2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Linear Twin SVM for Learning from Label Proportions

Bo Wang[†‡], Zhensong Chen[§] and Zhiquan Qi[*†‡],
[†]Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing, 100190, China
[‡]The Key Research Lab on Big Data Mining and Knowledge Management, Chinese Academy of Sciences
Beijing, 100190, China
[§]School of Management, University of Chinese Academy of Sciences, Beijing, 100190, China
[*]Corresponding Author: qizhiquan@ucas.ac.cn

***Abstract*—In this paper, we study the problem of learning from label proportions in which label information of data is provided in bag level. In this kind of problem, training data is grouped into various bags and only the proportions of positive instances is known. Inspired by proportion-SVM, we propose a new classification model based on twin SVM, which is also in a large-margin framework and only needs to solve two smaller problems. Avoiding making restrictive assumptions of the data, our model can learn the labels of every single instance based on group proportions information. In order to solve the non-convex problem in our new model, we propose an alternative algorithm to obtain the optimal solution efficiently. Also, we prove the effectiveness of our method in theoretical and experimental way.**

Keywords: Learning from label proportions; Twin SVM; Alternative algorithm; Large-margin

I. Introduction

Learning from label proportions (LLP) has been intensively studied in recent years. It is important for its widely applications in real life scenarios.

For example, in political voting, candidates would be desirable to focus their attention on the swing voters who will make their decision dependent on what the candidates offer. Previous elections can directly reveal the profile of those who favor regardless, that is those who voted in favor where low campaign resources were committed. Those who voted in favor where substantial resources were committed can be either swing voters or always-favorable ones. In this way, we can tell the proportion of swing voters and always-favorable ones respectively, which can help us to categorize the voters into always-favorable, swing and opposite voters.

Another typical application is healthcare. Because of privacy protection, only the proportions of diagnosed diseases of each ZIP code area are available to public. Our aim is to learn a model to predict the individual labels (disease or nondisease) based on only this kind of group-level label proportions. Also, LLP has been applied to visual attribute modeling [9,2] and event detection [6] in computer vision.

In a word, for LLP, the training instances are provided as groups, or "bags". Here, a binary learning setting is considered: for each bag, only the proportion of the positive instances is available. We expect to acquire a classifier to predict the label for every single instance. Many efforts have been devoted in dealing with this problems [7,8,1].

This paper tries to apply extensively studied twin support vector machine (TWSVM)[4,3,5,10,11], which is a typical nonparallel hyperplanes based classifier, to LLP. An inherent drawback of known LLP (MeanMap[7] and Inverse Calibration[8]) have been indicated in [1] and a large-margin model for LLP has been studied. This paper follows the idea in [1] and proposes a new $\propto$SVM model for LLP based on TWSVM.

This paper is organized as follows. In Part II, we review the original $\propto$SVM model and TWSVM model for binary classification problem. Then, in Part III, a new large-margin model for LLP based on TWSVM is proposed. We will also analyze the correctness of new model and give an algorithm to solve the model effectively.

II. Preliminaries

In this section, $\propto$SVM for learning with label proportions and twin SVM will be introduced.

A. Linear $\propto$SVM for learning with label proportions

1) Learning with label proportions: Consider a binary classification problem. The training set $\{\boldsymbol{x_i}\}_{i=1}^{N}$ is given in the form of K disjoint bags, i.e.

$$\{\boldsymbol{x_i}|i \in \mathcal{B}_k\}_{k=1}^{K}, \bigcup_{k=1}^{K} \mathcal{B}_k = \{1, 2, \cdots, N\}, \mathcal{B}_k \cap \mathcal{B}_l = \emptyset, \forall k \neq l.$$

Without knowing the exact label for each instance, the proportion of positive class of every bag is available. Denote the proportion in k-th bag p_k and the unknown real label for every instance $y_i^* \in \{1, -1\}, i = 1, 2, \cdots, N$. Then, $p_k := \frac{|\{i|i \in \mathcal{B}_k, y_i^* = 1\}|}{|\mathcal{B}_k|}, \forall k$. Our goal is to find a classifier to predict every new instance.

