A Maximum K-Min Approach for Classification

Mingzhi Dong and Liang Yin
Beijing University of Posts and Telecommunications, China
mingzhidong@gmail.com, yin@bupt.edu.cn

Abstract

In this paper, a general Maximum K-Min approach for classification is proposed, which focuses on maximizing the gain obtained by the K worst-classified instances while ignoring the remaining ones. To make the original optimization problem with combinational constraints computationally tractable, the optimization techniques are adopted and a general compact representation lemma is summarized. Based on the lemma, a Nonlinear Maximum K-Min (NMKM) classifier is presented and the experiment results demonstrate the superior performance of the Maximum K-Min Approach.

Introduction

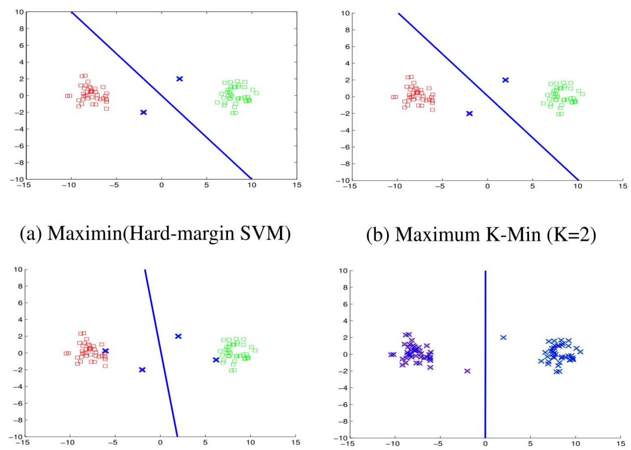
In the realm of classification, maximin approach, which pays strong attention to the worst situation, is widely adopted and it is regarded as one of the most elegant ideas. Hard-margin Support Vector Machine (SVM) (Vapnik 2000) is the most renowned maximin classifier and it enjoys the intuition of margin maximization.

However, maximin methods based on the worst instance may be sensitive to outliers/noisy points near the boundary, as shown in Figure 1(a). Therefore, after the previous work on a special case of naive linear classifier (Dong et al. 2012), in this paper, we propose a general Maximum K-Min approach for classification, which focuses on maximizing the gain obtained by the K worst-classified instances while ignoring the remaining ones, as exemplified in Figure 1(c)(d).

Maximum K-Min Approach

Maximum K-Min Gain can be readily solved via Minimum K-Max Loss. Since Minimum K-Max are more frequently discussed in the realm of optimization, in the following sections, the Maximum K-Min approach will be discussed and solved via minimizing K-Max Loss.

With the prediction function $f(\boldsymbol{x}_n, \boldsymbol{w})$ of a classifier, $g_n = t_n f_n = t_n f(\boldsymbol{x}_n, \boldsymbol{w})$ represents the classification confidence of the nth training instance $\boldsymbol{x}_n$; $g_n \geq 0$ indicates the correct classification of $\boldsymbol{x}_n$ and the larger g_n is, the more confident the classifier outputs the classification result. Therefore, we can define a vector $\boldsymbol{g} = (g_1, \ldots, g_N)^T \in R^N$ to represent a kind of classification gain of all training instances. Then, $\boldsymbol{l} = -\boldsymbol{g} = (-g_1, \ldots, -g_N)^T = (-t_1 f_1, \ldots, -t_N f_N)^T \in R^N$ can be introduced to represent a kind of classification loss. After that, by sorting $\boldsymbol{l}_n$ in descent order, we can obtain $\boldsymbol{\theta} = (\theta_{[1]}, \ldots, \theta_{[N]})^T \in R^N$ where $\theta_{[n]}$ denotes the nth largest element of $\boldsymbol{l}$, i.e. the ith largest loss of the classifier. Based on the above discussion, K-Max Loss of a classifier can be defined as $\Theta_K = \sum_{n=1}^{K} \theta_{[n]}$, with $1 \leq K \leq N$ and it measures the sum loss of the worst K training instances during classification. Therefore, the minimization of K-Max Loss with respect to the parameters can result in a classifier which performs best with regard to the worst K training instances.



(a) Maximin(Hard-margin SVM) (b) Maximum K-Min (K=2)

(c) Maximum K-Min (K=4) (d) Maximum K-Min (K=N)

Figure 1: A comparison of Maximin Approach (Hard-Margin SVM) and linear Maximum K-Min Approach(Support Vectors/Worst K Instances are marked $\times$).

Original Objective Function

Thus, the original objective function of Maximum K-Min approach is to minimize the K-Max Loss and has the following formulation

$$\min \quad \Theta_K = \sum_{i=1}^{K} \theta_{[i]}, \tag{1}$$

Equation 1 is a convex optimization problem as follows

$$\begin{array}{ll} \min\limits_{s} & s, \\ \text{s.t.} & y_{i_1} + \cdots + y_{i_K} \leq s, \\ \text{where} & 1 \leq i_1 \leq \cdots \leq i_K \leq N, \end{array}$$

Compact Representation of Θ_K

However, it's prohibitive to solve an optimization problem under C_N^K inequality constrains. We need to sketch out a compact formulation. Thus, with the convex optimization techniques (Boyd and Vandenberghe 2004), we introduce the following lemma.

