Kernel Extreme Learning Machine for Learning from Label Proportions

Hao Yuan[1,4,5], Bo Wang[3], and Lingfeng Niu[2,4,5,*]

[1] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China
[2] School of Economics and Management, University of Chinese Academy of Sciences, Beijing, 100190, China
[3] School of Information Technology and Management, University of International Business and Economics, Beijing 100029, China
[4] Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing 100190, China
[5] Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China
yuanhao15@mails.ucas.ac.cn
wangbo@uibe.edu.cn
niulf@ucas.ac.cn

Abstract. As far as we know, Inverse Extreme Learning Machine (IELM) is the first work extending ELM to LLP problem. Due to basing on extreme learning machine (ELM), it obtains the fast speed and achieves competitive classification accuracy compared with the existing LLP methods. Kernel extreme learning machine (KELM) generalizes basic ELM to the kernel-based framework. It not only solves the problem that the node number of the hidden layer in basic ELM depends on manual setting, but also presents better generalization ability and stability than basic ELM. However, there is no research based on KELM for LLP. In this paper, we apply KELM and design the novel method LLP-KELM for LLP. The classification accuracy is greatly improved compared with IELM. Lots of numerical experiments manifest the advantages of our novel method.

Keywords: Learning from label proportions · Extreme learning machine · Kernel · Classifier calibration.

1 Introduction

In the age of big data, there are a huge number of varied data, but manually labeling these data is very difficult and expensive[10, 35, 34, 19]. In order to solve the situation, many machine learning techniques called weak-label learning are proposed. They don't require the complete labeling information and can achieve good generalization performance. There are many specific techniques of weak-label learning, such as semi-supervised learning (SSL)[3, 17, 33] , learning from partial labels [16, 5, 31], multi-instance learning (MIL)[1, 2, 7, 21, 32] and learning from label proportions (LLP) [18, 4, 22, 25–30, 8, 23, 6]. In this paper, the problems of LLP are concerned on and investigated.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

In LLP, the training instances are divided into bags and there are no labels for every instance. The only known information about labels is the label proportion for every bag. The goal of LLP is to get a instance-level classifier to give the predictions of the class labels for the new instances. An intuitive instruction of LLP problem is shown in Figure 1.

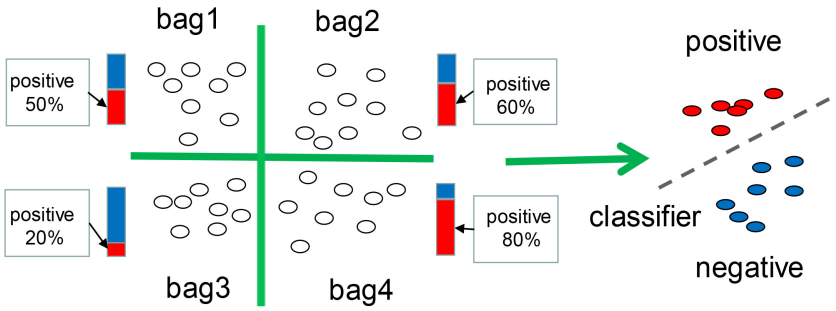

Fig. 1. Illustration of LLP. Consider the binary classification problem, each instance has the positive or negative label. Now there are 4 bags for training: bag1, bag2, bag3, bag4. The proportions of positive instances in each bag are $50\%, 60\%, 20\%, 80\%$, respectively. By using the 4 bags with the proportions, we get a instance classifier.

LLP attracts a lot of attention and has many applications, such as privacy protection, spam filter, computer vision and medical research. Let's take medical research as a detailed example. Of course, it is also an application of LLP for privacy protection. In medical research, we want to study the outbreak pattern of a new type of flu. Whether each patient is infected with this new type of flu virus is a private information between the patient and the doctor. However, the statistics can be obtained on the proportion of patients diagnosed with this novel flu who went to hospital for treatment. Some information about the basic physical condition of patients is also available. Based on these, LLP methods can be used to predict whether each patient is infected with the new flu virus. According to the prediction of LLP methods, the medical researchers can explore the specific relationship between this new type of flu and individual physical condition. After establishing the corresponding relationship, they can develop corresponding measures to better treat this disease and prevent large-scale infections.

