人工知能学会研究会資料
SIG-DMSM-A803-02 (3/3)

ラベル信頼度を利用したブースティング手法
A Boosting Method Utilizing Reliably Labeled Data

中田　康太 [1*]　櫻井　茂明 [1]
折原　良平 [1]
Kouta Nakata[1]　Shigeaki Sakurai[1]　Ryohei Orihara[2]

[1] 東芝研究開発センター　システム技術ラボラトリー
[1] System Engineering Laboratory, Corporate Research & Development Center TOSHIBA CORPORATION
[2] 東芝研究開発センター　ヒューマンセントリックラボラトリー
[2] Human Centric Laboratory, Corporate Research & Development Center TOSHIBA CORPORATION

Abstract: We address a novel and realistic Label Reliability Problem that belongs to the field of supervised learning, where confidence of labeling is different for each training set. Our main idea is to make more precise classifiers by dealing with reliably and not reliably labeled sets seperately. We focus on a novel boosting method that utilizes reliably labeled data. The theoretical investigation on the method makes clear its relation to soft margin approach, cost-sensitive learning and semi-supervised learning. We perform detailed experiments that include the boosting method and 8 related methods. The results suggest the superiority of our approach that counts on unreliable labels.

1 Introduction

In the field of supervised learning, accurate labeling is essential to accurate learning. Since such labeling requires the domain knowledge, enough experience and the fairness and consistency of judgement, it is desirable to have experts of the domain give labels to obtain accurate training data. But in reality, having such experts for all data is often impossible because of its economical and temporal costs or merely rare chance to come by capable experts. With realistic limitations, experts' labeling are often short for solving actual problems, and we instead have to make use of non-experts' labeling in order to obtain sufficient mount of training data. While lower costs and ubiquity of non-experts make it possible to obtain large training sets, the accuracy of labeling alternatively become a problem: since non-experts' labeling is not so reliable as experts', it is possible that a large amount of training data contains proportional mislabels. We frequently face a dilemma of fine/small and large/coarse training data. In genearal supervised learning, such background is not taken into account; each of data is dealt with equally as a member of single training set, no matter which of experts or non-experts are responsible for its labeling. But let us consider again that data of non-experts' labeling are realistically dominant in quantity. In that case, fine data of experts' labeling would be submerged and likely to have little influence on classifiers. On the other hand, as the ratio of non-experts labeling become large, their accompanying mislabeled data would make up some part of the training set in proportion, and they possibly makes the resultant classifier poor. It is highly likely that expensive labeling are not fully utilized. In this paper, we call this problem as the Label Reliability Problem (LRP).

When the LRP is assumed, several machine learning frameworks would be effective: soft margin algorithms of well-known Support Vector Machine [5] or boosting [14] will efficiently reduce the effect of noise data in non-experts' set. In framework of cost-sensitive learning, where boosting methods are often utilized [16], one can put higher costs on experts' labels, to make data of important and reliable sets

[*] 連絡先：東芝研究開発センター
〒 212-8582　神奈川県川崎市幸区小向東芝町１番地
E-mail: kouta.nakata@toshiba.co.jp

contribute to learning more. Somtimes it might be better to discard all non-experts' labels and use only their distribution, that is, to adopt semi-supervised approach [4].

In this paper, our basic idea is based on a novel approach of the previous work [12] which attempt to make more precise classifiers than exisiting frameworks by dealing with experts data and non-experts data seperately. We first review a boosting approach, where credits, the degrees of label confidence, are set on non-experts' data based on their neighboring reliable experts data, and the degrees are in turn reflected in learning as data with higher confidence contribute to the classifier the more (Section 2). In their work, a boosting based method is examined with artificial and publicly available data. Although the experimental results suggest that their novel method can make more precise classifier than traditional method that adopts all data equally, , there remain further investigation to be done. In this paper, we extend their previous work. We clarify how the boosting based method make more precise classifiers by investigating the relation to existing frameworks. Firstly, we focus on soft margin approach and show that introducing credits is closely related to setting slack variables without heuristic parameter setting search. We also refer to other related works of cost-sensitive learning and semi-supervised learning (Section 3). To confirm the effectiveness of the approach, we perform detailed experiments where the boosting method and other related methods are involved, and show the properties and the superiority of the approach (Section 4). Finally we discuss the further extension of the method (Section 5) and summarize our paper (Section 6).

