Transductive Optimization of Top k Precision

Li-Ping Liu[1] **Thomas G. Dietterich**[1] **Nan Li**[*2] **Zhi-Hua Zhou**[2]

[1]EECS, Oregon State University, Corvallis, Oregon 97331, USA
[2]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{liuli@eecs.oregonstate.edu, tgd@oregonstate.edu}, {lin, zhouzh}@lamda.nju.edu.cn

Abstract

Consider a binary classification problem in which the learner is given a labeled training set, an unlabeled test set, and is restricted to choosing exactly k test points to output as positive predictions. Problems of this kind—*transductive precision@k*—arise in many applications. Previous methods solve these problems in two separate steps, learning the model and selecting k test instances by thresholding their scores. In this way, model training is not aware of the constraint of choosing k test instances as positive in the test phase. This paper shows the importance of incorporating the knowledge of k into the learning process and introduces a new approach, Transductive Top K (TTK), that seeks to minimize the hinge loss over all training instances under the constraint that exactly k test instances are predicted as positive. The paper presents two optimization methods for this challenging problem. Experiments and analysis confirm the benefit of incoporating k in the learning process. In our experimental evaluations, the performance of TTK matches or exceeds existing state-of-the-art methods on 7 UCI datasets and 3 reserve design problem instances.

1 Introduction

In the *Transductive Precision@k* problem, the training set and the unlabeled test set are given, and the task is to predict exactly k test instances as positives. The precision of these selected instances—the fraction of correct positive predictions—is the only measure of importance. Our work is motivated by the problem of designing conservation reserves for an endangered species. Suppose a geographical region is divided into equal-sized cells of land. The species is present in positive cells and absent in negative cells. To protect the species, we seek to purchase some cells ("a conservation reserve"), and we want as many of those as possible to be positive cells. Suppose we have conducted a field survey of publicly-owned land to collect a training set of cells. With a fixed budget sufficient to purchase k cells, we want to decide which k privately-owned (and un-surveyed) cells to buy. In this paper, we assume that all cells have the same price. This is an instance of the Transductive Precision@k problem. Other instances arise in information retrieval and digital advertising.

The standard approach to this problem is to first train a classifier or ranker on the training data and then threshold the predicted test scores to obtain the k top-ranked test instances. Any model that outputs continuous scores (e.g., an SVM) can be employed in this two-step process. Better results can often be obtained by bipartite ranking algorithms [1, 4, 8, 10, 14, 15, 20], which seek to minimize a ranking loss (including ranking losses that put more weight on highly-ranked instances). Recent work focuses even more tightly on the top-ranked instances. The MPG algorithm [22] formulates the ranking problem as an adversarial game and can optimize several ranking measures. The Accuracy At The Top (AATP) algorithm [3] seeks to optimize the ranking quality for a specified top quantile of the training data. Maximizing accuracy on the top quantile is intractable, so AATP optimizes a relaxation of the original objective. However, none of the algorithms above explicitly considers the constraint of choosing k test instances in model training.

Unlike the ranking problems discussed so far, our problem is transductive, because we have the unlabeled test examples available. There is a substantial body of research on transductive classification [6, 9, 13, 17]. Most transductive classification algorithms are inspired by either the large margin principle or the clustering principle. The goal of these algorithms is to develop classifiers that will perform well on the entire test set. Some transductive classifiers [6, 9] have a parameter to specify the desired ratio of positive predictions. However, such parameter is mainly used for training a stable classifier, and the ratio not strongly enforced on the test set.

In this paper, we claim that the knowledge of k helps to learn a better model in terms of the measure of top-k precision and that the trained model should be constrained to output a selection of k test instances as positive. We call this constraint the *k-constraint*. The benefit of incorporating k into the learning process can be understood in three ways. First, the k-constraint greatly reduces the hypothesis space and thus reduces the structural risk of the trained model. Second, the algorithm can take advantage of the knowledge of k to jointly optimize the scoring model and the score threshold

[*]Nan Li is now working at Alibaba Group, Hangzhou China.

with respect to the k-constraint. As a comparison, a two-step method trains a ranker that is optimal by some standard, but the ranker together with the threshold may not be optimal for the selection task. Third, the selection of k test points is directly obtained through model training, instead of learning a general classifier or ranker as an intermediate step. Vapnik's principle [21] dictates that we should not solve a more difficult problem on the way to solving the problem of interest.

