Stability-based Stopping Criterion for Active Learning

Wenquan Wang Wenbin Cai Ya Zhang*
Shanghai Key Laboratory of Multimedia Processing and Transmissions
Shanghai Jiao Tong University, Shanghai, China
E-mail: {wangwenquan, cai-wenbin, ya_zhang}@sjtu.edu.cn

***Abstract*—While active learning has drawn broad attention in recent years, there are relatively few studies on stopping criterion for active learning. We here propose a novel model stability based stopping criterion, which considers the potential of each unlabeled examples to change the model once added to the training set. The underlying motivation is that active learning should terminate when the model does not change much by adding remaining examples. Inspired by the widely used stochastic gradient update rule, we use the gradient of the loss at each candidate example to measure its capability to change the classifier. Under the model change rule, we stop active learning when the changing ability of all remaining unlabeled examples is less than a given threshold. We apply the stability-based stopping criterion to two popular classifiers: logistic regression and support vector machines (SVMs). It can be generalized to a wide spectrum of learning models. Substantial experimental results on various UCI benchmark data sets have demonstrated that the proposed approach outperforms state-of-art methods in most cases.**

***Keywords*-Stopping criterion, Active learning, Stability**

I. Introduction

Active learning has been well-motivated in many machine learning domains [1] [2] [3] where unlabeled examples are easy to collect but labeling cost is high. A typical active learning process iterate through the following three steps: 1) Build a base model with a small initial training set; 2) Use a certain sampling function to sample from a large unlabeled pool set and query their labels; 3) Add them to the training set and retrain the model. This sample selection process is repeated until a certain stopping criterion is met.

In recent years, most researchers have focused on designing the sampling function such as uncertainty sampling [4] and query-by-committee [5]. However, a potentially important issue of the interactive learning process is the stopping criterion (SC), i.e., deciding when to stop active learning, which is a relatively less-researched area. In active learning, the model performance, which usually means prediction accuracy, increases gradually at the initial stages. With the active learning proceeds, the performance of the model starts to remain stable and cannot be improved considerably at the latter round of active learning. To avoid the huge annotation waste on the non-informative examples, it is highly desirable to define an appropriate stopping criterion to cease the learning. However, there are relatively few studies on active learning without having the test set [6], [7], [8].

In this paper, we propose a new stopping criterion for active learning with model stability, which considers the capability of each candidate examples to change the classifier. The main idea is that active learning should stop when the model cannot be changed too much even with more training examples. The intuition behind our stability-based method is that the examples cannot change the current model are useless for active learning.

The model change is quantified as the difference between the current model parameters and the new parameters obtained with the accumulated training set. Inspired by the stochastic gradient update rule, where the model parameters are updated repeatedly according to the negative (or positive) gradient of the objective function, we use the gradient of the loss function at each candidate example to measure its capability to change the model. Under the model change principle, we stop active learning when the changing ability of all remaining unlabeled examples is less than a given threshold, i.e., the model is stable. In this study, we apply the stability-based stopping criterion to two popular classifiers: logistic regression and SVMs. It can be generalized to other base learners such as linear regression. Extensive experimental results on various UCI benchmark data sets have demonstrated that the proposed approach outperforms state-of-art methods in most cases.

The main contributions of this paper are as follows. First, We propose a new stopping criterion (SC) framework with model stability, which considers the capability of each candidate example to change the classifier. It can be applied to a wide spectrum of learners. Secondly, under the stability-based principle, we apply it to two popular classifiers: logistic regression and SVMs. Substantial empirical results demonstrate the effectiveness of the proposed methods.

II. Related Work

In this section, we summarize existing studies on stopping criterion for active learning.

1) **Performance-based SC**: The general idea behind this strategy is intuitive and natural, i.e., active learning should stop when a desired performance threshold is reached [9]. However, the limitation is that an extra test set is needed to evaluate the model's performances, involving an extra cost of labeling.

2) **Gradient-based SC**: The underlying motivation here is to stop active learning when more examples do not contribute more information, e.g., the model has reached maximum performance or the uncertainty cannot be decreased further. Under this principle, Laws and Schutze [10] presented a novel stopping criterion based on gradient decrease with application to named entity recognition. But, it is non-trivial to accurately estimate the performance without having a relatively large test set.