2 Credit Sensitive Boosting

The first assumption of our approach is that experts' labeling is definitive but non-experts' labeling accompanies some mislabels. For simplicity, we call the former "expert data" and the latter "non-expert data". In this section, we briefly review Credit-sensitive AdaBoost (CAB) method that utilizes reliably labeled expert data [12]. The method mainly assumes two points. Firstly, CAB sets credits, the degrees of label confidence, on non-expert data based on the neighboring expert data. Secondly, the credits are reflected in learning such as data with the higher credits have the greater influence on a resultant classifier. Credits of non-expert data could be determined in many different ways and the previous work employs an intuitive way that exploits local information of expert and non-expert sets, that is, credits are calculated by

$$c_i = a \sum_{k=1, y_k^{\mathrm{ex}} = y_i^{\mathrm{ne}}}^{K} \frac{1}{d_k}, \tag{1}$$

where c_i is the value of a credit imposed on the non-expert data $(x_i^{\mathrm{ne}}, y_i^{\mathrm{ne}})$, $(x_k^{\mathrm{ex}}, y_k^{\mathrm{ex}})$ is the k-th nearest neighbor expert data of $(x_i^{\mathrm{ne}}, y_i^{\mathrm{ne}})$ and d_k is the Euclid distance between the instance x_k^{ex} and x_i^{ne}. Each of non-expert data refers to the K-nearest neighbors of expert data and can take credits which is inversely proportional to the distance to the expert data, but only from the expert data of same label. Thus, non-expert data with agreeing expert data around take higher credits.

The CAB method reflects the credits of non-expert data in learning by extending the algorithm of AdaBoost [7]. One advantage of utilizing AdaBoost is its high performance for various problems, but its compatibity with cost-sensitive learning seems also favorable [6]. The method of cost-sensitive learning can be applied to their 'credit-sensitive' learning as they accordingly incorporate credits of non-expert data into weights in AdaBoost algorithm. Algorithm 1 shows the learning algorithm of CAB. The extension from traditional AdaBoost is step (a), where credits are incorporated into the data weights. As many discussions in the field of cost-sensitive learning imply, how to incorporating credits is various and the best way seems open to question, and the previous work makes the simplest modification on CAB by multiplying weights by credits

$$D_t^{'}(i) = c_i D_t(i). \tag{2}$$

New weight $D_t^{'}(i)$ is incorporated in generating weak learners in the next step. Data with smaller c_i is obviously utilized less for learning however large the value of original weight $D_t(i)$ become. In the next section, we extend the previous work to make theoretical analysis on how CAB exploits the credits on non-expert data. We focus on the margin, and show the relation of CAB to other frameworks of soft margin, cost-sensitive learning and semi-supervised learning approaches.

```
Algorithm1: Credit-Sensitive AdaBoost
1.  Start with weights $D_1(i) = 1/N, i = 1, 2, ..., N$.
2.  Repeat for $t = 1, 2, ...,$:
    (a) Incorporate credits $c_i$ into weights $D_t(i)$ and obtain new weights $D^{'}_t(i)$.
    (b) Train weak learner $h_t(x)$ using weights $D^{;}_t(i)$ on the training data.
    (c) Set $\beta_t = \log \frac{\epsilon_t}{1-\epsilon_t}$, where $\epsilon_t$ is error function, $\epsilon_t = \sum_{y_i \neq h_t(x_i)} D_t(i)$.
    (d) Set $D_{t+1}(i) = D_t(i) \exp(\beta_t),\ i = 1, 2, ..., N$, when $y_i \neq h_t(x_i)$,
      and renormalize weights as $\sum_i D_t(i) = 1$.
3.  Output the final classifier: $sign(\sum_{t=1}^{T} \beta_t h_t(x))$
```

3 Properties and Related Works

3.1 Soft Margin Properties

To make clear how CAB learns better classifiers, we investigate the properties of CAB focusing on the margin. Margin is defined on each of training data (x_i, y_i) as

$$\text{margin}(x_i, y_i) = \frac{y_i \sum_t \alpha_t h_t(x_i)}{\sum_t \alpha_t}, \tag{3}$$

which represents the degree of agreement of the base classifiers. The high performance of AdaBoost is explained by its margin maximization property [15]. Schapire et al. theoretically and experimentally proved that the high smallest margins lead to low generalization error, and that AdaBoost is desigined to greedily maximize the margin of hard-to-classify data by putting higher weights on the misclassified training data. Such property also explains its sensitivity to noises. AdaBoost concentrate on data with the smallest margin even if they are indeed noise data and the resultant classifier become poor as AdaBoost adjust itself too well to noises. Therefore, AdaBoost is also known as a hard margin classifier and sometimes suffers the overfitting problem on noisy data. To avoid serious overfitting to noises, some boosting methods based on soft margin concept are proposed. In an intuitive algorithm of AdaBoost$_{Reg}$ [14], a soft margin is defined as

