Active Learning for Multiclass Cost-Sensitive Classification Using Probabilistic Models

Po-Lung Chen
Dept. of Computer Science & Information Engineering
National Taiwan University
Taipei, Taiwan
r99922038@csie.ntu.edu.tw

Hsuan-Tien Lin
Dept. of Computer Science & Information Engineering
National Taiwan University
Taipei, Taiwan
htlin@csie.ntu.edu.tw

***Abstract*—Multiclass cost-sensitive active learning is a relatively new problem. In this paper, we derive the *maximum expected cost* and *cost-weighted minimum margin* strategies for multiclass cost-sensitive active learning. The two strategies can be viewed as extended versions of the classical cost-insensitive active learning strategies. The experimental results demonstrate that the derived strategies are promising for cost-sensitive active learning. In particular, the cost-sensitive strategies out-perform cost-insensitive ones on many benchmark data-sets and justify that an appropriate consideration of the cost information is important for solving cost-sensitive active learning problems.**

***Keywords*-Active learning, Multiclass, Cost-sensitive**

I. Introduction

In many applications of machine learning [22], it is expensive to label the examples. For instance, it usually requires hiring a medical expert to make diagnoses (labels) for x-ray scans (examples). An active learning [16] setup aims to achieve decent learning performance by querying the labels of a few examples strategically, therefore saving the labeling expense. This setup has been studied in a variety of machine learning problems [13], [21]. Many existing works [9], [19] on active learning focus on binary classification. An intuitive and successful paradigm in active learning for binary classification is called uncertainty sampling, which queries the most ambiguous examples. i.e., the examples near the boundary of the two classes.

An extended problem of binary classification is multiclass classification, where examples can be associated with one of the $M > 2$ labels. The multiple boundaries (for different classes) of multiclass classification make it more difficult to define the uncertainty of examples for active learning by the closeness to the boundaries. Therefore, many recent works [12], [14] of active learning for multiclass classification resort to the use of a probabilistic model for calculating uncertainty.

Traditionally, the evaluation criterion of multiclass classification problem is error rate. Cost-sensitive classification is an extended problem that assigns different penalties for different types of misclassification. For instance, consider a three-class classification problem for predicting the state of patient as (1) SARS-infected, (2) cold-infected, (3) - healthy. When predicting a SARS-infected patient as healthy, the society could suffer from a relatively large cost when compared with, for instance, predicting a cold-infected patient as healthy. The multiclass cost-sensitive classification problem has been an important research direction for more than a decade [6]. Among many approaches [1], [18] for multiclass cost-sensitive classification, an important and pioneering family of approaches [6], [7] is based on probabilistic models. In short, the approaches estimate the class probability of each example, and aim at predicting with the lowest expected cost based on estimation.

It has been reported [1], [18] an appropriate consideration of cost information can improve the performance of cost-sensitive classification in a non-active learning setting. The problem of active learning for multiclass cost-sensitive classification, shortened as *cost-sensitive active learning*, is relatively new. And thus, it is not clear whether the addition of cost information would be advantageous. In particular, the strategies for appropriately including cost information for cost-sensitive active learning have not been thoroughly studied. Given the wide range of potential applications for cost-sensitive active learning, we are interested in exploring strategies for cost-sensitive active learning.

The probabilistic model has been an important component in both multiclass cost-sensitive classification and multiclass active learning. In case of multiclass cost-sensitive classification, the model provides a natural decision function by calculating expected cost [6], [7]. In case of multiclass active learning, the model is widely used for designing querying strategies that combine uncertainty values from multiple boundaries [11], [12]. Given the importance described above, we focus on designing strategies for cost-sensitive active learning using probabilistic models.

In this paper, we propose two novel strategies for cost-sensitive active learning using probabilistic models. The idea for deriving the strategies is *maximum expected cost reduction*, which aims at choosing the examples that will reduce expected cost the most. One strategy selects examples with *minimum expected cost*, and can be considered an extension of a classical (cost-insensitive) active learning strategy called *minimum confidence*. Another strategy extends from cost-insensitive strategy *minimum margin* and is called *cost-weighted minimum margin*. Experimental results demonstrate that the proposed cost-sensitive strategies generally outperform other strategies for cost-sensitive active learning.