In this paper, we jointly train the model and determine the threshold to obtain exactly k test instances as positive. We seek a decision boundary that predicts exactly k positives and has high precision on the training data. The paper proceeds as follows. We start by identifying a deterministic relation between the precision@k measure and the accuracy of any classifier that satisfies the k-constraint. This suggests that the learning objective should maximize classifier accuracy subject to the k-constraint. We adopt the space of linear decision boundaries and introduce an algorithm we call Transductive optimization of Top k precision (TTK). In the TTK optimization problem, the objective is to minimize the hinge loss on the training set subject to the k-constraint. This optimization problem is very challenging since it is highly non-convex. We first formulate this problem into an equivalent Mixed Integer Programming (MIP) problem and solve it with an off-the-shelf MIP solver. This method works for small problems. To solve larger problems, we also design a *feasible direction* algorithm, which we find experimentally to converge very rapidly. Finally, our theoretical analysis of the transductive precision@k problem shows that one should train different scoring functions for different values of k. As a byproduct of the work, We also find a problem in the optimization algorithm used in AATP, which solves a problem similar to ours.

In the experiment section, we first present a small synthetic dataset to show how the TTK algorithm improves the SVM decision boundary. Then, we show that our feasible direction method can find solutions nearly optimal as global optimal solutions. In the third part, we compare the TTK algorithm with five other algorithms on ten datasets. The results show that the TTK algorithm matches or exceeds the performance of these state-of-the-art algorithms on almost all of these datasets.

2 The TTK model

Let the distribution of the data be $\mathcal{D}$ with support in $\mathcal{X} \times \mathcal{Y}$. In this work, we assume $\mathcal{X} = \mathcal{R}^d$ and only consider the binary classification problem with $\mathcal{Y} = \{-1, 1\}$. By sampling from $\mathcal{D}$ independently, a training set $(\mathbf{x}, \mathbf{y}) = (x_i, y_i)_{i=1}^n$ and a test set $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = (\hat{x}_j, \hat{y}_j)_{j=1}^m$ are obtained, but the labeling $\hat{\mathbf{y}}$ of the test set is unknown. The problem is to train a classifier and maximize the precision at k on the test set. The hypothesis space is $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ (functions mapping from $\mathcal{X}$ to $\mathcal{Y}$). The hypothesis $h \in \mathcal{H}$ is evaluated by the measure precision@k.

When we seek the best classifier from $\mathcal{H}$ for selecting k instances from the test set $\hat{\mathbf{x}}$, we only consider classifiers satisfying the k-constraint, that is, these classifiers must be in the hypothesis space $\mathcal{H}_k(\hat{\mathbf{x}}) = \{h \in \mathcal{H} | \sum_{j=1}^m \mathcal{I}[h(\hat{x}_j) = 1] = k\}$, where $\mathcal{I}[\cdot]$ is 1 if its argument is true and 0 otherwise. All classifiers not predicting k positives on the test set are excluded from $\mathcal{H}_k$. Note that any two-step method essentially reaches a classifier in $\mathcal{H}_k(\hat{\mathbf{x}})$ by setting a threshold in the second step to select k test instances. With these methods, the model optimized at the first step may be optimal in the original task, however, the classifier obtained by thresholding is often not optimal within $\mathcal{H}_k$.

To maximize the precision of $h \in \mathcal{H}_k(\hat{\mathbf{x}})$ on the test set, we essentially need to maximize the classification accuracy of h. This can be seen by the following relation. Let m_- be the number of negative test instances, and let m_{tp}, m_{fp} and m_{tn} denote the number of true positives, false positives, and true negatives (respectively) on the test set as determined by h. Then the precision@k of h can be expressed as

$$\begin{aligned} \rho(h) &= \frac{1}{k} m_{\mathrm{tp}} = \frac{1}{k}(m_{\mathrm{tn}} + k - m_-) \\ &= \frac{1}{2k}(m_{\mathrm{tp}} + m_{\mathrm{tn}} + k - m_-). \end{aligned} \quad (1)$$

Since the number of negative test instances m_- is unknown but fixed, there is a deterministic relationship between the accuracy $(m_{\mathrm{tp}} + m_{\mathrm{tn}})/m$ and the precision@k on the test set. Hence, increasing classification accuracy directly increases the precision. This motivates us to maximize the accuracy of the classifier on the test set while respecting the k-constraint.

