ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Argonne, Illinois 60439

Mixed-Integer Support Vector Machine

Wei Guan, Alex Gray, and Sven Leyffer

Mathematics and Computer Science Division

Preprint ANL/MCS-P1697-1209

December 2, 2009

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

Mixed-Integer Support Vector Machine

Wei Guan
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
wguan@cc.gatech.edu

Alex Gray
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
agray@cc.gatech.edu

Sven Leyffer
Mathematics & Computer Science Division
Argonne National Laboratory
Argonne, IL 60439
leyffer@mcs.anl.gov

Abstract

In this paper, we propose a formulation of a feature selecting support vector machine based on the L0-norm. We explore a perspective relaxation of the optimization problem and solve it using mixed-integer nonlinear programming (MINLP) techniques. Given a training set of labeled data instances, we construct a max-margin classifier that minimizes the hinge loss as well as the cardinality of the weight vector of the separating hyperplane $|| w ||_0$, effectively performing feature selection and classification simultaneously in one optimization. We compare this relaxation with the standard SVM, recursive feature elimination (RFE), L1-norm SVM, and two approximated L0-norm SVM methods, and show promising results on real-world datasets in terms of accuracy and sparsity.

1 Introduction

Feature selection is necessary for many classification task such as microarray analysis, mass spectrometry analysis, biomedical image analysis, and other modern applications. The data sets from these domains are typically high dimensional, while only a few of the features may be needed or helpful for the learning task. Many of the features may be (i) noisy or irrelevant such that ignoring them would improve the generalization ability of the classifier; (ii) redundant (for example, linear combinations of other features) such that eliminating them would not drastically change the prediction performance. Other advantages of feature selection can include reduced computational cost in classifier learning, and better interpretability, for example, in microarray analysis, there is a need to find critical genes that are disease related.

We address feature selection in the context of linear SVM learning [21] for binary classification problems. The main goal is to select out an optimal feature subset with as few features as possible while preserving or improving the discriminative ability of a classifier. Existing approaches of feature selection for SVMs fall into three categories: filter-based methods, wrapper-based methods, and embedded algorithms. Filter-based methods adopt feature ranking strategies disjoint from SVM training [10], such as t2-statistics, and signal-to-noise ratio, etc. Wrapper-based methods order features based on the SVM hyperplane parameters or SVM performance on the training dataset [9, 19], such as backward/forward selection, and recursive feature elimination (RFE), etc. Embedded algorithms augment the SVM formulation, and seek to learn the optimal feature subset as well as the SVM classifier simultaneously [3, 7, 17, 18, 22, 23], such as Feature Selection Concave (FSV), and L1-norm SVM, etc.

In this paper, we propose a feature-selecting vector machine derived from the L0-norm SVM formulation. Unlike L1-norm SVM, which performs feature selection as a by-product because of the resulting sparse solution, L0-norm SVM directly minimizes on both hinge loss and cardinality of its weight vector. In the context of regression, where feature selection has been most throughly studied, it has been pointed out that though the L1 penalization yields sparse solutions, the estimates can be biased since larger penalties are imposed on larger weight coefficients [6]. A recent comparison study of least absolute shrinkage and selection operator (LASSO) regression [20] and forward stepwise regression, which is a greedy surrogate of L0-regularization, further indicated that L1-regularization never outperforms L0-regularization by more than a constant factor, and in some cases, using an L1-norm penalty is much worse than an L0-norm penalty [14]. This comparison analysis also pointed out that "an approximation solution to the right problem can be better than the exact solution to the wrong problem" [14]. Our study follows this guideline. However, optimizing the L0-norm SVM is a NP-hard problem [1]. Previous work in this direction includes adopting smoothed approximations of the L0-norm [3, 23, 24], using adaptive scaling parameters to control the sparsity [22, 7, 18], and exploring the convex relaxations of the cardinality constraint [4].

In this work, we apply mixed-integer nonlinear programming (MINLP) techniques to reformulate the L0-norm SVM with the introduction of indicator variables and perspective relaxation [8], and then solve the resulting mixed-integer quadratic program. Empirical comparison of our proposed mixed-integer SVM method with the standard SVM method [21], RFE method [9], L1-norm SVM method [3, 15], FSV method [3], and Weston's method [22], demonstrates either sparser solutions with roughly identical classification performance, or an increase in classification performance with similar or sparser representations.

In the next section, we briefly summarize the SVM learning problems, and L1-norm SVM. In section 3, we describe the mixed-integer SVM problem and its convex relaxation formulated as mixed-integer quadratic problems. Section 4 presents the comparison of the methods on four data sets from the UCI repository. Finally, in section 5, we conclude and provide several directions for future work.