The rest of this paper is organized as follows. In Section II, we formally introduce the setup of cost-sensitive active learning, and discuss some existing and related strategies for binary and multiclass active learning. Then, in Section III, we derive the proposed strategies. We present the experimental results in Section IV and the conclusion in Section V.

II. Multiclass and cost-sensitive active learning

A. Setup of multiclass active learning

In multiclass classification, we seek a classifier that maps each example $x \in \mathcal{X} \subseteq \mathcal{R}^d$ to a label $y \in \mathcal{Y} = \{1, 2, ..., M\}$, where M denotes the number of classes. Given a data set $\mathcal{D}$ that contains N training examples represented as $\{(x_n, y_n)\}, n = 1, ..., N$, a multiclass learner $\mathcal{F}$ learns a decision function $f^{\mathcal{D}} : \mathcal{X} \rightarrow \mathcal{Y}$ from $\mathcal{D}$, which can be used for predicting the label of any future test example.

The pool-based setup for (multiclass) active learning has been studied in many previous works [12], [20]. In this setup, an active learning algorithm starts from a given labeled pool $\mathcal{D}$ and then iteratively queries an oracle to obtain the label of some examples from an unlabeled pool $\mathcal{D}_u = \{x_i\}, i = 1, ..., N_u$. In each of the R rounds of queries, the algorithm can obtain the labels of K examples $\mathcal{D}^+$ from $\mathcal{D}_u$. Then, $\mathcal{D}^+$ and the corresponding labels are added into $\mathcal{D}$, and $\mathcal{D}^+$ is removed from $\mathcal{D}_u$. The method of selecting the K examples is called a query strategy $\mathcal{S}$, which takes $\mathcal{D}, \mathcal{D}_u$, and K into consideration. The learned model $f^{\mathcal{D}}$ is evaluated at the end of each round using an independently sampled test set $\mathcal{D}_t$. This setup aims to obtain a decision function $f^{\mathcal{D}}$ that achieves a low classification error on $\mathcal{D}_t$ after a small number of querying rounds.

B. Setup of cost-sensitive active learning

In this work, we assume that the cost is provided as an $M \times M$ cost matrix C, where $C(a, b)$ represents the cost to be paid when predicting an example of label a as label b.When adopting the cost-sensitive setting in active learning framework, there are three changes. Firstly, the example selection strategy $\mathcal{S}$ can take the given cost matrix as input. Secondly, instead of classification error, we use classification cost $\sum_{(x,y)\in\mathcal{D}_t} C(y, f^{\mathcal{D}}(x))$ to evaluate the strategies. Lastly, the learner $\mathcal{F}$ can also take the given cost matrix as input. The cost-sensitive active learning setup that we will study is shown in Algorithm 1.

Algorithm 1 Pool-based active learning for multiclass cost-sensitive classification

```
Input: a labeled pool $\mathcal{D}$, an unlabeled pool $\mathcal{D}_u$, the number of rounds $R$, the number of queries in a round $K$, a multiclass learner $\mathcal{F}$, a cost matrix $C$, a labeling oracle
for $i = 1...R$ do
    $\mathcal{D}^+ \leftarrow \mathcal{S}(\mathcal{D}, \mathcal{D}_u, K)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup (\mathcal{D}^+, \text{Oracle.label}(\mathcal{D}^+))$
    $\mathcal{D}_u \leftarrow \mathcal{D}_u \backslash \mathcal{D}^+$
    $f^{\mathcal{D}} \leftarrow \mathcal{F}(\mathcal{D}, C)$
    Evaluate$(f^{\mathcal{D}}, \mathcal{D}_t, C)$
end for
```

C. Probabilistic model

In this work, we consider probabilistic models to assist both $\mathcal{S}$ and $\mathcal{F}$. The probabilistic models output conditional distribution $P(y|x)$ after training. The use of probabilistic models for $\mathcal{F}$ has been established in earlier works [6], [7]. In particular, a reasonable $\mathcal{F}$ can use P to predict the class with minimum expected cost. That is, the decision function can be written as

$$f(x) = \underset{j\in\mathcal{Y}}{\text{argmin}} \sum_{k=1}^{M} P(y = k|x)C(k, j). \tag{1}$$

The probabilistic models have also been utilized in many previous cost-insensitive active learning works [11], [12], where P assists the combination of the confidence values associated with the multiple boundaries. In the next section, we will introduce some existing cost-insensitive active learning strategies that are based on probabilistic models.