$$\text{mg}(x_i, y_i) = \text{margin}(x_i, y_i) + C\mu_t(x_i, y_i)^p, \tag{4}$$

where $\mu_t(x_i, y_i) = \sum_{t'=1}^{t} \alpha_{t'} D_{t'}(x_i)$ is the average weight of instance (x_i, y_i). Generally, noises are difficult to classify and the values of $\mu_t(x_i, y_i)$ tend to be large. AdaBoost$_{Reg}$ maximizes margin $\text{mg}(x_i, y_i)$, which make it possible to leave Margin(x_i, y_i) small if the value of $\mu_t(x_i, y_i)$ becomes large. The variable $\mu_t(x_i, y_i)$ works as a slack variable to actual margin and enables AdaBoost$_{Reg}$ to avoid overfitting to noises. The variables P and C control the trade-off of how much a slack variable affects learning, and their best value is generally determined by some heuristic parameter search.

CAB sets the credits on non-expert data and reflect them in learning. Ideally, noise data should be set with smaller credits and contribute less to learning. As soft margin concept, our approach of introducing credits aims at adequately reducing the effects of noises in non-expert data. In this section, we prove that incorporating credits into learning is equal to setting slack variables. From weight update rule of traditional AdaBoost and $D^{'}_t(i) = c_i D_t(i)$,

$$\begin{aligned} D^{'}_{T+1}(i) &= c_i D_{T+1}(i) \\ &= \frac{c_i D^{'}_T(i) \exp[-y_i \alpha_t h_t(x_i)]}{Z_T} \\ &= \frac{c_i^T \exp[-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)]}{\Pi_{t=1}^{T} Z_t}. \end{aligned}$$

where Z_t is the normalization factor

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}, \tag{5}$$

and ϵ_t represents the training error rate of t-th weak classifier. Solving for the exponential term and taking logarithmic of both sides lead to

$$\begin{aligned} y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) &= -\ln[c_i^{-T} D^{'}_{t+1}(i) \cdot \Pi_{t=1}^{T} Z_t] \\ &= -\sum_{t=1}^{T} \ln(c_i^{-1} \cdot Z_t) - \ln D^{'}_{t+1}(i) \\ &= \sum_{t=1}^{T} \ln \frac{c_i}{Z_t} - \ln D^{'}_{t+1}(i), \end{aligned}$$

and credit-incorporated margin is represented as

$$\begin{aligned} \text{margin}(x_i, y_i) &= \frac{y_i \sum_t \alpha_t h_t(x_i)}{\sum_t \alpha_t} \\ &= \frac{\sum_{t=1}^{T} \ln \frac{c_i}{Z_t} - \ln D^{'}_{t+1}(i)}{\sum_t \alpha_t} \\ &= \frac{\ln c_i - \frac{1}{T} \sum_{t=1}^{T} \ln \sqrt{4\epsilon_t(1-\epsilon_t)} - \frac{1}{T} \ln D^{'}_{t+1}(i)}{\frac{1}{T} \sum_t^T \sqrt{\ln \frac{1-\epsilon_t}{\epsilon_t}}} \end{aligned}$$

Since the data weight is bounded as $0 \leq D^{'}_{t+1}(i) \leq 1.0$, the term $-\ln D^{'}_{t+1}(i)$ is non-negative and the equation imposes the smallest margin on each training data by

$$\text{margin}(x_i, y_i) \geq \frac{\ln c_i - \frac{1}{T}\sum_{t=1}^{T} \ln \sqrt{4\epsilon_t(1-\epsilon_t)}}{\frac{1}{T}\sum_{t}^{T} \sqrt{\ln \frac{1-\epsilon_t}{\epsilon_t}}}. \tag{6}$$

The relation (6) shows the lower boundary of margin depends on its credit value c_i. CAB takes data with higher credits to larger margin but leaves data with lower credits have smaller margin. These properties represent that incorporating credits is equivalent to setting slack variables based on the credits. CAB definitely assumes the nature of soft margin methods but determine its margin relaxation by the previous knowledge.

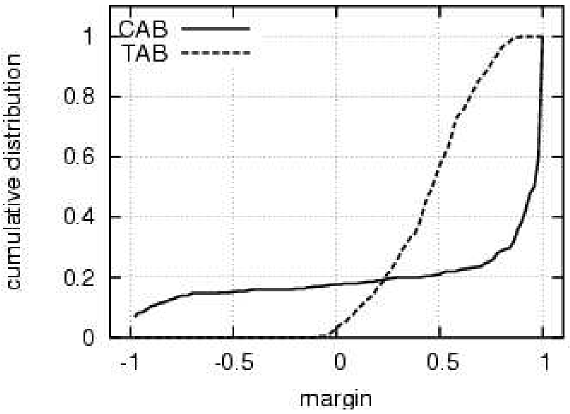

図 1: Margin Distribution of CAB and TAB (*artificial*)