2 Support Vector Machines

Given a dataset $S = \{x_i, y_i\}_{i=1}^m$ ($x_i \in R^n$ is the feature vector of ith training instance and $y_i \in \{0, 1\}$ is the corresponding label), for two-class classification problems, SVM learns the separating hyperplane $wx = \gamma$ that maximizes the margin distance $\frac{2}{||w||_2^2}$, where w is the weight vector and γ is the bias. Defining ξ as the slack parameters (for describing the training errors), $c > 0$ as the error penalty parameter, diagonal matrix $Y \in R^{m\times m}$ with $Y_{ii} = y_i$, data matrix $X = [x_1, x_2, \cdots, x_m]^T$, vector $e_k = [1, 1, \cdots, 1]^T \in R^k$, and identity matrix $I_k \in R^{k\times k}$, we can formulate SVM learning problem into the following convex optimization.

$$\begin{array}{ll} \min_{w,\gamma,\xi} & \frac{1}{2} \mid\mid w \mid\mid_2^2 + c \mid\mid \xi \mid\mid_1 \\ \text{s.t.} & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \end{array} \tag{1}$$

Defining $\alpha \in R^m$ as the Lagrange multiplier and H as the kernel matrix with $H_{ij} = y_i y_j x_i \cdot x_j$, then the dual problem can be represented as:

$$\begin{array}{ll} \min_{\alpha} & \frac{1}{2}\alpha^T H\alpha - e_m^T\alpha \\ \text{s.t.} & e_m^T Y\alpha = 0, 0 \leq \alpha \leq ce_m \end{array} \tag{2}$$

The optimal weight vector is then computed as $w = X^T Y\alpha$ (α is usually sparse vector, nonzero only on the support vectors). The optimal discrimination function for a data instance x is $f(x) = w \cdot x - \gamma = \alpha^T YXx = \sum_{i=1}^m y_i\alpha_i x_i \cdot x$. The prediction label is $+1$ if $f(x) > 0$ and -1 otherwise.

2.1 L1-norm SVM

L1-norm SVM proposed by Bradley & Mangasarian (1998) [3] solves the following optimization problem. It performs feature selection as a by-product of the resulting sparse solution.

$$\begin{array}{ll} \min_{w,\gamma,\xi} & \| w \|_1 + c \| \xi \|_1 \\ \text{s.t.} & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \end{array} \tag{3}$$

Define $w = p - q; p, q \geq 0$. Problem (3) is then equivalent to the linear programming problem.

$$\begin{array}{ll} \min_{p,q,\xi} & e_n^T(p+q) + c e_m^T \xi \\ \text{s.t.} & YXp - YXq - \gamma Y e_m + \xi \geq e_m \\ & p, q \geq 0, \;\; \xi \geq 0 \end{array} \tag{4}$$

Mangasarian [15] proposed a fast algorithm that solves the corresponding asymptotic exterior penalty problem of the dual problem through Newton's method. Because of its computational efficiency and the empirically sparse solutions, L1-norm SVM method and its variants have been applied to various problems in computation biology and many other domains.

3 Mixed-Integer SVM

In this paper, we consider the following L0-norm SVM formulation:

$$\begin{array}{ll} \min_{w,\gamma,\xi} & \| w \|_0 + c \| \xi \|_1 \\ \text{s.t.} & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \end{array} \tag{5}$$

However, the problem of minimizing the L0-norm is proved to be NP-hard [1]. Inspired by Gunluk and Linderroth's work on perspective relaxation of indicator-induced MINLP problems [8], we modify the above problem by introducing z_j, the indicator variable ($z_j \in \{0, 1\}$, $z_j = 0 \Rightarrow w_j = 0$); u_j, the squared upper bound of the weight element w_j; and the perspective constraints, represented by the conic constriants: $w_j^2 \leq z_j u_j$. These define a convex hull of $w_j^2 = z_j u_j$, which is the equality we want to enforce. The proposed mixed-integer SVM can then be formulated as the following mixed-integer quadratically constrained quadratic program.

$$\begin{array}{ll} \min_{z,u,w,\gamma,\xi} & a e_n^T z + \frac{1}{2} e_n^T u + c e_m^T \xi \\ \text{s.t.} & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \\ & w_j^2 - z_j u_j \leq 0 \\ & 1 - \sum_j z_j \leq 0 \quad j = 1, \cdots n \\ & z \in \{0, 1\}^n, \; u \geq 0, \;\; \xi \geq 0 \end{array} \tag{6}$$

where vector $z = [z_1, \cdots, z_n]^T$, $u = [u_1, \cdots, u_n]^T$, and constants $a, c > 0$ adjust the trade-off between the cardinality of the weight vector and the hinge loss.