D. Existing cost-insensitive active learning strategies

1) Binary active learning: Active learning for binary classification (binary active learning) has been studied in many works [9], [14]. One of the most popular and successful strategies in binary active learning is uncertainty sampling. It focuses on querying the labels of the most ambiguous examples. Most uncertainty sampling strategies rely on a measurement of uncertainty. For example, [17] defines uncertainty by the distance between an example and the decision hyperplane of the support vector machine [5]. The resulting strategy queries the label of the example that is closest to the decision hyperplane. In the case of probabilistic model, the uncertainty can be defined using conditional probability estimates [2]. For example, consider a probabilistic model that outputs $P(y = +1|x)$. A natural sampling strategy is used for querying the examples with $P(y = +1|x) \approx P(y = -1|x)$. Empirically, uncertainty sampling is a promising baseline approach for binary active learning.

Another important strategy for active learning is *maximum expected error reduction* [16], which queries examples that reduce the expected error the most. That is, *maximum expected error reduction* queries by solving

$$\underset{\mathcal{D}^+}{\text{argmax}} \sum_{\mathcal{D}\cup\mathcal{D}_u} E_{\text{error}}(f^{\mathcal{D}}) - E_{\text{error}}(f^{\mathcal{D}\cup\mathcal{D}^+}). \qquad (2)$$

where $E_{\text{error}}(f)$ denotes the expected error made by the classifier f. In the case of binary classification, the resulting strategy is usually similar to uncertainty sampling strategies.

2) Multiclass active learning: In binary active learning, many existing uncertainty sampling strategies consider the relation between examples and the decision boundary, and define uncertainty using this information. In multiclass active learning, there are multiple decision boundaries between classes. For example, there are in total $\frac{M\times(M-1)}{2}$ decision hyperplanes in a one-versus-one SVM classifier, and each of them may suggest a different uncertainty value for an example. When extending binary active learning strategies to a multiclass problem, an appropriate combination of uncertainty values is an ongoing research issue.

There are several existing works on multiclass active learning, and most of them focus on uncertainty-based strategies. For example, [21] computes the uncertainty using the loss associated with some binary classification subproblems that are constructed by output coding. Further, [11], [12], and [13] adopt a model for estimating the class probability distribution $P(y|x)$. Then, the uncertainty can be easily defined by using P asa confidence measurement of an example x.

Next, we introduce two of the most representative uncertainty sampling (cost-insensitive) active learning strategies, *minimum confidence* and *minimum margin*, which can be coupled with probabilistic models.

3) Minimum confidence: The *minimum confidence* strategy selects the least confident examples to label. In the case of probabilistic model, the confidence for an example can be defined as the class probability of the predicted class. That is, the strategy selects

$$\underset{\mathcal{D}^+}{\text{argmin}} \sum_{x\in\mathcal{D}^+} P(y = f^{\mathcal{D}}(x)|x, \mathcal{D}), \qquad (3)$$

where $f^{\mathcal{D}}(x)$ always chooses the most probable class $f^{\mathcal{D}}(x) = \underset{y}{\text{argmax}}\, P(y|x, \mathcal{D})$.

4) Minimum margin: In addition to the most probable class, *minimum margin* considers the information of the second probable class. This strategy chooses the examples with the minimum confidence difference between the most and the second most probable classes.

$$\begin{aligned}\underset{\mathcal{D}^+}{\text{argmin}} \sum_{x\in\mathcal{D}^+} \quad & \big(P(y = f^{\mathcal{D}}(x)|x, \mathcal{D}) \\ & -P(y = f^{\mathcal{D}}_{\text{second}}(x)|x, \mathcal{D})\big)\,, \qquad (4)\end{aligned}$$

where $f^{\mathcal{D}}(x)$ returns the most probable class and $f^{\mathcal{D}}_{\text{second}}(x) = \underset{y\neq f^{\mathcal{D}}(x)}{\text{argmax}}\; P(y|x, \mathcal{D})$ is the second most probable class. The use of *minimum margin* for multiclass active learning has been analyzed in [12].

III. Strategies for cost-sensitive active learning

A. Strategy: Maximum expected cost

The first strategy that we propose is *maximum expected cost*. We show the derivation steps for the cost-sensitive setting.

