Automatic discrimination of mislabeled training points for large margin classifiers

Rómer Rosales[1], Glenn Fung[1] , Wei Tong[2]

[1] IKM CKS, Siemens Medical Solutions, PA USA
[2] Department of Computer Science Michigan State University, MI USA

In the context of supervised learning of classifiers, label noise or labeling errors in the training data can affect significantly the performance of the learning method. This is particularly true in large margin hyperplane-based methods where the classifier is trained by minimizing a function that penalizes quantitatively how far a training point is from the 'correct' side of the hyperplane. In this case (as we will show empirically later) a few mislabeled data points in the training set could lead to a significant drop in classification performance, even when regularization is properly taken into account. This problem arises in basically any learning domain where the reliability of the training data (particularly the labels) has been compromised.

This problem has been approached in the context of *robust statistics* where the goal is to create estimators that can properly handle small deviations from the basic model assumptions. For example, the median is a simple robust estimate of the *center* of the distribution. In machine learning, the popular RANSAC (Random Sample Consensus)[1] method is an example of model estimation that is robust to some outliers in the data.

In this work we explore this problem in the context of large-margin classifiers. We use several large-margin formulations to take into account data points where the corresponding label has a high probability of being wrong given all the training data. The problem can be seen as a SVM-like formulation where k (provided by the user) represents the number of training points that are required to be misclassified with a large misclassification penalty M.

The resulting formulation can be written as the mixed integer programming (MIP) (presented next). We have m datapoints in the n-dimensional real space $\mathbb{R}^n$ which are represented by the $m \times n$ matrix A. The datapoints are in two classes $A+$ or $A-$ and the class labels of each data point is specified by a given $m \times m$ diagonal matrix D with plus ones or negative ones along its diagonal.

$$\begin{array}{ll} \min\limits_{(\mathbf{w},\gamma,\mathbf{y},\mathbf{y}^*)} & C\sum_i y_i + \frac{1}{2}(\mathbf{w}^\top\mathbf{w}) \\ \text{s. t.} & D(A\mathbf{w}-\gamma\mathbf{e}) + \mathbf{y} \geq \mathbf{e} \\ & y_i^* \in \{0,1\} \\ & My_i^* \leq y_i, \text{ M is a big number} \\ & \sum_i y_i^* = k \end{array} , \tag{1}$$

where C is the parameter with $C > 0$, $\mathbf{y} \in \mathbb{R}^m$ is the classification error associated with each datapoint. $\mathbf{e}$ is a column vector with all ones. $\mathbf{w}$ is the normal to the bounding planes:$\mathbf{x}^\top\mathbf{w} = \gamma + 1$ and $\mathbf{x}^\top\mathbf{w} = \gamma - 1$. Here the optimal solution(s) aim to assign the large M penalties to the points that are most unlikely to belong to its (labeled) class.

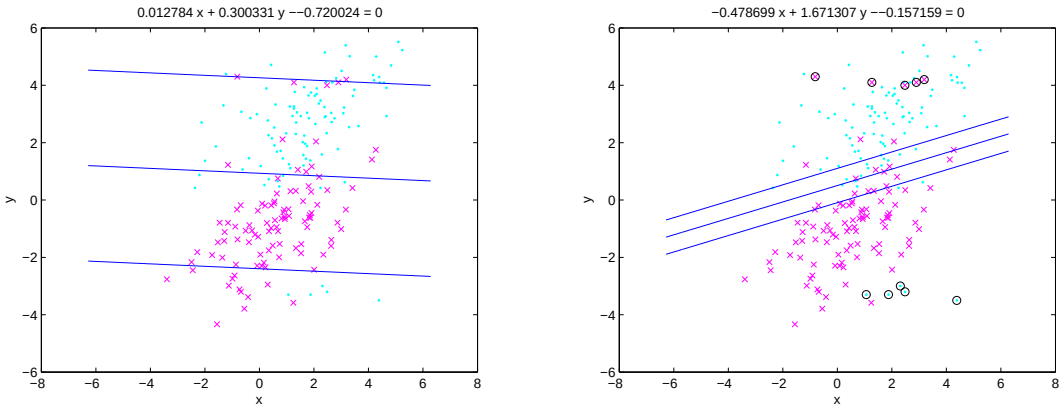
We can also think of the opposite formulation, where the classification errors of k points are allowed to be ignored in the objective function. This leads to the following formulation:

$$\begin{array}{ll} \min\limits_{(\mathbf{w},\gamma,\mathbf{y},\mathbf{y}^*)} & C\sum_i y_i^* y_i + \frac{1}{2}(\mathbf{w}^\top\mathbf{w}) \\ \text{s. t.} & D(A\mathbf{w}-\gamma\mathbf{e}) + \mathbf{y} \geq \mathbf{e} \\ & y_i^* \in \{0,1\} \\ & \sum_i y_i^* = k \end{array} \tag{2}$$

For both formulations presented above we explored several variations (depending on the cos function and regularization used) and relaxations that lead to very efficient approximate solutions including several convex relaxations: Semidefinite programming (similarly to the one presented in [2], Quadratic programming and Linear programming relaxations as well as several non-convex formulations and in particular we are very

encouraged by an effective branch and bound strategy that can be used to find good solutions when both the cost function and the regularization term have Gaussian priors (L2 regularizations) and the problem becomes an Integer least squares formulation.

In order to test the effectiveness of the original formulation, two hundred samples are generated from two Gaussian distribution, one hundred samples in each class. Results are shown in Fig. 1, where we can see the the outliers can be identified correctly in this toy data set.



(a) Decision boundary of standard svm (b) Result of the MIP problem

Fig. 1. Five data points in each class were manually moved away from their class center to consider them as (unknown) outliers. Figure shows toy data and the decision planes of (a) a standard SVM formulation (b) our original formulation

We now explore real datasets labeled 1-4 where text fragments from medical records are labeled to identify the appearance of certain medical conditions using a bag-of-words-like representation (congestive heart failure, smoking history, joint revision, and contraindication to a betablockers respectively, but more general concepts can be considered as well) and $k < 2\%$ of the available data. These results indicate that the original MIP formulation is effective in both improving the classification error (*Err* column) and also at reducing the number of support vectors necessary to achieve better performance (*num of SV* column). The solution to the optimization problem includes identification of potentially mislabeled data that can be used as a feedback mechanism for the labeling process. Current work includes the design of convex/non-convex relaxations to the two problem formulatios above.

References

[1] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 1981.

[2] L. Xu, Koby Crammer and Dale Schuurmans. Robust Support Vector Machine Training via Convex Outlier Ablation. Proceedings of theTwenty-First National Conference on Artificial Intelligence. AAAI 2006.

Topic: learning theory. Preference: oral/poster

(no. examples × dims)	SVM Err	SVM num of SV	SVM optC	MIP Err	MIP num of SV	MIP optC
dataset 1 (219 × 103)	0.0548	18	0.1	0.0183	4	1
dataset 2 (584 × 103)	0.0240	22	10	0.0171	11	100
dataset 3 (582 × 103)	0.0258	30	1	0.0172	11	10
dataset4 (269 × 103)	0.0967	26	100	0.0409	12	100

Table 1. Standard SVM results (columns 2-4) and new formulation results (columns 5-7)