In this section, we develop a learning algorithm for linear classifiers and thus $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathcal{Y}, h(x; w, b) = \mathrm{sign}(w^\top x + b)\}$. Our learning objective is to minimize the (regularized) hinge loss on the training set, which is a convex upper bound of the zero-one loss. Together with the k-constraint, the optimization problem is

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n [1 - y_i\,(w^\top x_i + b)]_+, \quad (2)$$

$$s.t. \qquad \sum_{j=1}^m \mathcal{I}[w^\top \hat{x}_j + b > 0] = k\,,$$

where $[\cdot]_+ = \max(\cdot, 0)$ calculates the hinge loss on each instance. Due to the piece-wise constant function in the constraint, the problem is very hard to solve.

Let us relax the equality constraint to an inequality constraint. The optimization problem becomes

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^n [1 - y_i\,(w^\top x_i + b)]_+, \quad (3)$$

$$s.t. \qquad \sum_{j=1}^m \mathcal{I}[w^\top \hat{x}_j + b > 0] \le k\,.$$

This relaxation generally does not change the solution to the optimization problem. If we neglect the constraint, then the solution that minimizes the objective will be an SVM. In our applications, there are typically significantly more than k positive test points, so the SVM will usually predict more than k positives. In that case, the inequality constraint will be active, and the relaxed optimization problem will give the same solution as the original problem [1].

[1] In the extreme case that many data points are nearly identical, the original problem may not have a solution while the relaxed problem always has one.

Even with the relaxed constraint, the problem is still hard, because the feasible region is non-convex. We first express the problem as a Mixed Integer Program (MIP). Let G be a large constant and η be a binary vector of length m. Then we can write the optimization problem as

$$\min_{w,b,\eta} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}[1 - y_i\,(w^\top x_i + b)]_+, \quad (4)$$

$$s.t. \quad w^\top \hat{x}_j + b \le \eta_j G,\ \ j = 1,\ldots,m$$
$$\eta_j \in \{0,1\},\ \ j = 1,\ldots,m$$
$$\sum_{j=1}^{m} \eta_j = k$$

Theorem 1 *Optimization problems* (3) *and* (4) *are equivalent.*

Proof: Since the two problems have the same objective that involves (w, b) only, we just need to show that two constraint sets of the two problems are equivalent in terms of (w, b). Suppose (w^1, b^1) is a solution of (3), then at most k instances have positive scores. Let η^1 be a binary vector with length m. Set $\eta^1_j = 1$ if instance j has positive score, and set other entries of η^1 to get $\sum_j \eta^1_j = k$. Then (w^1, b^1, η^1) is a solution of (4). If (w^2, b^2, η^2) is a solution of (4), then at most k test instances get positive scores, and then (w^2, b^2) is a solution of (3). □

For this MIP, a globally optimal solution can be found for small problem instances via the branch-and-bound method with an off-the-shelf package. We used Gurobi[5].

For large problem instances, finding the global optimum of the MIP is impractical. We propose to employ a *feasible direction* algorithm [2], which is an iterative algorithm designed for constrained optimization. In each iteration, it first finds a descending direction in the feasible direction cone and then calculates a step size to make a descending step that leads to an improved feasible solution. The feasible direction algorithm fits this problem well. Because the constraint is a polyhedral cone, a step in any direction within the feasible direction cone will generate a feasible solution provided that the step length is sufficiently small. Since our objective is convex but the constraint is highly non-convex, we want to avoid making descending steps along the constraint boundary in order to avoid local minima.

In each iteration, we first need to find a descending direction. The subgradient $(\nabla w, \nabla b)$ of the objective with respect to (w, b) is calculated as follows. Let $\xi_i = 1 - y_i\,(w^\top x_i + b)$ be the hinge loss on instance i. Then

$$\nabla w = w - C\sum_{i:\xi_i>0} y_i x_i\ , \quad \nabla b = -C\sum_{i:\xi_i>0} y_i. \quad (5)$$

We need to project the negative subgradient $(-\nabla w, -\nabla b)$ to a feasible direction to get a feasible descending direction. Let L, E, and R be the sets of test instances predicted to be positive, predicted to be exactly on the decision boundary, and predicted to be negative:

$$\begin{aligned} L &= \{j:\ w^\top \hat{x}_j + b > 0\},\\ E &= \{j:\ w^\top \hat{x}_j + b = 0\},\\ R &= \{j:\ w^\top \hat{x}_j + b < 0\}. \end{aligned}$$

With the constraint in (3), there can be at most k instances in L.