3) **Confidence-based SC**: The intuition behind this strategy is that active learning should be stopped when the classifier has enough confidence on its classification w.r.t. the remaining unlabeled data. Zhu et al. [8] proposed several alternatives to estimate the confidence. The *Max-confidence* strategy uses the uncertainty of the most selected instances as criteria for stopping active learning. The *Min-error* method stops active learning when the accuracy of the current classifier is larger than a given threshold. The *Overall-uncertainty* approach stops if the overall uncertainty value w.r.t. all remaining unlabeled examples is less than a predefined threshold. The *Classification-change* strategy stops active learning when the prediction does not change for the remaining unlabeled examples between two consecutive learning cycles. However, these confidence-based stopping criterions are specific to a set of probabilistic learning models, which limits their applicability.

Besides the above three frameworks, Schohn and Cohn *et al*. [6] presented a specific method for SVMs model, which suggested that active learning should stop when there are no unlabeled data points lying in the margin.

III. The Framework of Stability-based Stopping Criterion

In machine learning, the ultimate goal is to learn a classifier with good generalization performance on the future unseen data. We believe that the model stability is a reasonable indicator for stopping the active learning iteration with the following reasons. First, the model's generalization capability on the test set is changed if and only if the model is changed, and thus the examples that cannot change the model is actually useless for active learning. Secondly, as discussed before, active learning should stop when the performance remains stable, indicating that the remaining unlabeled examples cannot change the current classifier considerably. Hence, the model stability based stopping criterion (denoted as SC_{MS} hereafter) can be formulated as:

$$SC_{\text{MS}} = \begin{cases} 1, & ||\theta - \theta^{+}|| < \lambda,\ \forall x^{+} \in \mathcal{U}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where λ is a predefined constant. θ and θ^{+} denote the current model parameter and the updated parameter learned from the expanded training set, respectively. $\mathcal{U}$ stands for the unlabeled pool set. The active learning process ceases only if stopping criterion SC_{MS} is equal to 1. The main problem here is how to calculate the parameter change (i.e., $||\theta - \theta^{+}||$). In the following, we present our method with stochastic gradient rule.

We start with the well-known Empirical Risk Minimization (ERM) principle. To minimizes the empirical error, a widely used search method is stochastic gradient update rule, which updates the parameter θ repeatedly according to the negative gradient of the loss w.r.t. each training example:

$$\theta \leftarrow \theta - \alpha \frac{\partial \mathcal{L}_{\theta}(x_i)}{\partial \theta}, \qquad i = 1, 2, ..., n, \tag{2}$$

where α is the learning rate.

According to this rule, we approximate the parameter change as the derivative of the loss at the candidate example (x^{+}, y^{+}):

$$||\triangle\theta|| = ||\theta - \theta^{+}|| = ||\alpha \frac{\partial \mathcal{L}_{\theta}(x^{+})}{\partial \theta}||. \tag{3}$$

Putting together (1) and (3), SC_{MS} can be formulated as:

$$SC_{\text{MS}} = \begin{cases} 1, & ||\partial \mathcal{L}_{\theta}(x^{+})/\partial\theta|| < \lambda,\ \forall x^{+}, \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

In the above, we suppose that the learning rate α is identical for each candidate example.

In practice, because the true class label y^{+} of the example x^{+} is not known in advance, we are not able to calculate the derivative inside (4) directly. Instead, we use the expectation calculation over all possible labels $y^{+} \in Y$ to estimate the true parameter change. Our final SC_{MS} can be expressed as:

$$SC_{\text{MS}} = \begin{cases} 1, & \mathbb{E}_{y^{+}}||\partial \mathcal{L}_{\theta}(x^{+})/\partial\theta|| < \lambda,\ \forall x^{+}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

where

$$\mathbb{E}_{y^{+}}||\frac{\partial \mathcal{L}_{\theta}(x^{+})}{\partial \theta}|| = \sum_{y^{+} \in Y} p(y^{+}|x^{+})||\frac{\partial \mathcal{L}_{\theta}(x^{+})}{\partial \theta}||, \tag{6}$$

and $p(y^{+}|x^{+})$ is the conditional probability of label y^{+} given the example x^{+} estimated by the current model.

IV. Stability-based Stopping Criterion for Classification

In this section, we apply the proposed framework to classification tasks. We apply SC_{MS} to logistic regression and SVMs, two of the most popular classification methods. While we focus on the binary discrimination problems, it can be generalized to multi-class problems.