Basically, the objective function tries to minimize the sum of the L0-norm $\sum_j z_j$, the L2-norm upper bound $\sum_j u_j$, and the hinge loss $\sum_i \xi_i$. The first type of constraints $Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0$ regulates the classification error for each training instance. The second type of constraints $w_j^2 \leq u_j z_j$ enforce that i) $w_j = 0$ when $z_j = 0$, and ii) $u_j = w_j^2$ at optimal. The third type of constraints $\sum_j z_j \geq 1$ ensure that at least one feature should be selected (having nonzero indicator variable z_j).

However, solving this problem directly with the existing MINLP tools such as Bonmin [2], Cplex [12] or MINLP [13] fails. The experiments of optimizing this problem over even small datasets resulted in either infeasible states or unsatisfying solutions with only one nonzero indicator variable. We believe that the failure of the nonlinear solvers is due to a failure of a constraint qualification: whenever $z_j = 0$ during the tree-search or in the solution of continous subproblems in Bonmin, the relaxation contains a constraint $w_j^2 \leq 0$, which violates Slater's constraint qualification [11]. While it is in principle straightforward to remedy this situation by preprocessing the constraint $w_j^2 \leq 0$ and replacing it by $w_j = 0$, current nonlinear solvers do not perform this operation. The errors that we observe from the nonlinear solvers are consistent with a failure of a constraint qualification.

To remedy this adverse situation, we further relax the conic constraints $w_j^2 \leq u_j z_j$ into big-M constraints $| w_j | \leq M z_j$, where M is a fixed large number (M was set to 10^4 in our experiments), resulting in a mixed-integer quadratic problem (P_1) .

$$
\begin{array}{ll}
\min_{z,w,\gamma,\xi} & ae_n^T z + \frac{1}{2} e_n^T u + c e_m^T \xi \\
\text{s.t.} & Y(Xw - \gamma e_m) + \xi \geq e_m, \xi \geq 0 \\
 & \mid w_j \mid - M z_j \leq 0 \\
 & 1 - \sum_j z_j \leq 0 \quad j = 1, \cdots n \\
 & z \in \{0,1\}^n, \ \xi \geq 0
\end{array}
\qquad (P_1)
$$

4 Results and Discussion

We compared the performance of the proposed MI-SVM (Eqn. (P_1)), with the standard SVM (Eqn. (1)), RFE method [9], L1-norm SVM (Eqn. (3)), and two most commonly-cited L0-norm SVM approximation methods: FSV [3] and R2W2 (Weston's method) [22] on four data sets from the UCI repository [16]. We used Chang and Lin's LibSVM packages [5] and Mangasarian's L1-norm SVM code [15]. We wrote our own RFE, FSV, R2W2 code in MATLAB. CPLEX11.1 has been used for solving the mixed-integer quadratic problem, Eqn. (P_1).

- **Ionosphere** dataset consists of 351 instances with 34 features. There are 225 radar returns termed "good" or showing some type of structure in the ionosphere, and 126 radar returns termed "bad"; their signals pass through the ionosphere.
- **Wisconsin prognostic breast cancer** dataset consists of 198 instances with 32 numerical features representing follow-up data of the patients. Two of its variants are used.

  - The first data set includes 28 patients who had a cancer recurrence in less than 24 months and 127 patients who didn't have a cancer recurrence in less than 24 months.

  - The second variant of the data set contains 41 patients with a cancer recurrence in less than 60 months, and 69 patients which cancer had not recurred in less than 60 months.
- **SPECTF heart** dataset: the training dataset consists of 80 instances with 44 features (40 instances labeled with "1" and "0", respectively); the testing dataset consists of 187 instances with 172 instances labeled with "1" and 25 labeled with "0".

We estimate the generalization ability of each method via 10-fold cross validation (10-fold CV), except for SPECTF as its training and testing split are given. Note that we need to tune the parameter c of the standard SVM method and the RFE method, parameters δ, c of L1-norm SVM, parameter λ for FSV method, and parameters a, c of the MI-SVM methods for the performance evaluation. We employ the following tuning procedure on each data set: for each parameter setting, we perform a 10-fold CV, and the score for this parameter setting is the averaged training accuracy over cross-validation; while for SPECTF data set, we use the training accuracy as its score. Then we select the parameter setting with the best score (ties are broken by choosing the sparser solutions). The candidate parameter values used for the experiments were $c \in \{2^{-7}, \cdots, 2^{-1}, 1, 2^1, \cdots, 2^7\}$, $\delta \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3\}$, $a \in \{2^{-3}, \cdots, 2^{-1}, 1, 2^1, \cdots, 2^3\}$, $\lambda \in \{0.05, 0.1, 0.15, \cdots, 0.85, 0.9, 0.05\}$. Finally, due to numerical reasons, for FSV and MI-SVM methods (denoted this approach as MI-SVM^1), the optimal weight elements with small relative magnitude, i.e. $\frac{|w_j|}{max_k(|w_k|)} < 10^{-4}$, are set to zero. For MI-SVM method, since we also obtain the optimal indicator variable assignment after solving Eqn. (P_1), thus we would apply standard SVM to the selected feature subset, consisting of features with non-zero indicator variables, and obtain the final weight values (denoted this approach as MI-SVM^2).