Lemma 1 For a fixed integer K, $\Theta_K(\boldsymbol{y})$, the sum of K largest elements of a vector $\boldsymbol{y}$, is equal to the optimal value of a linear programming problem as follows,

$$\begin{array}{ll} \max\limits_{\boldsymbol{z}} & \boldsymbol{y}^T\boldsymbol{z}, \\ \text{s.t.} & 0 \leq z_i \leq 1, \\ & \sum_i z_i = K, \end{array}$$

where $i = 1, \ldots, N$,$\boldsymbol{z} = \{z_1, \ldots, z_N\}$. $\blacksquare$

According to Lemma 1, we can obtain an equivalent representation of Θ_K. The dual of this equal representation can be written as

$$\begin{array}{ll} \min\limits_{s,u_i} & Ks + \sum_i u_i \\ \text{s.t.} & s + u_i \geq y_i \\ & u_i \geq 0 \\ \text{where} & i = 1, \ldots, N. \end{array}$$

According to the above conclusion, the minimization of $\Theta_K(\boldsymbol{y})$, i.e. the original Maximum K-Min problem, is equal to the minimization of $Ks + \sum_i u_i$ under $2N$ constraints.

A Nonlinear Maximum K-Min Classifier

To utilize kernel in non-linear separable situations, we make a linear assumption of the prediction function $f(\boldsymbol{x})$

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{n=1}^{N} a_n k(\boldsymbol{x}, \boldsymbol{x}_n) t_n + b, \quad n = 1, \ldots N \tag{2}$$

where $k(\boldsymbol{x}, \boldsymbol{x}_n)$ denotes a similarity measurement between $\boldsymbol{x}$ and $\boldsymbol{x}_n$ via a kind of kernel function. $\boldsymbol{a} = \{a_1, \ldots, a_N\}$ denotes the weight of training samples in the task of classification. b indicates the bias. t_n indicates the corresponding target value and $t_n \in \{-1, 1\}$. The prediction of $\boldsymbol{x}$ is made by taking weighted linear combinations of the training set target values, where the weight is the product of a_n and $s(\boldsymbol{x}, \boldsymbol{x}_n)$. Therefore, the classification confidence measurement $g(\boldsymbol{x})$ equals $t_n f(\boldsymbol{x})$.

According to the above discussion, we can obtain a compact representation for this Nonlinear Maximum K-Min Classifier (NMKM) as follows

$$\begin{array}{ll} \min\limits_{\boldsymbol{a},s,\boldsymbol{u}} & Ks + \sum_{i=1}^{N} u_i, \\ \text{s.t.} & \sum_{n=1}^{N} a_n t_n k(\boldsymbol{x}_i, \boldsymbol{x}_n) + t_n b - s - u_i \leq 0, \\ & u_i \geq 0, \\ & \boldsymbol{a}^T\boldsymbol{a} \leq 1, \\ \text{where} & i = 1, \ldots, N. \end{array}$$

Table 1: Experiment Results of NMKM (Accuracy, the higher the better)

Algorithm	DataSet				
	australian	breast	colon	diabetes	fourclass
$NMKM$	**85.07%**	**96.35%**	84.40%	**74.84%**	**99.88%**
SVM	76.67%	93.54%	**87.20%**	67.37%	99.65%
LR	84.02%	95.07%	84.40%	66.62%	72.03%
	german	heart	ionosphere	liver	splice
$NMKM$	76.11%	80.65%	**94.39%**	**71.42%**	82.25%
SVM	65.75%	73.98%	88.33%	62.56%	**83.50%**
LR	**76.38%**	**81.67%**	82.20%	67.61%	79.38%

[i] NMKM and SVM are deployed with RBF kernel;
[ii] The best result on each dataset is shown in boldface.

As a Quadratically Constrained Linear Programming (QCLP) problem, standard convex optimization methods, such as interior-point methods (Boyd and Vandenberghe 2004), can also be adopted to solve the above problem efficiently and a global maximum solution is guaranteed.

Experiment

The algorithm of NMKM (with RBF kernel) is compared with two state-of-the-art classifiers of Support Vector Machine (SVM) (with RBF kernel) and Logistic Regression (LR) (L2-regularized). Ten UCI datasets for binary classification are adopted in the experiment. Our algorithm is deployed using cvx toolbox[1] in matlab environment and kernel SVM is deployed using libsvm toolbox[2]. The hyperparameters of NMKM and SVM are chosen via cross-validation.

As shown in Table 1, among 10 different datasets, NMKM achieves best accuracy in 6 datasets. SVM performs best in 2 datesets and LR shows best performance in 2 datasets. Among all datasets which SVM or LR performes better than NMKM, NMKM performs slightly weaker than SVM or LR. While for datasets which NMKM gains better performance, the accuracy of NMKM may be much larger than MSVM and LR. The accuracy gap of SVM in dataset *liver* and LR in dataset *fourclass* from NMKM is 8.86% and 27.85% respectively.

Thus we can conclude, compared with SVM and LR, NMKM has shown competitive classification performance, which verifies the efficiency of Maximum K-Min approach.

References

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge Univ Press.

Dong, M.; Yin, L.; Deng, W.; Wang, Q.; Yuan, C.; Guo, J.; Shang, L.; and Ma, L. 2012. A linear max k-min classifier. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2967–2971. IEEE.

Vapnik, V. 2000. *The nature of statistical learning theory*. Springer-Verlag New York Inc.

[1] http://cvxr.com/

[2] www.csie.ntu.edu.tw/ cjlin/libsvm