Soft margin properties of CAB are also confirmed by the observaton of margin distribution. To illustrate the properties, we perform a preliminary experiment with *artificial* dataset. The *artificial* dataset is made as follows. The 2-dimensional field ($-2 < x < 2, -2 < y < 2$) and the border $f(x) = 0.2x^3 + exp(-6.0(x - 0.3)^2)$ are prepared. Random 1000 points are generated on the field and each of them is labeled '1' if it falls on the area above the border ($y >= f(x)$), and labeled '-1' otherwise. Of course, this *artificial* set has no information about the persons or situations of labeling. Thus, in accordance with experiments in the previous work, we simulate the sets of expert and non-expert data by dividing the original data into subsets. First, we make training data of $\alpha\%$ of the original data and test data of the rest, as in the traditional validation process. The training data are again divided into simulated expert and non-expert data with the ratio of $\beta\%$ and $(100 - \beta)\%$, respectively, and noises are randomly added only to $\gamma\%$ of non-expert data by flipping their labels. In our preliminary experiment, α, β and γ are fixed at 80%, 20% and 20%. We examine CAB and TAB with *artificial* dataset. Both methods make 100 weak classifiers in learning. We repeat 5-fold Cross Validation (CV) process 10 times and averaged errors are as 4.90% for CAB and 8.98% for TAB, which result clearly represents the improvement by CAB. Figure 1 shows an example of distribution of margins after 100 iterations of CAB and TAB on *artificial* dataset in the preliminary experiments. The distribution of TAB is exemplary. TAB makes margins of all the data above a certain value around 0, even though a non-expert set includes 20% of noises in the population. Larger error rate of TAB can be explained by its overfitting to noises. On the other hand, CAB makes quite diffrent distribution. A portion of data has margin lower than 0 and margins of some instances remain as low as -1, which means they are neglected even when they keep being misclassified all through the learning process. Such distribution is known to appear when a soft magin method is utilized for noisy datasets. The margin distribution and low error rate of CAB suggest that noise data in non-expert data are rightly set with the lower credits and contribute to learning the less. As shown by (6), however, CAB is different from general soft margin methods in that it determines the smallest magin of each data based on its credit. Figure 2 graphically shows credit of each data in *artificial* data and its ob-

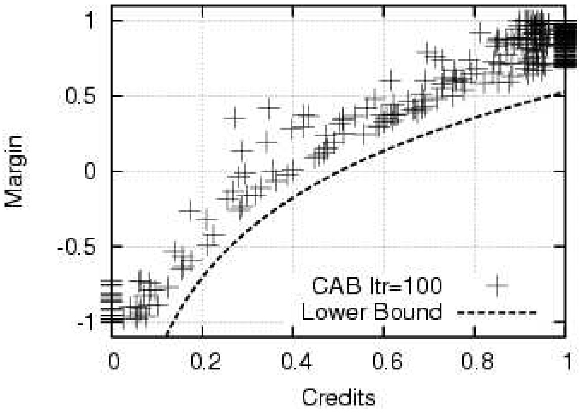

図 2: Credits and Observed Margin (CAB, *artificial*, 4.9% test error) served final margin. The lower boundary calculated by (6) is also shown. We assign observed training errors to ϵ_t for calculation of (6) at each step. The dependency on credits is clear: margin of each data is above the estimated lower boundary which is dependent on $-\ln c_i$. The log-shaped boundary also seems suitable because the smallest margin continuously becomes larger in accordance with its credit. Once credits are estimated, CAB reasonably reflects the credits without any parameter setting. Now that CAB is found to maximize the margins with credit-based slack variables, our interest is in which of introducing credit-based slack variables or general soft margin makes better classifiers. We examine CAB with soft margin methods in the next section.

3.2 Related Works

In addition to soft margin methods, several frameworks are closely related to CAB. One of the most apparently-related is cost-sentitive learning, in which boosting methods are often utilized because of its compatibility with costs. Various cost-sensitive extensions have been proposed as, for example, AdaCost [6], CSB [16] and Asymmetric Boosting [11]. Especially, binary CSB2 [16] reflects costs in updating data weights in AdaBoost process as