When (w, b) changes in a descending direction, no instance can move directly from L to R or from R to L without going though set E due to the continuity of its score. A feasible direction only allows (w, b) to moves no more than $k - |L|$ from E to L. Therefore, the feasible direction cone is

$$\mathcal{F} = \left\{(d_w, d_b) : \sum_{j\in E} \mathcal{I}[\hat{x}_j^\top d_w + d_b > 0] + |L| \le k\right\}. \quad (6)$$

To avoid that the descending direction moves exessive instances from E to L, we project the direction into the null space of a set $B \subseteq E$ of test instances, then the instances in B will stay in E. We also need to guarantee that no more than $k - |L|$ instances in $E \setminus B$ moves from E to L. Now the problem is how to find the set B.

We first sort the instances in E in descending order according to the value of $-\hat{x}_j^\top \nabla w - \nabla b$. Let $j' : 1 \le j' \le |E|$ re-index the instances in this order. To construct the set B, we start with $B = \emptyset$ and the initial direction $(d_w, d_b) = -(\nabla w, \nabla b)$. The starting index is $j_0 = 1$ if $|L| = k$, and $j_0 = 2$ if $|L| < k$. Then with index j' starting from j_0 and increasing, we consecutively put instance j' into B and project (d_w, d_b) into the null space of $\{(\hat{x}_{j^\circ}, 1) : j^\circ \in B\}$. We stop at $j' = j_1$ when all the remaining instances in E have negative inner product[2] with (d_w, d_b). The final projected direction is denoted by $(d^\star_w, d^\star_b)$. The direction $(d^\star_w, d^\star_b)$ has non-positive inner product with all instances with indices from $j' = j_0$ to $j' = |E|$, so these instances will not move into the set L when (w, b) moves in that direction. Only when $|L| < k$, is the first instance allowed to move from E to L. It is easy to check that the final projected direction $(d^\star_w, d^\star_b)$ is in the feasible cone $\mathcal{F}$ and that it is a descending direction. This subgradient projection algorithm is summarized in Algorithm 1.

In this design, we have the following considerations. When $|L| < k$, the instance in E that has the largest inner product with the negative subgradient is allowed to enter set L. We allow at most one instance to move from E to L to reduce the chance that (w, b) hits the boundary. In the projecting iterations, instances with large inner products are selected first to reduce the number of projections.

Once a descending direction is chosen, we perform a line search to determine the step size. We first find the maximum step size α that guarantees the feasibility of the descending step. That is, no points in R will cross the decision boundary and enter L with the step length α.

$$\alpha = \min_{j\in R\,:\,\hat{x}_j^\top d^\star_w + d_b > 0} \frac{-(\hat{x}_j^\top w + b)}{\hat{x}_j^\top d^\star_w + d_b}. \quad (7)$$

[2]By "inner product" between a direction (d_w, d_b) and an instance x, we mean $x^\top d_w + d_b$.

Algorithm 1 Find a descending feasible direction

Input: subgradient $(\nabla w, \nabla b)$, instance set $\{\hat{x}_j : j \in E\}$, size $|L|$, k
Output: descending feasible direction $(d_w^\star, d_b^\star)$
Sort instances in E in descending order according to $-\hat{x}_j^\top \nabla w - \nabla b$
Initialize $(d_w, d_b) = -(\nabla w, \nabla b)$
Initialize $B = \emptyset$
$j_0 = \min(k - |L|, 1) + 1$
for $j' = j_0$ **to** $|E|$ **do**
  if $\exists j'' :\ j' \le j'' \le |E|,\ \hat{x}_{j''}^\top d_w + d_b > 0$ **then**
    $B = B \cup \{j'\}$
    project (d_w, d_b) into the null space of $(\hat{x}_B, \mathbf{1})$
  else
    break
  end if
end for
$d_w^\star = d_w$, $d_b^\star = d_b$

Then we do a line search in $[0, 0.5\alpha]$ to find the best step length $\alpha^\star$. Note that the objective function is a convex piecewise quadratic function, so we only need to check these elbow points plus a minimum between two elbow points to find the best step length. We omit the details. The shrinkage 0.5 of α reduces the chance of (w, b) hitting the boundary.