2) Linear $\propto$SVM: When explicitly labeling the training by $\boldsymbol{y} = (y_1, y_2, \cdots, y_N), y_i \in \{1, -1\}, \forall i$, we can obtain an estimation of the proportion of positive instances in every bag as follows:

$$\tilde{p}_k(\boldsymbol{y}) = \frac{|\{i|i \in \mathcal{B}_k, y_i^* = 1\}|}{|\mathcal{B}_k|} = \frac{\sum\limits_{i \in \mathcal{B}_k} y_i}{2|\mathcal{B}_k|} + \frac{1}{2}.$$

978-1-4673-9618-9/15 $31.00 © 2015 IEEE
DOI 10.1109/WI-IAT.2015.130

As a result, $\propto$SVM can be formulate based on large-margin principle as follows:

$$\min_{\boldsymbol{y}\in\mathcal{Y},\boldsymbol{w},b} \tfrac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N} L(y_i, \boldsymbol{w}^T\boldsymbol{x_i}+b) + C_p\sum_{k=1}^{K} L_p(\tilde{p}_k(\boldsymbol{y}), p_k). \tag{1}$$

Here, $\boldsymbol{w}$ and b are both unrestricted and $\mathcal{Y} = \{(y_1, y_2, \cdots, y_N), y_i \in \{-1, 1\}\}$.

Remark II.1. In (1), $L(\cdot)$ and $L_p(\cdot)$ are both non-negative valued functions. More specific, $L(\cdot)$ is a loss function for supervised learning problem, which is hinge loss in standard SVM, i.e. $L(y_i, \boldsymbol{w}^T\boldsymbol{x_i}+b) = max(0, 1-y_i(\boldsymbol{w}^T\boldsymbol{x_i}+b))$.

In addition, $L_p(\cdot)$ is distance function between explicit proportions and real proportions of positive instances in every single bag. Absolute loss will be a reasonable choice for $L_p(\cdot)$.

B. Linear Twin SVM

Consider typical binary classification problem with training set:

$$T = \{(\boldsymbol{x_1}, y_1), \cdots, (\boldsymbol{x_p}, y_p), (\boldsymbol{x_{p+1}}, y_{p+1}), \cdots, (\boldsymbol{x_{p+q}}, y_{p+q})\}.$$

For the linear case, Twin SVM (TWSVM) seeks a pair of nonparallel hyperplanes:

$$\boldsymbol{w_+}^T\boldsymbol{x} + b_+ = 0, \boldsymbol{w_-}^T\boldsymbol{x} + b_- = 0,$$

to classify data based on the perpendicular distance of point $\boldsymbol{x}$ from the two hyperplanes.

1) Formulation: Here, we build the primal problem for TWSVM, which contains a pair of QPPs as follows:

$$\begin{aligned}\min_{\boldsymbol{w_+},b_+,\boldsymbol{\xi_-}} \ & \tfrac{1}{2}(\boldsymbol{A_y w_+} + \boldsymbol{e_+}b_+)^T(\boldsymbol{A_y w_+} + \boldsymbol{e_+}b_+) + c_1\boldsymbol{e_-}^T\boldsymbol{\xi_-} \\ s.t. \ & -(\boldsymbol{B_y w_+} + \boldsymbol{e_-}b_+) + \boldsymbol{\xi_-} \geq \boldsymbol{e_-},\ \boldsymbol{\xi_-} \geq \boldsymbol{0},\end{aligned} \tag{2}$$

and

$$\begin{aligned}\min_{\boldsymbol{w_-},b_-,\boldsymbol{\xi_+}} \ & \tfrac{1}{2}(\boldsymbol{B_y w_-} + \boldsymbol{e_-}b_-)^T(\boldsymbol{B_y w_-} + \boldsymbol{e_-}b_-) + c_2\boldsymbol{e_+}^T\boldsymbol{\xi_+} \\ s.t. \ & (\boldsymbol{A_y w_-} + \boldsymbol{e_+}b_-) + \boldsymbol{\xi_+} \geq \boldsymbol{e_+},\ \boldsymbol{\xi_+} \geq \boldsymbol{0},\end{aligned} \tag{3}$$

where $c_i, i = 1, 2$ are the penalty parameters, e_+ and e_- are vectors of ones of appropriate dimensions, ξ_+ and ξ_- are slack vectors of appropriate dimensions.