Basic idea. Because the goal of cost-sensitive active learning is to achieve a low cost, a reasonable cost-sensitive active learning strategy is used for obtaining a query set of unlabeled examples $\mathcal{D}^+$, which can minimize the expected cost

$$\begin{aligned}& E_{\text{cost}}(f_C^{\mathcal{D}\cup\mathcal{D}^+}) \\ = & \sum_{x\in\mathcal{D}\cup\mathcal{D}_u} \sum_{k=1}^{M} P(y = k|x) C(k, f_C^{\mathcal{D}\cup\mathcal{D}^+}(x_i)). \quad (5)\end{aligned}$$

The difficulty for the minimization of (5) is the need of checking every possible $\mathcal{D}^+$ with the unknown labels. Even if we want to approximate $f_C^{\mathcal{D}\cup\mathcal{D}^+}(x_i)$ by assuming the labels in $\mathcal{D}^+$, we still need to train classifiers $f_C^{\mathcal{D}\cup\mathcal{D}^+}$ for all $M^{|\mathcal{D}^+|}$ combinations, which is computationally infeasible in many cases.

Maximum expected cost reduction. To seek a computationally feasible approach, we start by mimicking the derivation of the maximum error reduction. Because the size of $\mathcal{D}^+$ is usually considerably smaller than $\mathcal{D}$, $f_C^{\mathcal{D}}$ and $f_C^{\mathcal{D}\cup\mathcal{D}^+}$ shall be similar. Thus, we attempt to determine the query set $\mathcal{D}^+$ with the help of $f_C^{\mathcal{D}}$. By considering the expected cost of $f_C^{\mathcal{D}}$, the optimal $\mathcal{D}^+$ shall maximize the amount of the expected cost reduction. That is,

$$\mathcal{D}^+_{opt} = \underset{\mathcal{D}^+}{\text{argmax}} \left(E_{\text{cost}}(f_C^{\mathcal{D}}) - E_{\text{cost}}(f_C^{\mathcal{D}\cup\mathcal{D}^+})\right).$$

Further, we replace $E_{\text{cost}}(f_C^{\mathcal{D}})$ and $E_{\text{cost}}(f_C^{\mathcal{D}\cup\mathcal{D}^+})$ with (5) and use $E^x_{\text{cost}}(f_C^{\mathcal{D}})$ to denote the expected cost of $f_C^{\mathcal{D}}$ on x. Thus $\mathcal{D}^+$ can be calculated as follows:

$$\mathcal{D}^+_{opt} = \underset{\mathcal{D}^+}{\text{argmax}} \sum_{x\in\mathcal{U}} E^x_{\text{cost}}(f_C^{\mathcal{D}}) - E^x_{\text{cost}}(f_C^{\mathcal{D}\cup\mathcal{D}^+}). \qquad (6)$$

Approximation. We first attempt to eliminate the most computationally expensive component $E^x_{\text{cost}}(f_C^{\mathcal{D}\cup\mathcal{D}^+})$. For $\{x|x \in \mathcal{U}, x \notin \mathcal{D}^+\}$, $f_C^{\mathcal{D}}(x)$ and $f_C^{\mathcal{D}\cup\mathcal{D}^+}(x)$ are expected to be close. This is attributed to the fact that x is either shared by both sets or lies out of both sets. Hence, the two classifiers should share similar predictions on x. Then, we can substitute $\mathcal{U}$ with $\mathcal{D}^+$ and obtain

$$\mathcal{D}^+_{opt} \approx \underset{\mathcal{D}^+}{\text{argmax}} \sum_{x\in\mathcal{D}^+} E^x_{\text{cost}}(f_C^{\mathcal{D}}) - E^x_{\text{cost}}(f_C^{\mathcal{D}\cup\mathcal{D}^+}).$$