For the sake of addressing the problems of LLP, many algorithms have been proposed[18, 4, 22, 25–30, 8, 23, 6]. Recently, Cui et al. [6] presented an approach based on extreme learning machine (ELM)[14, 15, 13, 12] called inverse extreme learning machine (IELM).Compared with the existing methods, it speeds up the training process and achieves competitive classification results. However, the LLP methods based on ELM have not been well studied. In this paper, we design a new LLP method LLP-KELM, which links inverse classifier calibration [24, 27, 6] to kernel extreme learning machine (KELM) [13]. It overcomes the

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

disadvantage that the node number of hidden layer need to be manually set. Moreover, the performance has been significantly improved than IELM.

2 Related Work

2.1 Classifier Calibration Methods

On a dataset, an probability distribution $P(X, Y)$ is given to describe the generation process of data instances. where Y and X denotes the label set and the sample space, respectively. Without loss of generality, Y is set to a binary set $\{+1, -1\}$. Generally speaking, $P(X, Y)$ is an oracle and we don't known it. For a general classification task, we usually obtain the classification result by using the sign value of numerical decision function, i.e.,

$$class(x) = sign(f_{Num}(x)).$$

In order to produce a probabilistic prediction , we can use a probabilistic classifier f_{Prob}. It estimates the class probability conditioned on the given sample x, i.e.

$$f_{Prob}(x) \approx P(Y = 1|x).$$

When we want to get the probability output and expect to improve the performance of numerical classifier, calibrating the numerical classifier is a standard approach. Therefore, we need to find a good appropriate function to scale the numerical decision values into the probability values:

$$\sigma(f_{Num}(x)) \approx P(Y = 1|x).$$

Platt calibration [24] is one of probabilistic calibration techniques. It has been validated that this method is quite efficient for many numerical decision functions. It can transform decision outputs to posterior probabilities by the equation:

$$\sigma_{Platt}(f(x)) = \frac{1}{\exp(B - Af(x)) + 1}.$$

In above equation, the parameters B and A can be solved by maximum likelihood estimation.

In LLP, "Inverse Calibration" [27] is the first method to adopt the process of inverting calibration for a classifier. In this paper, the idea also will be used.

2.2 Extreme Learning Machine

The work[14] proposed a single-hidden-layer feed-forward networks (SLFNs) learning system called ELM. ELM unifies the classification and regression in the same framework. ELM runs extremely fast and can be easily implemented.

We briefly describe the special form of ELM models, which only have one output node as follows: Given N training instances $(\boldsymbol{x}_i, y_i)_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^n$

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

denotes the feature vector, $y_i \in \mathbb{R}$ denotes the corresponding target value. Suppose that the SLFNs with M hidden nodes have activation function $z(\boldsymbol{x})$, then the model of SLFNs can be represented as

$$\sum_{j=1}^{M} \beta_j z(\boldsymbol{w}_j.\boldsymbol{x}_i + b_j) = o_i, \quad i = 1, 2, .., N. \tag{1}$$

Here, $\boldsymbol{w}_j \in \mathbb{R}^n$ and $\beta_j \in \mathbb{R}$ denote the input and output weight, respectively. In order to make the output o_i be as close as possible to the target y_i, the loss function of SLFNs is build as

$$\min_{\{\beta_j, \boldsymbol{w}_j, b_j\}_{j=1}^{M}} \sum_{i=1}^{N} |o_i - y_i|^2. \tag{2}$$

(2) can be transformed compactly as:

$$\min_{\boldsymbol{\beta}, \{\boldsymbol{w}_j, b_j\}_{j=1}^{M}} \|\mathbf{Q}\boldsymbol{\beta} - Y\|_2^2. \tag{3}$$