$$D_{t+1}(i) = C_{\{+,-\}} D_t(i) \exp(-\alpha y_i h_t(x_i)), \qquad (7)$$

where $C_{\{+,-\}}$ are the misclassification costs for class +1 and class -1. The face of (7) is identifical to CAB's reflection of credits in (2), but our approach differs mainly in two points. Firstly, our approach is taking advantage of the knowledge of which set is more reliably labeled, while cost-sensitive learning algorithms make a difference utilizing the knowledge of which class is more important. Secondly, our method can set continuous credits, while cost-sensitive methods can generally set discrete costs as misclassification of class +1 weighs 100 times more than that of class -1. One possible application of cost-sensitive approach to the LRP is to set a larger cost C_{ex} on expert data and a smaller cost C_{ne} on non-expert data as $(C_{\mathrm{ex}}, C_{\mathrm{ne}}) = (100, 1)$. Our question is which of introducing continuous credits or simple discrete costs makes better classifiers, and we examine this cost-sensitve extension (CSE) to the LRP in the next section. However, when it comes to the LRP, we can presume the superiority of CAB in consideration of training margin again. In the process of updating data weights in cost-sensitive process, the expert and non-expert costs of $(C_{\mathrm{ex}}, C_{\mathrm{ne}}) = (100, 1)$ are normalized and become equivalent to setting the credits of $c_{\mathrm{ex}} = \frac{C_{\mathrm{ex}}}{C_{\mathrm{ex}}+C_{\mathrm{ne}}} \sim 0.99$ and $c_{\mathrm{ne}} = \frac{C_{\mathrm{ne}}}{C_{\mathrm{ex}}+C_{\mathrm{ne}}} \sim 0.01$. The lower boundary in Figure 2 indicates that expert data with $c_{\mathrm{ex}} \sim 0.99$ contribute to learning much as their margins become large, while non-expert data with $c_{\mathrm{ex}} \sim 0.01$ are hardly involved in learning as their margins need not to be larger than -1. As a result, learning will be equivalent to that utilizing only a small set of expert data. Figure 2 also suggests that the less biased cost, for example, $(C_{\mathrm{ex}}, C_{\mathrm{ne}}) = (5, 1)$ which is $(c_{\mathrm{ex}}, c_{\mathrm{ex}}) \sim (0.83, 0.17)$ in terms of credits, makes little difference from $(C_{\mathrm{ex}}, C_{\mathrm{ne}}) = (100, 1)$ because lower boundaries of both credit sets are similar in Figure 2 and non-expert data contribute little in learning in both cases, at least on *artificial* dataset.

In the preliminary experiment, we have compared CAB method with TAB that adopts both of expert and non-expert data equally. Another way to address the LRP is to utilize distribution of non-expert data but none of their labels, that is, semi-supervised learning approach. Roughly, framework of semi-supervised learning includes two main approaches of cluster based algorithms and manifold based algorithms. In cluster based algorithms that assumes data lies in some clusters, learning algorithms generally train the classifiers not only with labeled data but also with unlabeled data in the way of maximizing the margin of unlabeled data. In boosting category, ASSEMBLE (ASBL) [2], which gives pseudo labels to unlabeled data and maximizes the pseudo margins, is one of the most representative methods. ASBL takes advantage of distribution of unlabeled data as they maximize the margin of labeled and unlabeled data. Transductive Support Vector Machine (TSVM) [9] is another representative method to exploit unlabeled data. Our interest is which of which of setting credits or exploiting only their distributions makes better classifiers, and we include semi-supervised learning methods in detailed experiments in the next section.

4 Experimental Results

In this section, we perform detailed experiments to compare CAB with learning methods of related frameworks that can deal with the LRP. As in the previous work, we utilize as a baseline performance

表 1: Summery of Original Datasets

Dataset	Size	Features		Err.
datasets		N	C	
artifcial	1000	2	-	2.4
votes	435	-	16	5.5
breast cancer	683	10	-	3.4
mushroom	5644	-	22	0.2
hypothyroid	2000	6	18	1.9
kr-vs-kp	3196	-	36	1.8
ionosphere	351	34	-	6.7
tic-tac-toe	958	-	8	1.2
spambase	4601	57	-	5.7
credit approval	690	6	9	14.9
heart disease	303	5	9	20.9
iris	150	4	-	4.0
wine	178	13	-	2.4
vehicle	946	18	-	12.3

Traditional AdaBoost that adopt both of expert and non-expert data equally (TAB1), but also adopt TAB that adopt only fine/small expert data (TAB2) to see if small but accurate data will be enough to make a precise classifier without any help of large non-expert data. Another hard margin method, Arc-gv [3] is examined in order to investigate margin properties. In section 3, we have shown that CAB is closely related with frameworks of soft margin approach, cost-sensitive learning and semi-supervised learning. As a method to compare with in soft margin approach. we here utilize a boosting based method of ν-LPBoost (νLPB) [13]. We also utilize general soft margin SVM (CSVM) [5] for another method. We prepare 20 possible parameters of ν in range of 0.02-0.4 for νLPB and $2^k, k = -10, ..., 10$ for CSVM, respectively, and the best parameters are determined by 5-fold cross validation of a training set. We extend cost-sensitive approach to the LRP by setting higher cost on the misclassification of expert data. In the experiments, two sets of costs are introduced as $(C_{\mathrm{ex}}, C_{\mathrm{ne}}) = (100, 1)$ for CSE1 and $(C_{\mathrm{ex}}, C_{\mathrm{ne}}) = (5, 1)$ for CSE2. Learning algorithms of CSE1 and CSE2 are identical to that of usual AdaBoost except that they update data weights in manner of (7). As discussed in the previous section, TSVM and ASBL, which maximizes margins to unlabeled data, are utilized as methods from semi-supervised learning field. We implicitly assume that the methods will be applied to unknown data, TSVM is trained with expert (labeled) and non-expert (unlabeled) data, but not with test (unknown) data, and exmined with its predictions on test data.

