.1	50.0	50.0	82.6	61.5	79.0	77.5	64.8
200	5	50.0	83.2	83.2	74.5	83.2	84.3	83.4	74.9	82.9	48.9	53.4
500	5	50.0	83.6	84.0	77.4	84.2	82.6	84.0	79.4	83.7	72.9	48.1
60	10	64.4	82.0	82.0	50.0	50.0	83.5	84.1	66.3	78.2	64.2	51.5
100	10	57.3	82.9	82.9	50.0	81.1	84.0	84.1	50.0	78.9	82.7	56.0
200	10	47.0	82.5	82.3	46.0	50.0	64.4	82.5	42.4	81.1	51.1	46.6
500	10	50.0	83.6	82.5	68.7	50.0	76.4	83.0	71.0	83.2	57.1	54.0

Examples of the decision rules generated by the various methods is given in Figure 2. The figure is for $n = 500$ and $d = 2$. For this dataset, traditional SVM places all observations in one group and therefore cannot be plotted. In Figure 2(a)-(c), the ramp loss versions of the SVM formulations provide better decision rules than the QIP SVM formulation SVMIP1-RL. In Figure 2(a) and (c), we can see that GSVM1 and LPSVM provide some robustness to outliers and unlike traditional SVM, produce a decision rule. When the ramp loss is added, GSVM1-RL and MILPSVM-RL provide better robustness and a rule that is similar to SVMIP1-RL. In Figure 2(b), GSVM2 produces a decision rule but is dramatically affected by the outlier observations. The addition of the ramp loss with GSVM2-RL provides robustness to outliers. In Figure 2(d) both rSVM and rSVM-LP produce poor classifiers for this data set. The margins for both are extremely large and do not fit into the scale of the plot.

5. Conclusions

In this paper, we presented four new MILP formulations for SVM. The formulations are for ramp loss SVM with L_1-norm regularization. When the norm for the SVM regularization term is changed, incorporating the kernel trick is non trivial which is why several possible L_1-norm SVMs are possible.

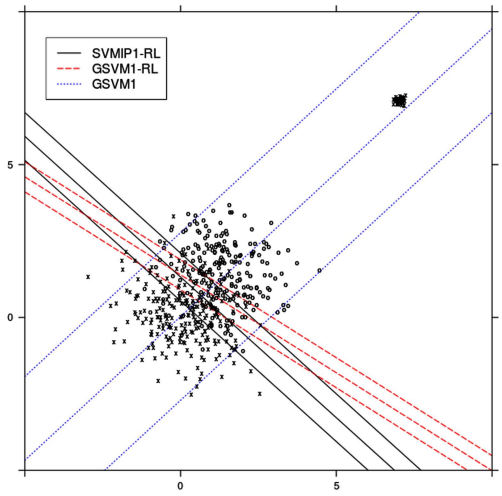
This paper represents to our knowledge the first time that the LP-based SVM versions have been compared head-to-head the same datasets. Further, we explore the benefits of modifying these formulations to increase robustness.

Ramp loss SVM is used to provide robustness. We observed that some LP-based SVM formulations can exhibit some robustness to outliers without the ramp loss on some datasets. However, adding the ramp loss provided increased accuracy for most datasets.

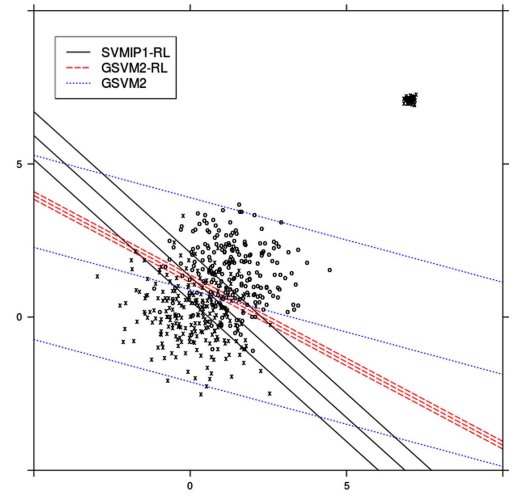
In this paper, we only compared the new SVM versions using outlier-contaminated data with the linear kernel. As demonstrated in [6], it is not clear if additional robustness is needed when using nonlinear discriminant such as SVM with the Gaussian kernel. Because of the bias-variance tradeoff, robust methods for SVM with the linear kernel remain valuable tools when the ratio of the number of observations to the number of attributes is small. In such cases, simpler models tend to provide better generalization.

Software for solving MILPs is more readily available and the technology is more mature than for MIQPs. Though we anticipated computational improvements when switching from

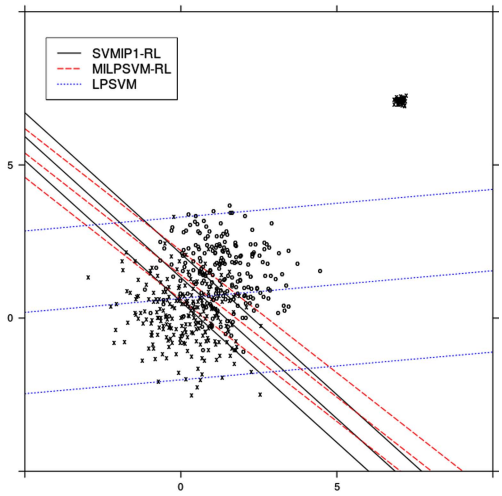
FIGURE 2. Plots of L_1-norm SVM decision rules and margin boundaries with (red, dashed) and without (blue, dotted) ramp loss for a dataset with $n = 500$ observations and $d = 2$ attributes. For each plot, the rule generated by SVMIP1-RL is plotted as solid black lines.