In case of $x \in \mathcal{D}^+$, $E^x_{\text{cost}}(f_C^{\mathcal{D}\cup\mathcal{D}^+})$ is expected to be small because x is within the training set of $f_C^{\mathcal{D}\cup\mathcal{D}^+}$. Thus, $f_C^{\mathcal{D}\cup\mathcal{D}^+}$ can probably make a low-cost prediction for any $x \in \mathcal{D}^+$. Thus, we can approximate the difference $E^x_{\text{cost}}(f^{\mathcal{D}}_C) - E^x_{\text{cost}}(f^{\mathcal{D}\cup\mathcal{D}^+}_C)$ by $E^x_{\text{cost}}(f^{\mathcal{D}}_C)$, and obtain

$$\begin{aligned} \mathcal{D}^+_{opt} &\approx \underset{\mathcal{D}^+}{\text{argmax}} \sum_{x\in\mathcal{D}^+} E^x_{\text{cost}}(f^{\mathcal{D}}_C) && (7) \\ &= \underset{\mathcal{D}^+}{\text{argmax}} \sum_{x\in\mathcal{D}^+} \sum_{k=1}^{M} P(y=k|x) C(k, f^{\mathcal{D}}_C(x)). && (8) \end{aligned}$$

Relationship between minimum confidence and maximum expected cost. The *maximum expected cost* is very similar to *minimum confidence*. In our setting, *minimum confidence* can be written as

$$\mathcal{D}^+_{opt} = \underset{\mathcal{D}^+}{\text{argmax}} \sum_{x\in\mathcal{D}^+} \sum_{k=1}^{M} P(y=k|x) Err(k, f^{\mathcal{D}}(x)), \quad (9)$$

where $Err(k, f^{\mathcal{D}}(x)) = 1$ if $k \neq f^{\mathcal{D}}(x)$, otherwise $Err(k, f^{\mathcal{D}}(x)) = 0$. The only difference between (8) and (9) is C and Err. That is, *minimum confidence* can be viewed as a special case of *maximum expected cost* that considers Err to be C.

B. Strategy: Cost-weighted minimum margin

As discussed in Section II, another popular strategy for cost-insensitive active learning is *minimum margin*, which focuses on the first and second most probable classes.

Similar to the above relationship between *minimum confidence* and *maximum expected cost*, we replace the confidence terms with the expected cost in (4) to obtain a cost-sensitive version of *minimum margin* strategy. The resulting strategy is

$$\mathcal{D}^+_{opt} = \underset{\mathcal{D}^+}{\text{argmin}} \sum_{x\in\mathcal{D}^+} \left(E^x_{\text{cost}}(f^{\mathcal{D}}_{C,\text{second}}) - E^x_{\text{cost}}(f^{\mathcal{D}}_C) \right). \quad (10)$$

Note that the $f^{\mathcal{D}}_{C,\text{second}}$ here is different from the one in (4). It predicts the class with second lowest expected cost. This strategy chooses the examples with the smallest gap between the first-choice and the second-choice classes in terms of the expected cost. Thus we call this strategy as *cost-weighted minimum margin* in the following discussions.

IV. Experiments

Next, we evaluate the proposed cost-sensitive strategies and compare them with cost-insensitive ones on real-world multiclass datasets.

A. Experiment settings

Data set. We consider ten benchmark multiclass datasets in the experiment. Every dataset is split into a training set and a test set, with 50% examples in the training set and the rest examples in the test set. The training set is used as the pool $\mathcal{U}$ and the test set is used for evaluating the performance of the strategies. Table I shows the summary of these datasets.

Table I
Summary of multiclass datasets

Name	Class	# of examples	Feature
Wine	3	178	13
Glass	6	264	9
SVMguide4	6	612	10
Vehicle	4	846	18
Vowel	11	990	10
Segment	7	2310	19
Dna	3	3186	180
Satimage	6	6435	36
USPS	10	9298	256
Pendigits	10	10992	16

Probabilistic model. We present the results on two types of probabilistic models. One is the one-versus-one SVM with probability estimation, which has been in use for many years and its application for active learning is studied in [12]. This model applies Platt's method [15] on multiclass (one-versus-one) SVM to obtain a probabilistic output from discrete classification results. The details of the multiclass probability estimation step are described in [10]. We use the implementation in LIBSVM [4] Tools to realize the probabilistic learner and take $C = 1$ with a linear kernel.

The other type of model is random forest [3]. In our setting, we use the implementation in Weka [8]. Random forest determine the probability values by the voting of trees.

Query strategy. We compare five query strategies in our experiment. In addition to *maximum expected cost* and *cost-weighted minimum margin* proposed in Section III, we include *minimum confidence* and *minimum margin* mentioned in Section II, to observe the influence of adding the cost information. We also consider the *random* strategy, which randomly selects examples to label. In the past active learning research, the *random* strategy usually shows a certain strength that is comparable to many active learning strategies [12].