For experiments, we utilize artificial and UCI repository datasets [1]. Datasets are summerized in Table 1, where the size of datasets and dimensions of numerical (N) and categorical (C) features are shown. For the purpose of reference, error rates of original datasets, which are estimated by 10 time repeated 5-fold CVs of Tradional AdaBoost, are also shown in the last column. The experiments involves *artificial* data in the preliminary experiment. In working on *ionosphere* and *tic-tac-toe*, we also introduce the manipulation of features in accordance with the previsou work [12], where the number of features alters from 34 to 3 on *ionosphere* and from 8 to 9 on *tic-tac-toe* and examine each methods on both original and modified feature sets. Better accuracy is adopted as a result of each method. In this experiments, multiclass problems of *iris*, *wine* and *vehicle* are converted to one-to-the-rest binary problems, and we simulate expert and non-expert datasets in the same way of the preliminary experiment. All methods are simultaneously examined with 5 time repeated 5-fold CVs. We employ decision stump as the weak learner for boosting based methods. Table 2 shows the average error rates of each method. On whole datasets CAB generally performs well and gets the top rating on 7 of 14 datasets. The relation between credits and margins as seen in Figure 2 are also observed in all of datasets (not shown). On the other hand, the performance of TAB1 and TAB2 clearly illustrates the dilemma of the LRP. TAB1 suffers from proportinal noises in large non-expert data, while TAB2 suffers from smallness of fine expert data. As expected from its hard margin property, Arc-gv suffers more from noises and the performance is inferior to TAB1. Soft margin methods of CSVM and νLPB are competitive and attain the best performance on 5 datasets. But as the performance of νLPB on *ionosphere* and *heart disease*, they are sometimes prone to the properties of datasets and heavily fail if datasets happen not to match. The high computational cost in the search of good parameters is another problem and discussed later. Table 3 shows the smallest margins of boosting based algorithms CAB, νLPB, TAB1 and Arc-gv for first 5 datasets. Properties of hard and soft margin classifiers are obeserved clearly. While CAB and νLPB can leave the smallest margin, TAB1 and Arc-gv manage to maximize the smallest margin on all datasets. Larger margins and higher error rates of Arc-gv obviously represent the overfitting problem, as the more Arc-gv persists in maximizing the smallest margin, the poorer the performance of the final

表 2: Error Rates

		Hard Margin			Soft Margin		Semi-Supervised		Cost-Sensitive	
	CAB	TAB1	TAB2	Arc-gv	CSVM	νLPB	TSVM	ASBL	CSE1	CSE2
ARTIFICIAL	4.90	8.98	5.41	9.78	6.45	18.9	-	14.4	5.25	5.23
VOTES	5.87	6.17	6.26	6.83	4.99	5.59	7.36	5.20	6.27	6.10
BREAST CANCER	4.40	5.92	5.27	6.30	4.22	7.32	3.91	9.19	5.33	5.24
THYROID	1.80	2.55	1.90	2.80	4.25	2.64	2.75	3.40	1.90	1.95
KR-VS-KP	4.85	6.36	5.00	6.98	6.10	12.5	11.3	31.8	5.05	5.11
MUSHROOM	0.00	0.72	0.00	0.20	5.33	10.2	4.31	10.2	0.00	0.00
IONOSPHERE	12.6	14.8	14.0	15.2	18.7	24.3	15.6	18.4	13.7	13.6
TIC-TAC-TOE	1.84	3.37	3.65	2.46	30.3	0.0	27.3	25.0	3.45	6.78
SPAMBASE	6.80	6.94	7.02	7.75	11.7	6.87	10.2	21.5	7.00	7.81
CREDIT APPROVAL	16.7	17.4	18.2	17.7	16.7	27.7	16.8	15.8	18.2	17.6
HEART DISEASE	22.2	28.1	25.1	28.0	20.9	39.0	21.9	28.0	24.1	25.2
IRIS	5.22	9.67	6.67	11.1	12.3	11.1	17.2	13.1	6.89	7.89
WINE	4.03	14.5	7.57	15.7	3.27	10.1	5.27	12.2	6.35	6.24
VEHICLE	15.2	15.5	15.8	15.6	18.7	13.2	23.9	19.3	15.6	16.3