Here, $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_M]^T$, $Y = [y_1, y_2, ..., y_N]^T$, Q ($q(\boldsymbol{x})$ can be regard as the feature mapping) can be expressed as:

$$\begin{aligned} \mathbf{Q} &= \begin{bmatrix} \boldsymbol{q}(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{q}(\boldsymbol{x}_N) \end{bmatrix} = \begin{bmatrix} q_1(\boldsymbol{x}_1) & \dots & q_M(\boldsymbol{x}_1) \\ \vdots & \vdots & \vdots \\ q_1(\boldsymbol{x}_N) & \dots & q_M(\boldsymbol{x}_N) \end{bmatrix} \\ &= \begin{bmatrix} z(\boldsymbol{w}_1\boldsymbol{x}_1 + b_1) & \dots & z(\boldsymbol{w}_M\boldsymbol{x}_1 + b_M) \\ \vdots & \vdots & \vdots \\ z(\boldsymbol{w}_1\boldsymbol{x}_N + b_1) & \dots & z(\boldsymbol{w}_M\boldsymbol{x}_N + b_M) \end{bmatrix}_{N \times M} \end{aligned} \tag{4}$$

In the training phase of ELM, $\{\boldsymbol{w}_j, b_j\}_{j=1}^{L}$ are randomly produced and don't need to be learned. So, the tuned parameters $\boldsymbol{\beta}$ of the learning system ELM can be solved by means of the least squares methods. The solution $\boldsymbol{\beta}^\star$ is

$$\mathbf{Q}^\dagger Y, \tag{5}$$

where the notation † operates the Moore-Penrose generalized inverse of a matrix. Finally, ELM is represented as

$$f(\boldsymbol{x}) = \boldsymbol{q}(\boldsymbol{x})\boldsymbol{\beta}^\star. \tag{6}$$

According to the theory of matrix computation [9], when $\mathbf{Q}$ is full row-rank,

$$\boldsymbol{\beta}^\star = \mathbf{Q}^T(\mathbf{Q}\mathbf{Q}^T)^{-1}Y, \tag{7}$$

when $\mathbf{Q}$ is full column-rank,

$$\boldsymbol{\beta}^\star = (\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T Y. \tag{8}$$

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

3 KELM for LLP

A binary classification situation is considered as follows. The training instances $\{x_i, y_i^\star\}_{i=1}^N$ are expressed as the form of K bags: $\{B_k, P_k\}_{k=1}^K$, where B_k represents the k-th bag including N_k instances$\{x_i, y_i^\star\}_{i=1}^{N_k}$.

We denote $y_i^\star \in \{+1, -1\}$ the unknown ground truth label of each instance x_i. Then, for the k-th bag B_k, the proportions of positive instances (i.e. the conditional probability)can be calculated by

$$P_k = \frac{|\{i|x_i \in B_k, y_i^\star = +1\}|}{|B_k|}, k = 1, 2 \dots K. \tag{9}$$

If the instance labels are modeled as $\{y_i\}_{i=1}^N$, for the k-th bag, the modeled label proportion can be expressed as

$$p_k = \frac{|\{i|x_i \in B_k, y_i = +1\}|}{|B_k|}, k = 1, 2 \dots K. \tag{10}$$

Here, we can treat p_k as the estimate value of P_k.

Now, LLP problem formulation has been completed and we know the bags $\{B_k\}_{k=1}^K$ and the corresponding proportions $\{p_k\}_{k=1}^K$. Next, We will inverse the process of classifier calibration. Firstly, each bag B_k is regarded as an instance X_k, which is called the "super-instance". The super-instance X_k is presented as the mean value of all instances in B_k , i.e., $X_k = (\sum_{x_i \in B_k} x_i)/|B_k|$. Secondly, a soft label $\sigma^{-1}(p_k)$ are generated. The generation process is described as follows: (1)We fix the scaling function in classifier calibration methods. Here, we use the scaling function σ_{Platt} in the Platt calibration and let the parameter $A = 1, B = 0$. (2) We calculate the inverse of the scaling function $\sigma_{Platt}^{-1}(p) = -log(1/p - 1)$ and get the soft label $y_k^s = \sigma_{Platt}^{-1}(p_k)$ of each super-instance X_k.