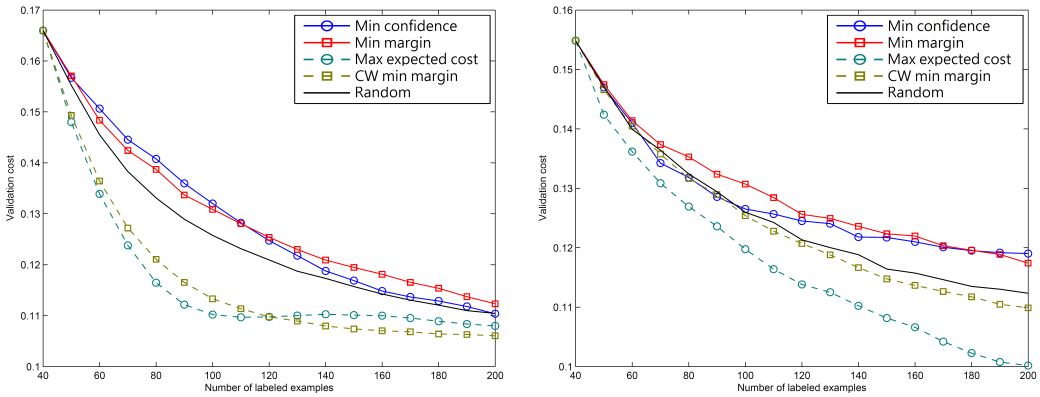
Cost matrix. We adopt the most popular setup in cost-sensitive classification, randomized proportional, as used by [1].

Evaluation. We evaluate the strategies on the test set. A classifier with a relatively low average classification cost $\frac{1}{|\mathcal{D}_t|}\sum_{(x,y)\in\mathcal{D}_t} C(y, f(x))$ is judged as a better classifier.

B. Comparison of strategies

In this section, we compare the performance of strategies under several situations. Figure 1 shows the active learning results on the dataset Vehicle. From this figure we can see that the cost-sensitive strategies behave better than the other strategies and are saturated when approximately 100 out of 400 labels are known in Figure 1(a). The cost-insensitive active learning strategies (*minimum confidence* and *minimum margin*) act worst. The *random* strategy is a stable choice in between.

Next, we compare the strategies on all datasets by randomly choosing 10% of the examples in pool $\mathcal{U}$ as the initial training set and then run 10 rounds of active learning. In each round, we select 1% of the examples for querying. That is, after 10 rounds, 20% of the examples in the pool will have



(a) One-versus-one SVM with probability estimation

(b) Random forest

Figure 1. Cost comparison on dataset `Vehicle`

been labeled. We present two values, the cost at the end of active learning and the AUC (area under curve) of cost, in the following experiments.

Comparison of strategies using one-versus-one SVM with probability estimation. In Table II, we list the results using one-versus-one SVM with probability estimation as the probabilistic model. The best ending cost and AUC for each dataset is marked in bold. The results show that *maximum expected cost* and *cost-weighted minimum margin* usually perform better than the other criteria. *Minimum margin* is often promising in terms of AUC but not good at the ending cost. The other two strategies generally fall behind.

Comparison of strategies using random forest. We also carry out a comparison for strategies using random forest as the probabilistic model. The results are shown in Table III. The *cost-weighted minimum margin* strategy achieves the best average rank, while *maximum expected cost* shows similar performance. Cost-insensitive strategies do not perform well in this experiment, and can be even worse than *random*. Note that, when using random forest as the probabilistic model, some active learning strategies lead to an increase in the cost. We have marked the values whose ending cost is greater than the starting cost with *. This situation usually occurs when *minimum confidence* and *minimum margin* are used for active learning, and will be an interesting future research direction to study the cause.

The above comparisons confirm the importance of using cost-sensitive strategies for cost-sensitive active learning.

V. Conclusion

In this research, we studied cost-sensitive active learning using probabilistic models. We derived the *maximum expected cost* strategy based on the idea of *maximum expected cost reduction*. Then, we replace the confidence term in *minimum margin* with the expected cost to get a cost-sensitive version of *minimum margin* strategy called *cost-weighted minimum margin*. We compared the performance of cost-sensitive strategies with that of cost-insensitive ones and a *random* strategy. The results revealed that cost-sensitive strategies often outperformed cost-insensitive ones, and thus, it was important to use cost-sensitive strategies for cost-sensitive active learning.