Table 1: Feature Selection Performance (Number of Features Maintained)

	SVM	RFE	L1-SVM	FSV	R2W2	MI-SVM^1	MI-SVM^2
Ionosphere	33 ± 0.0	**22.7** ± 3.9	29.8 ± 1.2	30.3 ± 2.0	30.1 ± 3.8	30.8 ± 1.1	30.7 ± 1.2
WPBC24	32 ± 0.0	27.7 ± 2.5	24.3 ± 1.2	**5.4** ± 0.8	27 ± 7.8	19.1 ± 1.1	24.3 ± 2.1
WPBC60	32 ± 0.0	23.5 ± 3.7	25.2 ± 1.8	19.3 ± 1.6	21.8 ± 5.7	21.2 ± 1.5	**16.1** ± 1.2
SPECTF	44	14	28	34	21	**12**	**12**
on-average	35.5	22	26.8	22.3	25	**20.8**	**20.8**

Table 2: Classification Performance (Accuracy)

	SVM	RFE	L1-SVM	FSV	R2W2	MI-SVM^1	MI-SVM^2
Ionosphere	88.4 ± 6.4	87.9 ± 5.9	**88.7** ± 5.5	88.1 ± 7.3	**88.7** ± 5.5	**88.7** ± 6.1	**88.7** ± 6.1
WPBC24	78.8 ± 5.0	78.1 ± 8.7	81.5 ± 8.8	70.8 ± 10.6	78.2 ± 5.6	81.3 ± 6.8	**82.1** ± 7.6
WPBC60	66.2 ± 8.0	65.6 ± 11.7	58.2 ± 9.5	61.9 ± 13.4	**66.5** ± 8.2	60.9 ± 14.3	63.6 ± 8.9
SPECTF	72.2	75.4	58.8	73.8	73.8	**76.5**	**76.5**
on-average	76.4	76.8	71.8	73.7	76.8	76.8	**77.7**

Table 1 summarizes the feature selection performance, measured by sparsity, that is the number of features selected by each method, while Table 2 describes the classification performance, measured by testing accuracy of each classifier. For the first three data sets, we give the average sparsity and testing accuracy as well as their standard deviations over the 10-fold CV. Overall, the experiment results show that MI-SVM methods are able to learn sparser representations with roughly identical or increased classification performance. And MI-SVM^2, the approach of applying standard SVM onto the feature selection results of Eqn. (P_1) (features with non-zeros indicator variables) had a higher prediction performance than MI-SVM^1, the approach of simply thresholding out the optimal weights of Eqn. (P_1) that having relative small magnitude.

MI-SVM^2 approach achieved the best evaluation performance (77.7% testing accuracy averaged over the four datasets) with an average sparsity of 20.8. MI-SVM^1, RFE, R2W2 methods had the second best testing accuracy (76.8% on average) with MI-SVM^1 having the smallest average sparsity (20.8). While L1-norm SVM had the worst evaluation performance (71.8% averaged testing accuracy) with the an average sparsity of 26.8, and FSV method had the second worst performance (73.7%). In datasets such as WPBC24, SPECTF, the testing accuracy increase significantly when using MI-SVM, which indicates that some of the features in these datasets may be noise or irrelevant. In other dataset like IONOSPHERE, WPBC60, the accuracy remains roughly the same while sparsity increases, which suggests that these datasets may contain redundant features. In both cases, MI-SVM is able to learn a lower dimensional representation with at least comparable classification performance. The comparison analysis indicates that our MI-SVM method realizes a suitable trade off between the classification errors and the number of selected features. Moreover, the sparse representations learned with the MI-SVM method are generally more predictive than those produced by the L1-norm SVM or other L0-norm SVM approximations such as FSV method.