After obtaining the super-instance X_k and the soft label y_k^s , the LLP problem is converted to a supervised learning problem, i.e., a regression problem. We expect the regression model f can fits y_k^s well over each super-instance X_k. In this paper, KELM is adopted to better solve the regression problem. In KELM, Mercer's conditions are applied and the ELM kernel function is defined as: $\kappa(x_i, x_j) = \langle \boldsymbol{q}(x_i), \boldsymbol{q}(x_j) \rangle = K_{i,j}$, where $\langle , \rangle$ represents the inner product operation, q is the feature mapping function in formula (4) . Here, lots of kernel function can be used, such as polynomial kernel, RBF kernel and so on. In equation (7), a positive number C can be added to all the diagonal elements of $\mathbf{Q}\mathbf{Q}^T$ motivated by the ridge regression theory[11], and $\mathbf{Q}\mathbf{Q}^T$ can be replaced by the kernel matrix $\mathbf{K}$. When a new instance $\boldsymbol{x}$ comes, we compute the kernel matrix $\boldsymbol{k_x}$ between $\boldsymbol{x}$ and $\{\boldsymbol{X}_k\}_{k=1}^K$ and then get the corresponding response $y_{\boldsymbol{x}}$ by formula (11), the label y of $\boldsymbol{x}$ can be obtained by the sign of $y_{\boldsymbol{x}}$. The novel

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

model LLP-KELM can be formalized as:

$$\begin{aligned} f(\boldsymbol{x}) =& \boldsymbol{q}(\boldsymbol{x})\mathbf{Q}^T(\frac{\mathbf{I}}{C}+\mathbf{Q}\mathbf{Q}^T)^{-1}Y \\ =& \boldsymbol{k}_{\boldsymbol{x}}(\frac{\mathbf{I}}{C}+\mathbf{Q}\mathbf{Q}^T)^{-1}Y \\ =& \begin{bmatrix} \kappa(\boldsymbol{x},X_1) \\ \vdots \\ k(\boldsymbol{x},X_N) \end{bmatrix}^T (\frac{\mathbf{I}}{C}+\mathbf{K})^{-1}Y, \end{aligned} \tag{11}$$

Here, $\mathbf{K}$ represents the kernel matrix of the super-instances $\{X_k\}_{k=1}^K$, Y is the vector of the soft labels $\{y_k^s\}_{k=1}^K$. We summarize the process of LLP-KELM as Algorithm 1.

Algorithm 1 LLP-KELM

Require: A training set in bags $\{(B_k, p_k)\}_{k=1}^K$, the kernel function κ, the parameter C and a test instance $\boldsymbol{x}$

Ensure: The predicted label y of the instance $\boldsymbol{x}$

1: Compute the bag-level super-instances $\{X_k\}_{k=1}^K$
2: compute the kernel matrix $\mathbf{K}$ of $\{X_k\}_{k=1}^K$by the kernel function κ
3: compute $Y = [y_1^s, ..., y_K^s]^T$ by the inverse function of σ: $\sigma^{-1}(p_k), k = 1, 2, .., K$.
4: compute the inverse of $\frac{\mathbf{I}}{C}+\mathbf{K}$
5: compute $\boldsymbol{\beta}^\star = (\frac{\mathbf{I}}{C}+\mathbf{K})^{-1}Y$
6: compute the kernel matrix $\boldsymbol{k}_{\boldsymbol{x}}$ between $\boldsymbol{x}$ and the super-instances $\{X_k\}_{k=1}^K$
7: compute the responding $y_{\boldsymbol{x}} = \boldsymbol{k}_{\boldsymbol{x}}\boldsymbol{\beta}^\star$
8: get the predicted label $y = sign(y_{\boldsymbol{x}})$
9: **return** the predicted label y