A. Stopping Criterion for Logistic Regression

The logistic regression model can be formulated as:

$$f(x) = \frac{1}{1 + e^{-\theta^{\text{T}} x}}, \tag{7}$$

where the class labels are represented as $Y = \{0, 1\}$ and θ is the parameter vector characterizing the model. The model

is learned by maximizing the log likelihood on a training set $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$:

$$\mathcal{L}_\theta(\mathcal{D}^+) = \sum_{i=1}^{n} y_i \log f(x_i) + (1 - y_i)\log(1 - f(x_i)). \quad (8)$$

Assume a candidate example x^+ with a given label y^+ is added to the training set. The log likelihood on the expanded training set $\mathcal{D}^+ = \mathcal{D} \cup (x^+, y^+)$ then becomes:

$$\begin{aligned}\mathcal{L}_\theta(\mathcal{D}^+) = &\sum_{i=1}^{n} y_i \log f(x_i) + (1 - y_i)\log(1 - f(x_i)) \\ &+ y^+ \log f(x^+) + (1 - y^+)\log(1 - f(x^+)).\end{aligned} \quad (9)$$

The derivative of the log likelihood $\mathcal{L}_\theta(x^+)$ at the candidate example (x^+, y^+) is calculated as:

$$\frac{\partial \mathcal{L}_\theta(x^+)}{\partial \theta} = (y^+ - f(x^+))x^+. \quad (10)$$

Note that the gradient ascent update rule is used here because we are maximizing rather than minimizing the objective function.

Since the true class label y^+ is unknown before querying, we employ bootstrap to create an ensemble $\mathcal{B}(K)$ to estimate the prediction distribution $y^+ \in \{y_1, y_2, ..., y_K\}$, and then use the expectation to approximate the true parameter change. The relationship between bootstrap and prediction distribution is formulated as [11]:

$$\mathbb{E}_{y^+} || \frac{\partial \mathcal{L}_\theta(x^+)}{\partial \theta} || = \frac{1}{K} \sum_{k=1}^{K} ||(y_k^+ - f(x^+))x^+||. \quad (11)$$

SC_{MS} for logistic regression can then be expressed as:

$$SC_{\text{MS}} = \begin{cases} 1, & \frac{1}{K}\sum_{k=1}^{K} ||(y_k^+ - f(x^+))x^+|| < \lambda, \ \forall x^+, \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

An interpretation behind SC_{MS} is explained as follows. The parameter change is proportional to the error term $(y^+ - f(x^+))$. Hence, if the classifier is good enough to accurately predict the examples, the changing ability of data instances will be small, indicating that the model is stable.

B. Stopping Criterion for Support Vector Machines

Support vector machines [12], with its superior properties of excellent generalization performance, robustness to the noise, and ability to handle high dimensional data, play a significant role in the machine learning community. For SVMs, the class labels are represented as $Y = \{-1, 1\}$. The linear SVM model is represented by a hyperplane:

$$f(x) = w^{\text{T}}x + b = 0, \quad (13)$$

where w is the weight vector parameterizing the classifier. For simplicity, we omit the bias term b throughout this study.

We consider the update rule in active learning. If a candidate point x^+ is incorporated into the training set with a given label y^+, the objective function on the enlarged training set $\mathcal{D}^+ = \mathcal{D} \cup (x^+, y^+)$ then becomes:

$$\min_w \ \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n}[1 - y_i w^{\text{T}} x_i]_+ + [1 - y^+ w^{\text{T}} x^+]_+. \quad (14)$$

We estimate the effect of adding the new candidate point on the training loss to approximate the parameter change, and hence the derivative of the loss at the example (x^+, y^+) is:

$$\frac{\partial \mathcal{L}_w(x^+)}{\partial w} = \begin{cases} -y^+ x^+, & \text{if } y^+ w^{\text{T}} x^+ < 1, \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

It shows that the SVM classifier updates its weight solely on those examples that satisfy the inequality $y^+ w^{\text{T}} x^+ < 1$, which is straightforward.