表 3: Smallest Margin of Hard/Soft Margin Classifiers

	CAB	ν-LPB	TAB1	Arc-gv
ARTFICIAL	-0.86	-0.54	-0.29	-0.01
VOTES	-0.86	-0.90	-0.34	-0.01
BREAST-CANCER	-0.80	-0.51	-0.25	-0.01
THYROID	-0.77	-1.00	-0.31	-0.01
KR-VS-KP	-0.82	-1.00	-0.59	-0.01

表 4: Computation Time (Normalized to TAB1)

	CAB	νLPB	CSVM	TSVM
VOTES	1.7	68.1	591.1	16.2
BREAST-CANCER	1.3	311.3	926.9	32.9
TIC-TAC-TOE	1.0	336.7	2208.1	3372.3
SPAMBASE	1.0	229.6	408.8	52.1
CREDIT APPROVAL	1.0	1721.8	827.9	22092
HEART DISEASE	1.0	1031.7	2276.8	38.4
WINE	2.6	92.4	861.3	222.1
VEHICLE	1.8	34.1	391.2	30.4

classifier becomes.

TSVM and ASBL performs best on *breast cancer* and *credit approval*, respectively, but they also show dependencies on the properties of datasets. For examle, on *artificial* where cluster assumption does not holds because of its random generation process, TSVM cannot learn a resultant classifier in reasonable learning time. Except for a few datasets, the error rates suggest that utilizing non-experts' labeling as in CAB or soft margin methods leads to better classifiers than discarding them as in TAB2 or semi-supervised methods, even if non-expert data includes as high ratio of noise labels as $\gamma = 20\%$. CSE1 and CSE2, applications of cost-sensitive learning methods, show the expected performances. The large difference of $C_{\mathrm{ex}} = 100$ in CSE1 and $C_{\mathrm{ex}} = 5$ in CSE2 does not make proportional difference in error rates, and they are actually accordant with TAB2, as presumed by the lower boundary of the smallest margin in section 3.

In addition to its stable accuracy, CAB has an advantage that it does not have to set heuristic regularization parameter as needed for general soft margin classifiers. For example, νLPB needs parameter ν which represents the fraction of data with margin errors above a constant ρ. The meaning of ν is more intuitive than purely abstract parameters, but it is still hard to determine appropriate value of ν definitely and νLPB generally has to depend on CV process to determine ν. CV process imposes costly computations, as learning must be repeated until all possible parameters are examined. In these experiments, νLPB and CSVM search 20 possible values and require high costs for the detemination of the best parameter. On the other hand, CAB needs calculation of credits only once before starting learning. In addition, the values of credits represent confidences of the labels and thus are intuitive. Table 4 shows the averaged computation time of CAB, νLPB, CSVM and TSVM for datasets on which the performances of soft margin methods or TSVM are competitive. For purpose of illustration, the computation times of TAB1 are normalized to 1 for all shown data. Other boosting based methods as ASBL and CSE requires similar costs with TAB1 and abbreviated here. The computation is performed on the PC with Intel Pentium 4 2.4GHz CPU, and CAB uses kd-tree algorithms to determine credits. Obviously, intensive searches in CSVM and νLPB processes require much higher costs, even up to 100-1,000 times as long as TAB1 to learn final classifiers. The costs of TSVM also reaches to dozens of times the cost of TAB1. Depending on distribution of training data, TSVM takes much longer time to learn a classifier, above 10,000 times as long at its worst, which results in extraordinary costs on *tic-tac-toe* and *credit approval* data. On the contrary, the cost of CAB is not so much different as that of TAB1, showing the extra of calculating credits does not cause serious additional costs. Similar results are obtained on the other datasets. Though costs of soft margin classifiers may vary on their ways of parameter determination, the results suggests that CAB perform as well as soft margin classifiers with reasonable computational costs.