4 Numerical Experiment

We conduct the experiment to verify the performances of our novel method LLP-KELM in this section. From the paper [6], we know that IELM can produce the comparable classification accuracy and has very fast training speed compared with other advanced methods in many public datasets. Therefore, it is appropriate that we choose IELM as the only baseline. We run the experiment code on a server with Windows Server OS. Its configurations are Intel(R) Xeon(R) CPU E5-2640 at 2.6GHz and 128GB main memory. MATLAB R2013a 64-bit version is used as the programming IDE.

4.1 Benchmark Datasets

Different algorithms are compared in real-world classification datasets obtained from the UCI repository [20] and LibSVM collection [1]. We only consider the

[1] https://www.csie.ntu.edu.tw/c̃jlin/libsvmtools/datasets/

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

binary classification datasets. For each dataset, the features are scaled. The statistics for these used datasets are shown in Table 1.

Table 1. Statistics of the used data. According to the sample number, data are list one by one.

Datasets	# of Samples	# of Features
spect	267	22
heart	270	13
liver-disorders	345	5
vote	435	16
credit-a	690	15
diabetes	768	8
fourclass	862	2
splice	1000	60
german.numer	1000	24
a1a	1605	119

4.2 Experimental Settings

In order to generate the data form of LLP problems, we randomly split various datasets in Table 1 into lots of bags with fixed bag size. In this paper, the bag sizes which are used are 2, 4, 8, 16, 32, 64. We utilize grid search and 5-fold cross validation to find the best parameters and evaluate the performance. The performance are evaluated based on test accuracy on the instance level.

For the baseline IELM, we follow the paper [6] and adopt the same parameter setting rule. The node number of hidden layer ranges from the set $\{5, 10, 15, 20, 25, ..., 200\}$. For our proposed method, the RBF kernel $\kappa(u, v) = \exp(-\gamma\|u-v\|_2^2)$ is considered. The logarithm of parameters, $\log_{10} C$ and $\log_{10} \gamma$, are adjusted from the set $\{-3, -2, -1, 0, 1, 2, 3\}$.

4.3 Results and Analysis

The experiment results on the various datasets are reported in Table 2. We use the bold figures to state the best accuracy of our experiments. Table 2 displays the mean test accuracies of 5-fold cross validation with standard deviation. As shown in Table 2, our novel method LLP-KELM overwhelmingly outperforms the baseline IELM on the test accuracy in most situations. We take some examples to illustrate the results. The datasets "splice", we observe that the accuracy of LLP-KELM and IELM are respectively 82.80% ,75.10% in the bag size 2. In the setting of bag size 4, the accuracy value are respectively 79.30% ,70.10%. It is obvious that LLP-KELM is much better than IELM on the test accuracy. In other bag size setting, this is also true. We can also notice that the accuracies of

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

Table 2. Mean test accuracies (mean ± std %) of 5-fold cross validation with different bag sizes: 2, 4, 8, 16, 32, 64.