References

[1] N. Abe, B. Zadrozny, and J. Langford, *An iterative method for multi-class cost-sensitive learning*, In Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2004.

[2] B. Anderson and Andrew Moore, *Active learning for hidden markov models: objective functions and algorithms*, In Proceedings of the 22nd International Conference on Machine Learning, 2005.

[3] L. Breiman, *Random forests*, Machine Learning, 2001.

[4] C.-C. Chang and C.-J. Lin, *LIBSVM : a library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, 2011.

[5] C. Cortes and V. Vapnik, *Support-vector networks*, Machine Learning, 1995.

[6] P. Domingos, *MetaCost: A General Method for Making Classifiers Cost-Sensitive*, In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.

[7] C. Elkan, *The foundations of cost-sensitive learning*, In Proceedings of the 17th International Joint Conference on Artificial Intelligence, 2001.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, *The WEKA data mining software: an update*, SIGKDD Explorations, 2009.

[9] S. J. Huang, R. Jin, and Z. H. Zhou, *Active learning by querying informative and representative examples*, In Advances in Neural Information Processing Systems 23, 2010.

[10] T. K. Huang, R. C. Weng, and C. J. Lin, *Generalized Bradley-Terry models and multi-class probability estimates*, Journal of Machine Learning Research, 2006.

Table II
COMPARISON OF STRATEGIES USING ONE-VERSUS-ONE SVM WITH PROBABILITY ESTIMATION

dataset	starting cost	ending cost / area under curve				
		Min confidence	Min margin	Max expected cost	CW min margin	Random
Wine	7.71 ± 0.40	2.37 ± 0.19	2.34 ± 0.19	1.57 ± 0.13	**1.46 ± 0.11**	2.35 ± 0.10
		43.4495	43.0833	41.0985	**36.6338**	43.1978
Glass	26.85 ± 0.15	26.03 ± 0.13	26.02 ± 0.15	25.49 ± 0.13	**25.42 ± 0.13**	25.57 ± 0.14
		264.3530	264.6724	261.4862	**261.0046**	261.7933
SVMguide4	12.62 ± 0.58	10.20 ± 0.46	9.07 ± 0.23	9.37 ± 0.27	**8.57 ± 0.16**	8.97 ± 0.19
		337.3102	316.2988	314.8814	**295.6578**	315.5661
Vehicle	8.97 ± 0.25	6.42 ± 0.21	6.34 ± 0.15	**5.70 ± 0.21**	6.63 ± 0.25	6.69 ± 0.21
		292.0806	**286.4367**	290.2878	303.7575	308.0229
Vowel	12.80 ± 0.46	10.18 ± 0.29	9.89 ± 0.23	10.06 ± 0.35	**8.91 ± 0.26**	9.21 ± 0.26
		564.4970	545.5735	559.8923	**519.3619**	530.1122
Segment	3.10 ± 0.09	2.14 ± 0.07	**1.88 ± 0.06**	1.98 ± 0.07	2.24 ± 0.06	2.56 ± 0.07
		303.5622	**280.4852**	296.4099	305.5039	333.2711
DNA	3.35 ± 0.08	2.37 ± 0.04	2.54 ± 0.06	**2.26 ± 0.04**	2.32 ± 0.05	2.70 ± 0.05
		461.9455	468.5468	**434.6247**	441.6270	481.9276
Satimage	3.67 ± 0.06	3.27 ± 0.05	3.55 ± 0.05	**2.99 ± 0.02**	3.11 ± 0.04	3.26 ± 0.04
		1067.1777	1150.7329	1091.4725	**1042.9394**	1101.3152
USPS	2.50 ± 0.03	1.55 ± 0.01	1.59 ± 0.03	**1.53 ± 0.02**	1.57 ± 0.02	1.85 ± 0.02
		896.1013	**876.7218**	902.9421	887.5385	961.1685
Pendigits	1.36 ± 0.02	**0.58 ± 0.01**	0.59 ± 0.00	0.64 ± 0.01	0.94 ± 0.02	0.88 ± 0.01
		470.1182	**466.1121**	502.4575	617.2564	591.6024