2) Solving: Generally speaking, we solve the Lagrange dual problems for (2) and (3), which can take advantage of SVM in prediction, i.e. only support vectors will be used in building corresponding hyperplanes. Consequently, dual problems are shown as follows:

$$\begin{aligned}\max_{\boldsymbol{\alpha}} \ & \boldsymbol{e_-}^T\boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{G}\boldsymbol{H}^T\boldsymbol{H}\boldsymbol{G}^T\boldsymbol{\alpha} \\ s.t. \ & \boldsymbol{0} \leq \boldsymbol{\alpha} \leq c_1\boldsymbol{e_-},\end{aligned} \tag{4}$$

and

$$\begin{aligned}\max_{\boldsymbol{\gamma}} \ & \boldsymbol{e_+}^T\boldsymbol{\gamma} - \tfrac{1}{2}\boldsymbol{\gamma}^T\boldsymbol{H}\boldsymbol{G}^T\boldsymbol{G}\boldsymbol{H}^T\boldsymbol{\gamma} \\ s.t. \ & \boldsymbol{0} \leq \boldsymbol{\gamma} \leq c_2\boldsymbol{e_+},\end{aligned} \tag{5}$$

where $\boldsymbol{H} = [\boldsymbol{A}\ \boldsymbol{e_+}] \in \mathbb{R}^{p\times(n+1)}$ and $\boldsymbol{G} = [\boldsymbol{B}\ \boldsymbol{e_-}] \in \mathbb{R}^{q\times(n+1)}$. After solving the dual problems (4) and (5), the solutions of problems (2) and (3) can be obtained by

$$(\boldsymbol{w_+}, b_+)^T = -(\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{G}^T\boldsymbol{\alpha},$$

$$(\boldsymbol{w_-}, b_-)^T = -(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{H}^T\boldsymbol{\gamma}.$$

Then, for an unclassified point $\boldsymbol{x} \in \mathbb{R}^n$, we can predict its label according to the value of $arg\min_{k=-,+} |\boldsymbol{w_k}^T\boldsymbol{x} + b_k|$.

III. Linear TWSVM for Learning for Label Proportions

In this section, we propose linear TWSVM for learning with label proportions first. Then, an alternative algorithm for solving this model will be introduced.

A. Linear TWSVM for Learning with Label Proportions

We write the problems of linear TWSVM as follows.

$$\begin{aligned}\min_{\boldsymbol{y},\boldsymbol{w_+},b_+,\boldsymbol{\xi_-}} \ & \tfrac{1}{2}\|\boldsymbol{A_y w_+} + \boldsymbol{e_+}b_+\|^2 + c_1\boldsymbol{e_-}^T\boldsymbol{\xi_-} + c_p\boldsymbol{e}^T\boldsymbol{L_p(y)} \\ s.t. \ & -(\boldsymbol{B_y w_+} + \boldsymbol{e_-}b_+) + \boldsymbol{\xi_-} \geq \boldsymbol{e_-},\ \boldsymbol{\xi_-} \geq \boldsymbol{0}.\end{aligned} \tag{6}$$

and

$$\begin{aligned}\min_{\boldsymbol{y},\boldsymbol{w_-},b_-,\boldsymbol{\xi_+}} \ & \tfrac{1}{2}\|\boldsymbol{B_y w_-} + \boldsymbol{e_-}b_-\|^2 + c_2\boldsymbol{e_+}^T\boldsymbol{\xi_+} + c_p\boldsymbol{e}^T\boldsymbol{L_p(y)} \\ s.t. \ & (\boldsymbol{A_y w_-} + \boldsymbol{e_+}b_-) + \boldsymbol{\xi_+} \geq \boldsymbol{e_+},\ \boldsymbol{\xi_+} \geq \boldsymbol{0}.\end{aligned} \tag{7}$$