Because the true label $y^+ \in \{1, -1\}$ of the example x^+ is unknown in advance, we therefore rewrite the inequality constraint $y^+ w^{\text{T}} x^+ < 1$ as $|w^{\text{T}} x^+| < 1$. Meanwhile, we take the expectation over each possible class labels $y^+ \in \{1, -1\}$ to approximate the true parameter change:

$$\begin{aligned}\mathbb{E}_{y^+} || \frac{\partial \mathcal{L}_w(x^+)}{\partial w} || &= \sum_{y^+ \in Y} p(y^+|x^+) || - y^+ x^+ || \\ &= \sum_{y^+ \in Y} p(y^+|x^+) ||x^+|| \\ &= ||x^+||. \end{aligned} \quad (16)$$

Therefore, the SC_{MS} for SVM can be reformulated as:

$$SC_{\text{MS}} = \begin{cases} 1, & ||x^+|| < \lambda, \ \forall x^+, \\ 0, & \text{otherwise.} \end{cases} \quad (17)$$

A nice interpretation is that the examples within the margin are the ones having the ability to change SVMs, i.e., $|w^{\text{T}} x^+| < 1$. Thus, a SVM classifier is sufficiently stable when there are no data examples lying in the margin, thereby resulting in a good generalization performance. Also, the SC_{MS} for SVM strategy can be regarded as a general case of previous work [6], which stops active learning when there are no unlabeled points within the margin.

V. Experiments

A. Data sets and Experimental Settings

To validate the effectiveness of the proposed strategy, we use eight benchmark data sets of various domains from UCI machine learning repository[1]: *Biodeg*, *Ionosphere*, *WDBC*, *Parkinsons*, *Letter*. Since *Letter* is a multi-class data set, we select four pairs of letters (i.e., *D-vs-P*, *E-vs-F*, *M-vs-N*, *U-vs-V*) that are relatively difficult to distinguish, and build a binary-class data set for each pair. Table I presents the information of the eight binary-class data sets.

We randomly divide each data set into three parts: the initial labeled training set (denoted as $\mathcal{D}$), the unlabeled pool

[1] http://archive.ics.uci.edu/ml/

Table I
THE INFORMATION OF EIGHT DATA SETS FROM UCI.

Data sets	# Examples	# Features	Class distribution
Biodeg	1055	41	356/699
Ionosphere	351	34	225/126
WDBC	569	30	357/212
Parkinsons	195	22	147/48
E-vs-F	1543	16	768/755
D-vs-P	1608	16	805/803
M-vs-N	1575	16	792/783
U-vs-V	1577	16	813/764

set (denoted as $\mathcal{U}$), and the test set (denoted as $\mathcal{T}$). We use the base labeled set $\mathcal{D}$ as the small labeled data set to train the initial models. The pool set $\mathcal{U}$ is used as a large size unlabeled data set to select the most informative examples, and the separate test set $\mathcal{T}$ is used to evaluate different SC methods. Further, we randomly split each data set as: $\mathcal{D}$(5%) + $\mathcal{U}$(75%) + $\mathcal{T}$(20%). The features are normalized with the function below:

$$f_{(i,j)}^{\text{Norm}} = \frac{f_{(i,j)} - min_{i\in n}\{f_{(i,j)}\}}{max_{i\in n}\{f_{(i,j)}\} - min_{i\in n}\{f_{(i,j)}\}}, \tag{18}$$

where n denotes the number of examples in each data set, and $f_{(i,j)}$ is the j-th feature from the i-th example.

Two learning models are used as the base learner: logistic regression and SVM. The active learning algorithm used here is uncertainty sampling, which is the most widely employed sample selection method. For logistic regression, the examples with posterior probabilities closest to 0.5 [13] are chosen. For SVMs, the examples closet to the separating hyperplane are queried [14]. 1% instances are selected in each iteration. We test the method with different threshold values: $\lambda \in \{0.5, 1, 2\}$.

B. Comparison Methods

To test effectiveness of our model stability based stopping criterion (MS), we compare it with the following three state-of-art competitors: 1) max-confidence (MC), 2) min-error (ME), and 3) overall-uncertainty (OU). As suggested in [8], the initial threshold for these competitors, i.e., MC, ME, and OU, are set to be 0.5, 0.95, and 0.05, respectively. The threshold update algorithm [8] is used to refine the initial thresholds.

Because these methods are specific to probabilistic models, we utilize the sigmoid function to transform SVMs outputs to posterior probabilities as suggested by [15]:

$$p(y|x) = \frac{1}{1 + \exp(w^{\mathrm{T}}x)}. \tag{19}$$

C. Evaluation Metrics

In active learning scenarios, when the model firstly reaches the highest performance, it is suggested that the active learning process can stop. We refer to this checking point as the Best Stopping Point (BST) for stopping active learning. In order to accurately achieve the BST, we further perform the sequential active learning, i.e., only a single example is added to the training set in each sampling cycle.