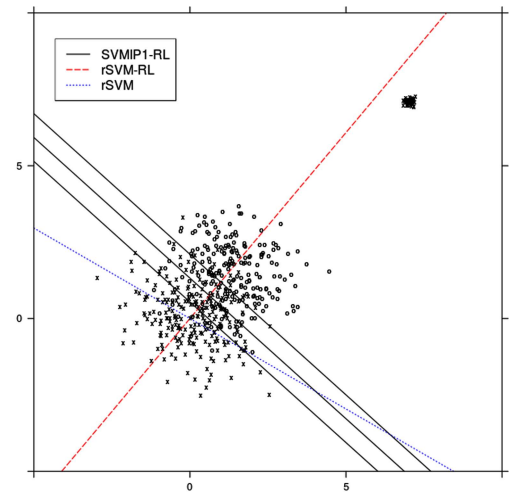
(a) GSVM1 and GSVM1-RL



(b) GSVM2 and GSVM2-RL



(c) LPSVM and MILPSVM-RL



(d) rSVM and rSVM-RL

MIQP formulations to MILP formulations for SVM, we did not see a dramatic effect. Perhaps tailored cutting planes and other enhancements for the formulations presented here can reduce the computation time for the MILP instances.

After the initial submission of this article, the authors became aware of recent work in handling indicator constraints in MILP; namely, the work of Belotti et al. [2]. With these advances, the formulations proposed here may prove to be more computationally tractable than the big-M formulations that we used in our experiment.

6. Acknowledgments

This work was supported in part by NIH awards 2P60MD002256-06 and 8U54HD080784 and an award from The Thomas F. and Kate Miller Jeffress Memorial Trust, Bank of America, Trustee. This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed

in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[2] P. Belotti, P. Bonami, M. Fischetti, A. Lodi, M. Monaci, A. Nogales-Gómez, and D. Salvagnin. On handling indicator constraints in mixed-integer programming. Technical Report OR-14-20, Università di Bologna, 2014.

[3] K.P. Bennett and O.L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[4] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, ACM, New York, NY, USA, 1992.

[5] L. Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–824, 1998.

[6] J.P. Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59(2):467–479, 2011.

[7] E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers & Operations Research*, 40:150–165, 2013.

[8] E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Heuristic approaches for support vector machines with the ramp loss. *Optimization Letters*, 8:1125–1135, 2014.

[9] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[10] I. Hadzic and V. Kecman. Support vector machines trained by linear programming: theory and application in image compression and data classification. In *Proceedings of the 5th Seminar on Neural Network Applications in Electrical Engineering*, pages 18–23, 2000.

[11] *IBM ILOG CPLEX 12.6 Reference Manual*. IBM. 2014.

[12] V. Kecman and T. Arthanari. Comparisons of qp and lp based learning from empirical data. In *Proceedings of the 14th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems: Engineering of Intelligent Systems*, IEA/AIE '01, pages 326–332, London, UK, 2001. Springer-Verlag.

[13] O.L. Mangasarian. Generalized support vector machines. In *Advances in Large Margin Classifiers*, pages 135–146. MIT Press, 1998.

[14] O.L. Mangasarian and D.R. Musicant. Lagrangian support vector machines. *J. Mach. Learn. Res.*, 1:161–177, September 2001.

[15] C. Orsenigo and C. Vercellis. Multivariate classification trees based on minimum features discrete support vector machines. *IMA Journal of Management Mathematics*, 14:221–234, 2003.

[16] C. Orsenigo and C. Vercellis. Evaluating membership functions for fuzzy discrete SVM. *Lecture Notes in Artificial Intelligence: Applications of Fuzzy Sets Theory*, 4578:187–194, 2007.

[17] C. Orsenigo and C. Vercellis. Softening the margin in discrete SVM. *Lecture Notes in Artificial Intelligence: Advances in Data Mining*, 4597:49–62, 2007.

[18] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge UP, 2004.

[19] X. Shen, G.C. Tseng, X. Zhang, and W.H. Wong. On ψ-learning. *Journal of the American Statistical Association*, 98:724–734, January 2003.

[20] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[21] B. Ustun, S. Tracà, and C. Rudin. Supersparse linear integer models for predictive scoring systems. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue, WA, 2013.

[22] Z. Wang and S. Vucetic. Fast online training of ramp loss support vector machines. In *IEEE International Conference on Data Mining*, pages 569–577, Los Alamitos, CA, USA, 2009. IEEE Computer Society.

[23] Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102:974–983, 2007.

[24] W. Zhou, L. Zhang, and L. Jiao. Linear programming support vector machines. *Pattern Recognition*, 35:2927–2936, 2002.