Here, $\boldsymbol{A_y}$ and $\boldsymbol{B_y}$ denote positive and negative data matrices corresponding to the selection of $\boldsymbol{y} \in \mathcal{Y}$, where $\mathcal{Y} = \{\boldsymbol{y}|y_i \in \{-1, +1\}, i = 1, 2, \cdots N\}$, $|\mathcal{Y}| = 2^N$. In addition, $\boldsymbol{e_+}, \boldsymbol{e_-}$ and $\boldsymbol{e}$ are vectors of ones of appropriate dimension, $\boldsymbol{\xi_+}$ and $\boldsymbol{\xi_-}$ are slack vectors of appropriate dimension.

B. Alternative Algorithm

We use alternating optimization method to solve our model.

- For a fixed $\boldsymbol{y}$, based on linear TWSVM, solve (6) and (7) w.r.t $\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-$.
- When $\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-$ are obtained, fix them and solve (6) and (7) w.r.t $\boldsymbol{y}$.

Remark III.1. For the first step of the above method, when $\boldsymbol{y}$ is fixed, the model degenerates to a standard linear TWSVM, which contains two QPPs. Instead of solving the primal problems, we can solve Lagrange dual problems (4) and (5) to find the optimal solutions for $\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-$.

Remark III.2. The second step of the above method is time consuming. However, there are some strategies to make it tractable. Firstly, when fixing $\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-$, solving (6) and (7) can be distributed, because the bags are disjoint to each other. Secondly, the optimal solution of the second step can be achieved by sorting of the influence of flipping of elements in $\boldsymbol{y}$. We will fulfil these strategies in the following part.

C. Solving Second Step

In this part, we assume that $\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-$ are fixed. Then, for the second step, we can solve the following twin optimization problems parallel for every single bag.

$$\begin{aligned}\min_{\{y_i|i\in\mathcal{B}_k\},\boldsymbol{\xi_-^k}} \ & \tfrac{1}{2}\|\boldsymbol{A_y^k w_+} + \boldsymbol{e_+^k}b_+\|^2 + c_1(\boldsymbol{e_-^k})^T\boldsymbol{\xi_-^k} + c_pL_p^k(\boldsymbol{y}) \\ s.t. \ & -(\boldsymbol{B_y^k w_+} + \boldsymbol{e_-^k}b_+) + \boldsymbol{\xi_-^k} \geq \boldsymbol{e_-^k},\ \boldsymbol{\xi_-^k} \geq \boldsymbol{0}.\end{aligned} \tag{8}$$

and

$$\min_{\{y_i|i\in\mathcal{B}_k\},\boldsymbol{\xi_+^k}} \frac{1}{2}\|\boldsymbol{B_y^k w_-} + \boldsymbol{e_-^k} b_-\|^2 + c_2 \boldsymbol{e_+^k}^T \boldsymbol{\xi_+^k} + c_p L_p^k(\boldsymbol{y})$$
$$s.t. (\boldsymbol{A_y^k w_-} + \boldsymbol{e_+^k} b_-) + \boldsymbol{\xi_+^k} \geq \boldsymbol{e_+^k},\ \boldsymbol{\xi_+^k} \geq \mathbf{0}. \tag{9}$$

For each bag $\mathcal{B}_k$, the solutions of (8) and (9) offer us the labels for training points. To find the solutions of (8) and (9) efficiently, we propose the following strategy.

- Initialize $y_i = -1, i \in \mathcal{B}_k$.
- For fixed c_1 and c_2, suppose the reduction of the first two terms in (8) and (9) are $\delta_i^{(a)}$ and $\delta_i^{(b)}$ respectively.
- Sort $\delta_i^{(a)} + \delta_i^{(b)}, i \in \mathcal{B}_k$ and flip the signs of the top-R y_i's, which have the highest reductions.
- When c_1 and c_2 are fixed, we only need to sort the reductions once. Then, we can incrementally flip the sign and pick the solution with the smallest objective value, which can be seen as the optimal solutions of (8) and (9).