Datasets	Method	Bag Size					
		2	4	8	16	32	64
spect	LLP-KELM	**80.91±3.55**	**80.13±4.00**	**81.30±3.56**	**81.66±6.69**	**80.13±4.41**	79.39±4.47
	IELM	80.17±3.46	77.13±5.15	73.81±8.31	72.61±8.34	78.25±7.57	**81.66±2.29**
heart	LLP-KELM	**82.96±4.01**	**85.93±3.84**	**81.85±2.41**	**77.04±5.65**	**74.44±5.77**	73.33±9.85
	IELM	82.22±4.46	80.37±2.81	76.30±7.90	70.74±12.92	67.04±6.60	**75.56±6.06**
liver-disorders	LLP-KELM	**69.28±7.20**	**63.48±5.27**	**62.32±5.42**	57.97±4.81	**58.84±5.29**	**58.26±5.16**
	IELM	68.41±4.40	62.03±7.20	60.00±6.28	**60.00±6.28**	53.91±8.84	52.46±11.10
vote	LLP-KELM	**96.32±1.89**	**95.40±1.41**	**94.48±0.96**	**93.56±2.38**	**91.95±1.82**	**87.82±5.05**
	IELM	95.63±0.51	92.87±3.08	89.43±4.48	91.03±4.48	**91.95±4.53**	87.36±7.22
credit-a	LLP-KELM	**86.38±3.05**	**85.36±3.05**	**85.22±4.43**	**85.22±3.22**	**83.48±5.48**	**79.71±7.19**
	IELM	85.94±1.89	81.16±5.45	81.30±2.48	79.86±2.73	74.64±5.66	77.54±4.23
diabetes	LLP-KELM	**77.73±1.30**	**75.78±1.92**	**74.87±0.96**	**74.48±2.24**	**70.32±2.18**	**69.27±3.92**
	IELM	76.43±0.58	74.35±1.28	72.39±3.37	69.66±4.99	69.80±4.48	60.78±13.32
fourclass	LLP-KELM	**81.22±6.12**	**79.01±3.61**	**77.03±4.11**	**76.22±7.66**	**75.18±3.95**	**74.02±5.76**
	IELM	78.54±4.11	75.88±4.03	67.52±4.91	62.41±12.23	59.98±8.58	55.21±9.45
splice	LLP-KELM	**82.80±1.96**	**79.30±2.14**	**74.00±5.42**	**71.00±5.33**	**61.50±3.18**	**60.50±7.95**
	IELM	75.10±3.21	70.10±4.60	66.10±4.16	63.40±2.13	61.30±2.46	60.20±6.39
german.numer	LLP-KELM	**75.00±2.62**	**74.60±2.04**	**73.30±3.93**	**70.90±1.85**	**70.10±1.64**	**71.00±3.54**
	IELM	73.70±2.36	71.60±3.07	66.60±4.38	67.00±2.85	64.50±4.43	63.40±2.58
ala	LLP-KELM	**83.12±1.68**	**82.18±1.48**	**80.12±2.98**	**78.69±2.60**	**78.38±2.61**	**76.88±2.30**
	IELM	79.31±3.17	75.26±2.58	72.40±3.97	68.72±5.56	71.21±7.83	74.58±4.90

the two methods LLP-KELM and IELM decrease with the bag size increasing in some ways. This indicates that the larger the bag size, the harder it is to correctly classify the instances in the bags. This is a great challenge in LLP.

5 Conclusion

We design a novel LLP method called LLP-KELM, which significantly improves the method IELM. In LLP-KELM, the kernel version of ELM and the inverse process of classifier calibration are fully utilized. In most situation of our experiments, it can gain the better performances than IELM. In conclusion, our novel method LLP-KELM is feasible for LLP and can be applied in many practical applications.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No.11331012, No.11671379) and UCAS Grant (No. Y55202LY00).

References

1. Amores, J.: Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence **201**(4), 81–105 (2013)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

2. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. Advances in Neural Information Processing Systems **15**(2), 561–568 (2002)
3. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. IEEE Transactions on Neural Networks **20**(3), 542–542 (2009)
4. Chen, B.C., Chen, L., Ramakrishnan, R., Musicant, D.R.: Learning from aggregate views. In: 22nd International Conference on Data Engineering (ICDE'06). pp. 3–3. IEEE (2006)
5. Cour, T., Sapp, B., Taskar, B.: Learning from partial labels. Journal of Machine Learning Research **12**(May), 1501–1536 (2011)
6. Cui, L., Zhang, J., Chen, Z., Shi, Y., Yu., P.S.: Inverse extreme learning machine for learning with label proportions. In: In Proceedings of IEEE International Conference on Big Data (2017)
7. Dietterich, T.G., Lathrop, R.H., Lozano-P, Rez, T.: Solving the multiple instance problem with axis-parallel rectangles. Elsevier Science Publishers Ltd. (1997)
8. Fan, K., Zhang, H., Yan, S., Wang, L., Zhang, W., Feng, J.: Learning a generative classifier from label proportions. Neurocomputing **139**, 47–55 (2014)
9. Golub, G.H., Van Loan, C.F.: matrix computations, 3rd (1996)
10. Hady, M.F.A., Schwenker, F.: Semi-supervised Learning. Springer Berlin Heidelberg (2013)
11. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (2000)
12. Huang, G.B., Wang, D.H., Lan, Y.: Extreme learning machines: a survey. International journal of machine learning and cybernetics **2**(2), 107–122 (2011)
13. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **42**(2), 513–529 (2012)
14. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. vol. 2, pp. 985–990. IEEE (2004)
15. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1), 489–501 (2006)
16. Jin, R., Ghahramani, Z.: Learning with multiple labels. In: Advances in neural information processing systems. pp. 921–928 (2003)
17. Joachims, T.: Transductive inference for text classification using support vector machines. In: Sixteenth International Conference on Machine Learning. pp. 200–209 (1999)
18. Kück, H., de Freitas, N.: Learning about individuals from group statistics. In: Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence. pp. 332–339. UAI'05, AUAI Press, Arlington, Virginia, United States (2005), `http://dl.acm.org/citation.cfm?id=3020336.3020378`
19. Li, Y.F., Tsang, I.W., Kwok, J.T., Zhou, Z.H.: Convex and scalable weakly labeled svms. Journal of Machine Learning Research **14**(1), 2151–2188 (2013)
20. Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
21. Maron, O., Lozano-P, Rez, T.: A framework for multiple-instance learning. In: Advances in Neural Information Processing Systems. pp. 570–576 (1997)
22. Musicant, D.R., Christensen, J.M., Olson, J.F.: Supervised learning by training on aggregate outputs. In: Seventh IEEE International Conference on Data Mining. pp. 252–261 (2007)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30

23. Patrini, G., Nock, R., Caetano, T., Rivera, P.: (almost) no label no cry. In: Advances in Neural Information Processing Systems. pp. 190–198 (2014)
24. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers **10**(4), 61–74 (1999)
25. Qi, Z., Wang, B., Meng, F., Niu, L.: Learning with label proportions via npsvm. IEEE Transactions on Cybernetics **PP**(99), 1–13 (2017)
26. Quadrianto, N., Smola, A.J., Caetano, T.S., Le, Q.V.: Estimating labels from label proportions. Journal of Machine Learning Research **10**(3), 2349–2374 (2009)
27. Rueping, S.: Svm classifier estimation from group probabilities. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 911–918 (2010)
28. Stolpe, M., Morik, K.: Learning from label proportions by optimizing cluster model selection. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 349–364. Springer (2011)
29. Yu, F.X., Choromanski, K., Kumar, S., Jebara, T., Chang, S.F.: On learning from label proportions. arXiv preprint arXiv:1402.5902 (2014)
30. Yu, F., Liu, D., Kumar, S., Jebara, T., Chang, S.: *propto*svm for learning with label proportions. In: Proceedings of the 30rd International Conference on Machine learning (2013)
31. Zhang, M.L., Yu, F., Tang, C.Z.: Disambiguation-free partial label learning. IEEE Transactions on Knowledge and Data Engineering **29**(10), 2155–2167 (2017)
32. Zhang, Q., Goldman, S.A.: Em-dd: an improved multiple-instance learning technique. In: International Conference on Neural Information Processing Systems: Natural and Synthetic. pp. 1073–1080 (2001)
33. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schlkopf, B.: Semi-supervised learning by maximizing smoothness. Journal of Machine Learning Research (2004)
34. Zhou, Z.H.: A brief introduction to weakly supervised learning. National Science Review (2017)
35. Zhu, X.: Semi-supervised learning literature survey. Computer Science **37**(1), 63–77 (2005)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_30