5 Conclusion

In this work, we propose a feature-selecting SVM that uses mixed-integer quadratic programming to solve a robust convex relaxation of the L0-norm SVM. Note that this study is in contrast to previous work, which either used a smoothed penalty function that approximates the L0-norm in the objective or used adaptive scale parameters, and then solved through convex optimization techniques. Empirical results show either an increase in sparsity with comparable classification performance, or an increase in classification accuracy compared to the most widely-used approaches: the standard SVM, RFE, L1-norm SVM, FSV and R2W2 (Weston's method) methods. We believe the approach is promising for feature-selecting SVM learning, and demonstrates effective MINLP techniques which have not previous been widely used in machine learning.

Several questions arise from this study. First, is the suggested way to approximate the conic constraints using the big-M method the best? We will investigated other convex functions for approximation. Secondly, the scalability is now constrained by the ability of Cplex to handle high dimensional data sets. We are currently investigating alternative tools and custom approaches to increase performance as well as the computational cost.

References

[1] Amaldi, E. & Kann, V. (1998) On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, **209**(1-2):237-260.

[2] Bonami, P., Biegler, L.T., Conn, A.R., Cornuéjols, G., Grossmann, I.E., Laird, C.D. & et al. (2008) An algorithmic framework for convex mixed integer nonlinear programs. *Discrete Optimization*, **5**(2):186-204.

[3] Bradley, P. & Mangasarian, O.L. (1998) Feature selection via concave minimization and support vector machines. *In Proceedings of the Fifteenth International Conference (ICML)*, pp. 82-90.

[4] Chan, A.B., Vasconcelos, N. & Lanckriet, G.R.G. (2007) Direct convex relaxations of sparse svm. In proceedings of the international conference on Machine learning, pp. 145–153, ACM New York, NY, USA.

[5] Chang, C.C. & Lin, C.J. (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456):1348-1360, 2001.

[7] Grandvalet, Y. & Canu, S. (2003) Adaptive scaling for feature selection in SVMs. *Advances in Neural Information Processing Systems*, pp. 569-576.

[8] Gunluk, O. & Linderoth, J. (2008) Perspective Relaxation of Mixed Integer Nonlinear Programs with Indicator Variables. *In Proceedings of International Conference on Integer Programming and Combinatorial Optimization (IPCO)*, pp. 1-16.

[9] Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002) Gene selection for cancer classification using support vector machines. *Machine learning*, **46**(1):389-422.

[10] Guyon, I. and Elisseeff, A. (2003) An introduction to variable and feature selection. *The Journal of Machine Learning Research*, **3**:1157-1182.

[11] Hiriart-Urruty, J.B. & Lemarechal, C. (1993) *Convex analysis and minimization algorithms*. Springer-Verlag

[12] ILOG CPLEX package. http://www.ilog.com/products/cplex/product/algorithms.cfm

[13] Leyffer, S. (199) User manual for MINLP BB. *University of Dundee Numerical Analysis Report*, **234**.

[14] Lin, D., Pitler, E., Foster, D.P. & Ungar, L.H. (2008) In Defense of l0. *In ICMLUAICOLT workshop on Sparse Optimization and Variable Selection*.

[15] Mangasarian, O.L. (2007) Exact 1-Norm Support Vector Machines Via Unconstrained Convex Differentiable Minimization (Special Topic on Machine Learning and Optimization). *Journal of Machine Learning Research*, **7**(2):1517-1530.

[16] Murphy, P.M. & Aha, D.W. (1992) UCI repository of machine learning databases. Technical report, Department of Information and Computer Science, Univerisity of California, Irvine, 1992. www.ics.uci.edu/mlearn/MLRepository.html

[17] Neumann, J., Schnörr, C. & Steidl, G. (2005) Combined SVM-based feature selection and classification, *Machine Learning*, **61**(1):129-150.

[18] Peleg, D., & Meir, R. (2004) A feature selection algorithm based on global minimization of a generalization error bound. *Advances in neural information processing systems*.

[19] Rakotomamonjy, A. (2003) Variable selection using SVM based criteria. *Journal of Machine Learning Research*, **3**:1357-1370.

[20] Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288, Blackwell Publishers.

[21] Vapnik, V. (1995) *The Nature of Statistical Learning Theory.* Springer.

[22] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. & Vapnik, V. (2001) Feature selection for SVMs. *Advances in neural information processing systems*, pp. 668-674.

[23] Weston, J., Elisseeff, A., Schölkopf, B. & Tipping, M. (2003) Use of the zero norm with linear models and kernel methods. *The Journal of Machine Learning Research*, **3**:1439-1461.

[24] Zhang, H.H., Ahn, J., Lin, X. & Park, C. (2006) Gene selection using support vector machines with nonconvex penalty. *Bioinformatics*, **22**(1):88-95.