Proposition III.1. *For a fixed proportion θ, the strategy above is reasonable and guarantee the optimal solution for (8) and (9).*

Proof: Note that, the influence of $y_i, i \in \mathcal{B}_k$ to the first two terms of the objective function of (8) and (9) is independent. In this case, the procedure of flipping of $y_i, i \in \mathcal{B}_k$ independently is reasonable.

For (8), we want to solve the following system:

$$\min_{\mathcal{B}_k^+} \sum_{i\in\mathcal{B}_k^+} \delta_i^{(a)} \quad s.t. |\mathcal{B}_k^+| = \theta|\mathcal{B}_k|. \tag{10}$$

The same as above, for (9), we want to solve the following system:

$$\min_{\mathcal{B}_k^+} \sum_{i\in\mathcal{B}_k^+} \delta_i^{(b)} \quad s.t. |\mathcal{B}_k^+| = \theta|\mathcal{B}_k|. \tag{11}$$

It is easy to prove that when $\delta_i^{(a)}, i \in \mathcal{B}_k$ and $\delta_i^{(b)}, i \in \mathcal{B}_k$ are sorted separately, we can obtain the optimal flipping results satisfying (10) and (11) respectively.

However, in our problem, for TWSVM formulation, we should consider both (10) and (11) at the same time, which leads to a combination of $\delta_i^{(a)}$ and $\delta_i^{(b)}$, i.e. $\delta_i^{(a)} + \delta_i^{(b)}, i \in \mathcal{B}_k$. We sort the combinations to select the collection of indices of the top $\theta|\mathcal{B}_k|$ instances as $\mathcal{B}_k^+$, the rest indices constituting the set of $\mathcal{B}_k^-$.

To see the justifiability of sort of addition combination $\delta_i^{(a)} + \delta_i^{(b)}, i \in \mathcal{B}_k$, we only need to notice that for $i \in \mathcal{B}_k^+$ and $\forall j \in \mathcal{B}_k^-$, there is at least one following two conditions holds, $\delta_i^{(a)} > \delta_j^{(a)}$ and $\delta_i^{(b)} > \delta_j^{(b)}$. In other words, $\mathcal{B}_k^+$ is optimal with respect to (8) and (9) simultaneously. ∎

Remark III.3. Because of the complexity of sort in every single bag is $\mathcal{O}(|\mathcal{B}_k|\log(\mathcal{B}_k))$, the overall complexity in solving the second step in alternative algorithm is $\mathcal{O}(N\log(J))$. Here, N is the amount of instances, $J = \max\{|\mathcal{B}_k|, k = 1, 2, ..., K\}$.

Remark III.4. Because of the objective function is lower bounded and our alternating between solving $(\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-)$ and $\boldsymbol{y}$ is non-increasing. But we may get a local solution because of the nonconvexity of problems (6) and (7). As a result, we increase c_1 and c_2 gradually to alleviate this kind of nonconvexity.

Algorithm 1 Twin alter-$\propto$SVM

Input: Randomly initialize $y_i \in \{-1, 1\}, i = 1, 2, \cdots, N$.
$c_1 = 10^{-5} p_k C, c_2 = 10^{-5}(1 - p_k)C$.
while $(c_1 < C)\&\&(c_2 < C)$ **do**
  $c_1 = min\{(1 + \Delta)c_1, C\}$;
  $c_2 = min\{(1 + \Delta)c_2, C\}$;
  repeat
    Fix $\boldsymbol{y}$ to solve $(\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-)$;
    Fix $(\boldsymbol{w_+}, b_+, \boldsymbol{w_-}, b_-)$ to solve $\boldsymbol{y}$ (using the strategy discussed in III-C);
  until
  The decrease of the objective function is smaller than a threshold (10^{-4});
end while

Remark III.5. In Algorithm 1, we set $\Delta = 0.5$ to achieve the "smoothing" step to avoid the local optimal solution for (8) and (9).