We initialize w by training a standard linear SVM (although any linear model can be used) and then initialize b to satisfy the k-positive constraint. This gives us a pair (w, b) that is a feasible solution to (3). Then (w, b) is updated in each iteration according to $(w, b) := (w, b) + \alpha^\star(d_w^\star, d_b^\star)$ until convergence.

We set the maximum number of iterations, T, to 500; the algorithm typically requires only 200-300 iterations to converge. In each iteration, the two most expensive calculations are computing the subgradient and projecting the negative subgradient. The first calculation requires $O(nd)$ operations, and the second one takes at most $O(ud^2)$ operations, where u is the largest size of E. The overall running time is the time of training an initial model plus $O(T(nd + ud^2))$.

Though the problem is highly non-convex, the proposed projected subgradient method is very effective in practice, which is indicated by the comparison between solutions obtained by this method and optimal solutions obtained by Gurobi in the experiment section.

The AATP algorithm [3] faces an optimization problem similar to (3) and uses a different relaxation to find an appoximate solution. Here we show that their relaxation is very loose. The AATP objective is equivalent to ours, and the difference is that the constraint is posed on the training set. Their constraint is that the top q quantile of training instances must receive positive scores and all others, negative scores. The AATP authors assume that the optimal decision boundary must go though a *single* training instance, so their relaxation of the optimization problem is constrained to require *one* instance to be on the decision boundary. However, their assumption is incorrect, since the optimal solution would put *multiple* instances on the boundary. So their relaxation is very loose, and their solutions classify much more than quantile q of the instances as positive. Our analysis is verified by the experiment results, which will be shown in the experiment section.

3 Analysis

Before presenting experiments, we first argue that different values of k require us, in general, to train different models. We work with the population distribution $\mathcal{D}$ instead of with samples, and we assume linear models. Suppose the distributions of positive instances and negative instances have probability measures μ_+ and μ_- defined on $\mathcal{R}^d$. The total distribution is a mixture of the two distributions, and it has measure $\mu = \lambda\mu_+ + (1-\lambda)\mu_-$ with $\lambda \in (0, 1)$. The classifier (w, b) defines a positive region $R_{w,b} = \{x \in \mathcal{R}^d, w^\top x + b > 0\}$. Assume $\mu_+(R_{w,b})$ and $\mu_-(R_{w,b})$ are both differentiable with respect to (w, b). If we consider classifiers that classify fraction q of the instances as positive, then $\mu(R_{w,b}) = q$. The precision of the classifier will be $\lambda\mu_+(R_{w,b}) \ / \ q$. The optimal classifier is therefore

$$(w^\star,\ b^\star) \quad = \quad \arg\max_{(w,b)} \quad \lambda\mu_+(R_{w,b}) \qquad (8)$$
$$s.t. \quad \lambda\mu_+(R_{w,b}) + (1-\lambda)\mu_-(R_{w,b}) = q.$$

If we change q, we might hope that we do not need to modify $w^\star$ but instead can just change $b^\star$. However, this is unlikely to work.

Theorem 2 *If $(w^\star, b_1)$ and $(w^\star, b_2)$ are two optimal solutions for* (8) *with two different quantile values q_1 and q_2, then $\exists s_1, t_1, s_2, t_2, \in \mathbf{R}$,*

$$s_1 \frac{\partial \mu_+(R_{w^\star,b_1})}{\partial(w^\star, b_1)} = t_1 \frac{\partial \mu_-(R_{w^\star,b_1})}{\partial(w^\star, b_1)}, \qquad (9)$$
$$s_2 \frac{\partial \mu_+(R_{w^\star,b_2})}{\partial(w^\star, b_2)} = t_2 \frac{\partial \mu_-(R_{w^\star,b_2})}{\partial(w^\star, b_2)}. \qquad (10)$$

The proof follows directly from the KKT conditions. Note that (9) and (10) are two vector equations. When b_1 is changed into be b_2, the vectors of partial derivatives, $\partial\mu_+(R_{w^\star,b_1})/\partial(w^\star, b_1)$ and $\partial\mu_-(R_{w^\star,b_1})/\partial(w^\star, b_1)$ must change their directions in the same way to maintain optimality. This will only be possible for very special choices of μ_+ and μ_-. This suggests that $(w^\star, b^\star)$ should be optimized jointly to achieve each target quantile value q.