Table III
COMPARISON OF STRATEGIES USING RANDOM FOREST

dataset	starting cost	ending cost / area under curve				
		Min confidence	Min margin	Max expected cost	CW min margin	Random
Wine	6.13 ± 0.77	1.86 ± 0.23	1.84 ± 0.16	1.48 ± 0.16	**1.42 ± 0.17**	3.14 ± 0.42
		31.5048	33.2569	**26.7969**	29.3776	42.8301
Glass	24.50 ± 0.22	21.06 ± 0.19	21.33 ± 0.17	22.06 ± 0.20	**20.86 ± 0.19**	21.35 ± 0.17
		226.3436	226.9277	233.9861	**224.1297**	229.1157
SVMguide4	9.27 ± 0.27	8.66 ± 0.16	8.73 ± 0.20	8.29 ± 0.18	**8.22 ± 0.17**	8.31 ± 0.09
		262.0197	262.8498	258.8623	**256.8489**	261.7847
Vehicle	8.90 ± 0.27	5.74 ± 0.16	5.73 ± 0.15	**5.59 ± 0.21**	6.60 ± 0.21	6.44 ± 0.22
		286.2498	283.9035	**280.7986**	304.2760	304.2550
Vowel	8.76 ± 0.18	7.85 ± 0.14	7.67 ± 0.13	7.38 ± 0.13	**7.33 ± 0.11**	7.77 ± 0.12
		412.7921	408.7147	399.7172	**397.0205**	412.1777
Segment	5.54 ± 0.21	4.35 ± 0.09	3.65 ± 0.15	**3.33 ± 0.13**	3.35 ± 0.14	5.84 ± 0.12
		571.7806	538.7604	**464.7054**	492.0996	691.7833
DNA	7.15 ± 0.37	10.90 ± 0.00*	10.85 ± 0.04*	**3.42 ± 0.12**	5.00 ± 0.08	7.61 ± 0.35*
		1574.4144	1491.8496	**788.3401**	852.3826	1166.6458
Satimage	3.07 ± 0.03	2.67 ± 0.01	2.75 ± 0.02	2.67 ± 0.03	**2.58 ± 0.02**	2.74 ± 0.02
		901.4025	912.9325	910.4718	**896.5614**	920.2444
USPS	10.50 ± 0.16	10.56 ± 0.11*	9.44 ± 0.12	9.77 ± 0.12	**8.08 ± 0.12**	10.33 ± 0.11
		4934.6092	4567.2638	4343.3769	**4124.0049**	4798.9406
Pendigits	2.05 ± 0.02	2.59 ± 0.03*	2.56 ± 0.03*	1.75 ± 0.02	**1.60 ± 0.02**	1.96 ± 0.02
		1235.8396	1253.2795	1004.4075	**972.4080**	1098.9694

[11] P. Jain and A. Kapoor, *Active learning for large multi-class problems*, IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[12] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, *Multi-class active learning for image classification*, IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[13] T. Luo, K. Kramer, D. B. Goldgof, L. O. Hall, S. Samson, A. Remsen, and T. Hopkins, *Active Learning to Recognize Multiple Types of Plankton*, Journal of Machine Learning Research, 2005.

[14] H. T. Nguyen and A. Smeulders, *Active learning using pre-clustering*, In Proceedings of the 21st International Conference on Machine Learning, 2004.

[15] J. C. Platt, *Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods*, In Advances in Large Margin Classifiers, 1999.

[16] B. Settles, *Active Learning Literature Survey*, Technical report, University of Wisconsin Madison, 2009.

[17] S. Tong and D. Koller, *Support vector machine active learning with applications to text classification*, Journal of Machine Learning Research, 2002.

[18] H. H. Tu and H.T. Lin, *One-sided support vector regression for multiclass cost-sensitive classification*, In Proceedings of the 27th International Conference on Machine Learning, 2010.

[19] A. Vlachos, *A stopping criterion for active learning*, Computer Speech & Language, 2008.

[20] R. Yan and A. Hauptmann, *Multi-class active learning for video semantic feature extraction*, In IEEE International Conference on Multimedia and Expo, 2004.

[21] R. Yan, J. Yang, and A. Hauptmann, *Automatically labeling video data using multiclass active learning*, In Proceedings of the Ninth IEEE International Conference on Computer Vision, 2003.

[22] X. Zhu, *Semi-supervised learning with graphs*, PhD thesis, Carnegie Mellon University, 2005.