Remark III.6. Here, we consider $L_p^k(\boldsymbol{y}) = |\tilde{p}_k(\boldsymbol{y}) - p_k|$, which is the absolute loss function.

IV. Experiment

In this section, an experiment result will be shown to compare our method with some known methods in solving LLP. Here, we tune the parameters $C \in \{0.1, 1, 10\}$ and $c_p \in \{1, 10, 100\}$. The UCI data set "Heart" is used to verify the effectiveness of our method. 10-fold cross-validation is applied. Here, we vary bag size in $\{4, 8, 16\}$.

TABLE I. Twin alter-$\propto$SVM

Method	4	8	16
MeanMap	80.39±0.47	79.63±0.83	79.46±1.46
InvCal	80.98±1.35	79.45±3.07	76.94±3.26
alter-$\propto$SVM	81.80±1.25	79.91±2.11	79.69±0.64
Twin alter-$\propto$SVM	**87.04±1.14**	**83.33±1.79**	**83.33±1.24**

In Table I, a superior result can be found comparing with known LLP methods. In particular, Twin alter-$\propto$SVM surpasses alter-$\propto$SVM stably.

V. Conclusion

In this paper, we study the problem of learning from label proportions in which label information of data is provided in bag level. Inspired by proportion-SVM, we propose a new classification model based on twin SVM, which is also in a large-margin framework and only needs to solve two smaller problems. Avoiding making restrictive assumptions of the data, our model can learn the labels of every single instance based on group proportions information. In order to solve the non-convex problem in our new model, we propose an alternative algorithm to obtain the optimal solution efficiently. Also, we prove the effectiveness of our method in this paper. Experimental result also show this superiority of

our method. In the future work, we will consider applying different SVM model combining with Empirical Proportion Risk Minimization (EPRM) and more efficient algorithms to solve these problems in large scale data sets.

ACKNOWLEDGMENT

This work has been partially supported by the following Grants: Key Project (No. 71331005) and Major International Joint Research Project (No. 71110107026) and Grants (No.61472390, No.11271361 and No. 61402429) from the National Natural Science Foundation of China.

REFERENCES

[1] F.X. Yu, D. Liu, Sanjiv K., T. Jebara, and S.-F. Chang, $\propto$*SVM for learning with label proportions*. In Proceedings of the 30th International Conference on Machine Learning, 2013.

[2] F.X. Yu, L. Cao, M. Merler, T. Chen, J.R. Smith, and S.-F. Chang, *Modeling attributes from category-attribute proportions*. In ACM Multimedia, 2014.

[3] R. Khemchandani, R.K. Jayadeva, and S. Chandra, *Optimal kernel selection in twin support vector machines*. Optim Lett, 3: 77C88, 2009.

[4] M.A. Kumar and M. Gopal, *Application of smoothing technique on twin support vector machines*. Pattern Recognit Lett, 29: 1842C1848, 2008.

[5] M.A. Kumar and M. Gopal. *Least squares twin support vector machines for pattern classification*. Expert Syst Appl, 36: 7535C7543, 2009.

[6] K.-T. Lai, F. X. Yu, M.-S. Chen, and S.-F. Chang, *Video event detection by inferring temporal instance labels*. In Computer Vision and Pattern Recognition, 2014.

[7] N. Quadrianto, A.J. Smola, T.S. Caetano, and Q.V. Le, *Estimating labels from label proportions*. The Journal of Machine Learning Research, 10: 2349-2374, 2009.

[8] S. Rüeping, *SVM classifier estimation from group probabilities*. In Proceedings of the 27th International Conference on Machine Learning, 2010.

[9] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, *Object-based visual sentiment concept analysis and application*. In ACM Multimedia, 2014.

[10] Y.J. Tian, Y. Shi, X.H. Liu, *Recent advances on support vector machines research*. Tech Econ Develop Econ, 18: 5-33, 2012.

[11] Z.Q. Qi, Y.J. Tian, Y. Shi. *Robust twin support vector machine for pattern classification*. Pattern Recognition, 46: 305-316, 2013.