4 Experimental Tests

4.1 An illustrative synthetic dataset

We begin with a simple synthetic example to provide some intuition for how the TTK algorithm improves the SVM decision boundary, see Figure 1. The dataset consists of 40 training and 40 test instances. The training and testing sets each contain 22 positive and 18 negative instances. Our goal is to select $k = 4$ positive test instances. The bold line is the decision boundary of the SVM. It is an optimal linear classifier both for overall accuracy and for precision@k for $k = 24$. However, when we threshold the SVM score to select 4 test instances, this translates the decision boundary to the dashed

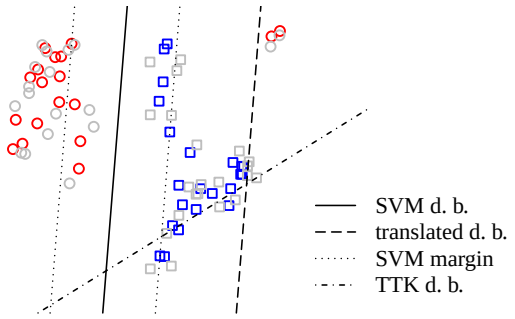

Figure 1: TTK improves the SVM decision boundary ("d. b."). Square/circle: positive/negative instance, colored/gray: training/testing instance. $k = 4$.

line, which gives very poor precision of 0.5. This dashed line is the starting point of the TTK algorithm. After making feasible direction descent steps, TTK finds the solution shown by the dot-dash-dot line. The k test instances selected by this boundary are all positive. Notice that if $k = 24$, then the SVM decision boundary gives the optimal solution. This provides additional intuition for why the TTK algorithm should be rerun whenever we change the desired value of k.

4.2 Effectiveness of Optimization

One way to compare different algorithms is to see how well they optimize the training and test surrogate loss functions. We trained a standard SVM, AATP, TTK (MIP) and TTK (feasible direction) on three UCI [12] data sets: *diabetes*, *ionosphere* and *sonar*. We set k to select 5% of the test instances. For the SVM and AATP methods, we fit them to the training data and then obtain a top-k prediction by adjusting the intercept term b. We compare the regularized hinge loss on the training set and the hinge loss on the test set of each model after adjusting b, since the model with b adjusted is the true classifier used in the task. The hyper-parameter C is set to 1 for all methods. The results in Table 1 show that TTK with either solver obtains much lower losses than the competing methods. The small difference between the third (MIP) and the fourth (feasible direction) columns indicates that the feasible direction method finds near-optimal solutions.

The results also show that the AATP method does not minimize the objective very well. Due to its loose relaxation of the objective, the original AATP solution often predicts many more positives than the target quantile of 5% of the test instances. This requires to change the intercept term b to satisfy the k-constraint.

To further understand and compare the behavior of AATP and TTK, we performed a non-transductive experiment (by making the training and test sets identical). We measured the number of training instances that fall on the decision boundary and the fraction of training instances classified as positive (see Table 2). The optimal solution given by the MIP solver always puts multiple instances on the decision boundary, whereas the AATP method always puts a single instance on the boundary. The MIP always exactly achieves the desired k, whereas AATP always classifies many more than k instances as positive. This shows that the AATP assumption that the decision boundary should pass through exactly one training instance is wrong.

Table 1: Training and test loss attained by different methods. The symbol "$+b_{adj}$" indicates that the bias term is adjusted to satisfy the k-constraint

dataset	SVM $+b_{adj}$	AATP $+b_{adj}$	TTK_{MIP}	TTK_{FD}
diabetes				
train obj.	311 ± 25	265 ± 23	224 ± 7	226 ± 7
test loss	323 ± 20	273 ± 24	235 ± 6	235 ± 5
ionosphere				
train obj.	325 ± 46	474 ± 88	127 ± 4	136 ± 4
test loss	338 ± 44	488 ± 85	146 ± 5	150 ± 7
sonar				
train obj.	167 ± 52	166 ± 41	20 ± 8	30 ± 10
test loss	216 ± 22	213 ± 30	103 ± 19	105 ± 24

4.3 Precision evaluation on real-world datasets

In this subsection, we evaluate our TTK method on ten datasets. Seven datasets, {*diabetes, ionosphere, sonar, spambase, splice*} from UCI repository and {*german-numer, svmguide3*} from the LIBSVM web site [11], are widely studied binary classification datasets. The other three datasets, NY16, NY18 and NY88, are three species distribution datasets extracted from a large eBird dataset [19]; each of them has 634 instances and 38 features. The eBird dataset contains a large number of checklists of bird counts reported from birders around the world. Each checklist is associated with the position of the observation and a set of 38 features describing the habitat. We chose a subset of the data consisting of checklists of three species from New York state in June of 2012. To correct for spatial sampling bias, we formed spatial cells by imposing a grid over New York and combining all checklists reported within each grid cell. This gives 634 cells (instances). Each instance is labeled with whether a species was present or absent in the corresponding cell.