5 Other Extensions

Experimetal results show that CAB can address the LRP, but several extensions should be discussed for applying our approach to real world data. Two major examples of extension can be considered for CAB. Firstly, for practical use, CAB has to be extended to multiclass problems. CAB has assumed binary class problems partially because the way of introducing manual noise to the simulated non-expert data become simple and fair for experiments. Extending CAB to multiclass problems can be done, for example, by adopting a standard multiclass algorithm of AdaBoost.M2 [7]. We incorporate estimated credits into the data weight of each class as $w_{i,y}^{\prime t} = c_i w_{i,y}^t$, where $w_{i,y}^t$ is a data weight on (x_i, y) for t-th weak classifier. This incorporation naturally extends CAB to multiclass problems. Secondly, a credit determination method can be appropriately chosen on the problems. CAB adopts standard Euclidean distance and calculates kNN based credits on non-expert data. One can easily replace the Euclidean estimation with appropriate measures, for example, cosine similarity for text classification problems. As seen in Table 2, the performances of CAB seems a little inferior on high dimensional data, and these degradation may be attributed to the difficulty in estimating accurate credits, as a kNN classification method often fails on high dimensional data. Some feature selections or alternations will be effective in estimation of credits as shown in the previous work [12]. To address the LRP, we have focused on the properties and performances of the simple boosting model in this paper, but there must be many possible ways to address the LRP. For example, Kamishima et al. recently propose a bagging based method in order to address 'Taming' problem which is of same concept as the LRP [10]. They work on the real world collaborative tagging data set and examine their method. In addition to the boosting based methods, adopting adequate approaches to the real world LRP is one of our future works.

6 Conclusion

We investigated a credit-based classification method CAB that utilizes reliably labeled data, mainly focusing on how CAB yields more precise classifiers. We have shown that introducing credits is equivalent to setting slack variables which is dependent on the values of credits. We performed detailed experiments that examine CAB with methods of closely related frameworks, that is, soft margin approach, cost-sensitive learning and semi-supervised learning. The results showed that CAB stably performs better in 7 of 14 datasets. Although soft margin methods are competative with CAB in several datasets, CAB are still advantageous in its computational costs, because it required no heuristic parameter search which is generally necessary for soft margin classifiers. As we have concentrated on investigating the properties and performances of CAB, it is possible to extend the approach to modifying CAB or adopting other learning methods. As a future work, we plan to apply the methods to real-world data that are actually labeled by experts and non-experts.

参考文献

[1] Asuncion A., Newman D.J.: UCI Machine Learning Repository, *http://www.ics.uci.edu/∼mlearn/MLRepository.html* (2007)

[2] Bennett, K.P., Demiriz, A., Maclin, R.: Exploiting unlabeled data in ensemble methods, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 289–296 (2002)

[3] Breiman, L.: Prediction Games and Arcing Algorithms, *Neural Computation*, Vol. 11, No. 7, oo. 1493–1517 (1999)

[4] Chapelle, O., Schölkopf, B., Zien, A., NetLibrary, I.: Semi-supervised learning (2006)

[5] Cortes, C., Vapnik, V.: Support-vector networks, *Machine Learning*, Vol. 20, No. 3, pp. 273–297 (1995)

[6] Fan, W., Stolfo, S., Zhang, J., Chan, P.K.: AdaCost: misclassification cost-sensitive boosting, *Proceedings of the 16th International Conference on Machine Learning*, pp. 97–105 (1999)

[7] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting, *European Conference on Computational Learning Theory* pp. 23–37 (1995)

[8] Freund, Y., Schapire, R.: Experiments with a New Boosting Algorithm, *Proceedings of the 13th International Conference on Machine Learning*, pp. 148–156 (1996)

[9] Joachims, T.: Transductive Inference for Text Classification using Support Vector Machines, *Proceedings of the 17th International Conference on Machine Learning*, pp. 200–209 (1999)

[10] Kamishima, T., Hamasaki, M., Akaho, S.: BaggTaming?Learning from Wild and Tame Data, *ECML/PKDD2008 Workshop: Wikis, Blogs, Bookmarking Tools–Mining the Web 2.0 Workshop* (2008)

[11] Masnadi-Shirazi, H., Vasconcelos, N.: Asymmetric boosting, *Proceedings of the 24th international conference on Machine learning*, pp. 609–619 (2007)

[12] Nakata, K., Sakurai, S., Orihara, R.: Classification Method Utilizing Reliably Labeled Data, *Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part I* , pp. 114–122 (2008)

[13] Rätsch, G., Scholkopf, B., Smola, A., Müller, K.R., Onoda, T., Mika, S.: ν-arc ensemble learning in the presence of outliers, *Advances in Neural Information Processing Systems*, Vol. 12 (2000)

[14] Rätsch, G., Onoda, T., Müller, K.R.: Soft Margins for AdaBoost, *Machine Learning*, Vol. 42, No. 3, pp. 287–320 (2001)

[15] Schapire, R.E., Freund, Y., Bartlett, P. Lee, W.S.: Boosting the margin: a new explanation for the effectiveness of voting methods, *Proceedings of the 14th International Conference on Machine Learning*, pp. 322–330 (1997)

[16] Ting, K.M.: A Comparative Study of Cost-Sensitive Boosting Algorithms, *Proceedings of the 17th International Conference on Machine Learning*, pp. 983–990, (2000)