We compare the proposed TTK algorithm with 5 other algorithms [3]. The SVM algorithm [16] is the baseline. The Transductive SVM (TSVM) [6] compared here uses the UniverSVM [18] implementation, which optimizes its objective with the convex-concave procedure. SVMperf [7] can optimize multiple ranking measures and is parameterized here to optimize precision@k. Two algorithms, Accuracy At The Top (AATP) [3] and TopPush [10], are specially designed for top precision optimization. The proposed TTK objective is solved by the MIP solver and the feasible direction method (denoted by TTK_{MIP} and TTK_{FD}). Each algorithm is run 10 times on 10 random splits of each dataset. Each of these algorithms requires setting the regularization parameter C. This was done by performing five 2-fold internal cross-validation runs within each training set and selecting the value of C from the set $\{0.01, 0.1, 1, 10, 100\}$ that maximized precision on the top 5% of the (cross-validation) test points. With the chosen value of C, the algorithm was then run on the full training set (and unlabeled test set) and the precision on the top 5% was

[3] One reviewer suggests comparing our algorithm with the MPG algorithm [22]. We will provide the result of the comparison separately.

Table 2: AATP and TTK solution statistics: Number of instances on the decision boundary ("# at d.b.") and fraction of instances predicted as positive ("fraction +")

dataset, dimension, ratio of positives	AATP		TTK_{MIP}	
	# at d.b.	fraction +	# at d.b.	fraction +
diabetes, $d = 8$, $n_+/n = 0.35$	1	0.12	5	0.05
ionosphere, $d = 33$, $n_+/n = 0.64$	1	0.53	21	0.05
sonar, $d = 60$, $n_+/n = 0.47$	1	0.46	40	0.05

Table 3: Mean Precision ($\pm$ 1 standard deviation) of classifiers when 5% of testing instances are predicted as positives.

dataset	SVM	TSVM	SVMperf	TopPush	AATP	TTK_{MIP}	TTK_{FD}
diabetes	**.86±.08**	**.86±.09**	.69±.20	**.80±.10**	.68±.28	**.85±.10**	**.86±.08**
ionosphere	.76±.13	.80±.17	.82±.22	.71±.16	**1.00±.00**	**.97±.05**	.84±.15
sonar	**.96±.08**	**.98±.06**	.85±.16	.88±.13	.90±.11	**.96±.08**	**1.00±.00**
german-numer	.70±.08	**.72±.08**	**.56±.17**	.63±.12	NA.	NA.	**.71±.06**
splice	**1.00±.00**	**1.00±.00**	**1.00±.01**	**1.00±.00**	NA.	NA.	**1.00±.00**
spambase	.97±.02	.97±.02	**.98±.01**	.96±.02	NA.	NA.	**.98±.01**
svmguide3	.86±.07	.85±.07	**.91±.04**	.83±.07	NA.	NA.	.87±.06
NY16	.64±.08	.64±.09	**.65±.12**	.62±.10	.62±.08	**.68±.07**	**.70±.09**
NY18	**.44±.11**	**.45±.10**	.36±.07	**.43±.13**	**.46±.12**	**.46±.08**	**.47±.12**
NY88	**.40±.08**	.33±.12	**.37±.15**	.34±.08	.31±.09	**.40±.09**	**.42±.07**
TTK_{MIP} w/t/l	1/5/0	2/4/0	2/4/0	1/5/0	2/4/0		
TTK_{FD} w/t/l	3/7/0	4/6/0	3/6/1	7/3/0	4/1/1		

measured. The achieved precision values were then averaged across the 10 independent runs.

Table 3 shows the performance of the algorithms. For datasets with more than 1000 instances, the AATP and TTK_{MIP} algorithms do not finish within a practical amount of time, so results are not reported for these algorithms on those datasets. This is indicated in the table by "NA". The results for each pair of algorithms are compared by a paired-differences t-test at the $p < 0.05$ significance level. If one algorithm is not significantly worse than any of the other algorithms, then it is regarded as one the best and its performance is shown in bold face. Wins, ties and losses of of TTK_{MIP} and TTK_{FD} with respect to all other algorithms are reported in the last two rows of Table 3.

On each of the six small datasets, the performance of TTK_{MIP} matches or exceeds that of the other algorithms. The TTK_{FD} method does almost as well—it is among the best algorithms on 8 of the 10 datasets. It loses once to SVMperf (on svmguide3) and once to AATP (on ionosphere). None of the other methods performs as well. By comparing TTK_{FD} with SVM, we see that the performance is improved on almost all datasets, so the TTK_{FD} method can be viewed as a safe treatment of the SVM solution. As expected, the transductive SVM does not gain much advantage from the availability of the testing instances, because it seeks to optimize accuracy rather than precision@k. The TopPush algorithm is good at optimizing the precision of the very top instance. But when more positive instances are needed, the TopPush algorithm does not perform as well as TTK.

5 Summary

This paper introduced and studied the transductive precision@k problem, which is to train a model on a labeled training set and an unlabeled test set and then select a fixed number k of positive instances from the testing set. Most existing methods first train a scoring function and then adjust a threshold to select the top k test instances. We show that by learning the scoring function and the threshold together, we are able to achieve better results.

We presented the TTK method. The TTK objective is the same as the SVM objective, but TTK imposes the constraint that the learned model must select exactly k positive instances from the testing set. This constraint guarantees that the final classifier is optimized for its target task. The optimization problem is very challenging, since it involves a set selection problem. We introduced two algorithms for solving it. First, we formulated it as a mixed integer program and solved it exactly via the branch-and-bound method. Second, we designed a feasible direction algorithm that is able to scale to larger datasets but that attempts to find a good solution. We compared both TTK algorithms to several state-of-the-art methods on ten datasets. The results indicate that the performance of the TTK methods matches or exceeds all of the other algorithms on most of these datasets.

Our analysis and experimental results show that the TTK objective is a step in the right direction. However, we believe that the performance can be further improved if we can minimize a tighter (possibly non-convex) bound on the zero-one loss.

Acknowledgments

This work has been supported primarily by CyberSEES: NSF Grant NSF 1331932. This work has also been partially supported by NSFC (61333014) and the Collaborative Innovation Center of Novel Software Technology and Industrialization of Nanjing University. Thank all reviewers for their comments and suggestions.

References

[1] S. Agarwal. The infinite push: a new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *Proceedings of the SIAM International Conference on Data Mining*, 2011.

[2] M. S. Bazaraa, H. D Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley Publishing, 3rd edition, 2006.

[3] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic. Accuracy at the top. In *Advances in Neural Information Processing Systems 25*. 2012.

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[5] Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2015.

[6] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning*, 1999.

[7] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[8] P. Kar, H. Narasimhan, and P. Jain. Surrogate functions for maximizing precision at the top. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[9] Y.-F. Li, I. Tsang, J. Kwok, and Z.-H. Zhou. Scalable and convex weakly labeled svms. *Journal of Machine Learning Research*, 14, 2013.

[10] N. Li, R. Jin, and Z.-H. Zhou. Top rank optimization in linear time. In *Advances in Neural Information Processing Systems 27*. 2014.

[11] Libsvm data: Classification (binary class). `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.

[12] M. Lichman. UCI machine learning repository. `http://archive.ics.uci.edu/ml`, 2013.

[13] D. Pechyony. *Theory and Practice of Transductive Learning*. PhD thesis, Department of Computer Science, Technion-Israel Institute of Technology, 2008.

[14] A. Rakotomamonjy. Sparse support vector infinite push. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[15] C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10, 2009.

[16] B. Schölkopf and A. J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press, 2002.

[17] V. Sindhwani and S. S. Keerthi. Large scale semi-supervised linear svms. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[18] F. Sinz and M. Roffilli. Universvm, 2012. `http://mloss.org/software/view/19/`.

[19] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. ebird: a citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142, 2009.

[20] N. Usunier, D. Buffoni, and P. Gallinari. Ranking with ordered weighted pairwise classification. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

[21] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[22] H. Wang, W. Xing, K. Asif, and B.D. Ziebart. Adversarial prediction games for multivariate losses. In *Advances in Neural Information Processing Systems 28*. 2015.