With BST, we use the following three metrics to measure the performance of each stopping criterion.

1) The difference between the BST point and the stopping point suggested by a certain stopping criterion.

$$\Delta_{point} = |\psi_{BST} - \psi_{sc}|, \tag{20}$$

where ψ_{BST} and ψ_{sc} denote the percentage of training set when active learning stops at the BST and the point suggested by a stopping criterion, respectively. The smaller Δ_{point}, the better stopping criterion.

2) The difference between the highest performance (accuracy) and the performance obtained at a certain stopping point with a stopping criterion:

$$\Delta_{Acc} = Acc_{BST} - Acc_{sc}, \tag{21}$$

where Acc_{BST} and Acc_{sc} are the classifier's accuracy performance obtained at the BST and the stopping point of a stopping criterion, respectively. The smaller Δ_{Acc}, the better stopping criterion.

3) To further evaluate the effectiveness of stopping criteria, we define a third metric by simultaneously considering the number of selected instances and the performance improvement:

$$\text{Average improvement} = \frac{Acc_i - Acc_{sc}}{\log(\text{\# selected examples})}, \tag{22}$$

where Acc_i denotes the accuracy obtained with the initial labeled set. The reason of using the logarithm function log(.) lies in the fact that the model's performance usually increases in a logarithm-like shape with the number of selected examples. The larger average improvement, the better stopping criterion.

D. Comparison Results and Discussions

To gain an overall insight of these four stopping criteria, we first present the results of ψ_{sc} (the percentage of training set when active learning stops) and Acc_{sc} (the accuracy obtained at the stopping point). Table II and Table III present the results for logistic regression and SVMs, respectively. The values having the best performance are highlighted. As shown in these tables, we see that the number of selected examples monotonically decrease with the value of the threshold λ, which agrees with the MS function defined in Eq.(1).

Table IV
THE NUMBER OF DATA SETS CONSIDERING Δ_{point}, Δ_{Acc}

	learn model	MS	MC	ME	OU
Δ_{point}	Logistic	**6**	2	5	3
	SVMs	**6**	3	2	**6**
Δ_{Acc}	Logistic	**4**	2	1	1
	SVMs	**4**	0	1	3

Table II
EXPERIMENTAL RESULTS ON UCI BENCHMARKS FOR LOGISTIC REGRESSION.

Dataset		BST	MS($\lambda = 2$)	MS($\lambda = 1$)	MS($\lambda = 0.5$)	MC	ME	OU
Biodeg	ψ_{sc}	19.84%	**6.06%**	12.36%	25.27%	44.50%	52.59%	52.59%
	Acc_{sc}	89.15%	**88.63%**	**88.63%**	87.20%	86.26%	86.73%	86.73%
Ionosphere	ψ_{sc}	63.59%	**7.12%**	8.26%	9.40%	37.50%	50.00%	27.23%
	Acc_{sc}	87.14%	74.29%	77.14%	78.57%	78.57%	**81.43%**	77.14%
WDBC	ψ_{sc}	61.41%	**6.15%**	13.53%	13.53%	29.51%	17.75%	34.62%
	Acc_{sc}	96.46%	85.84%	92.92%	92.92%	93.81%	**94.69%**	93.81%
Parkinsons	ψ_{sc}	11.79%	**5.64%**	9.74%	11.79%	33.57%	5.64%	31.28%
	Acc_{sc}	92.31%	71.79%	84.62%	**92.31%**	**92.31%**	71.79%	**92.31%**
D-vs-P	ψ_{sc}	8.54%	**7.96%**	9.40%	9.40%	10.31%	20.92%	10.31%
	Acc_{sc}	99.07%	**98.76%**	**98.76%**	**98.76%**	98.45%	**98.76%**	98.45%
E-vs-F	ψ_{sc}	37.84%	**5.84%**	6.85%	11.80%	15.68%	11.80%	11.80%
	Acc_{sc}	99.35%	85.79%	95.79%	**99.03%**	**99.03%**	**99.03%**	**99.03%**
M-vs-N	ψ_{sc}	10.41%	**6.03%**	8.06%	8.06%	15.17%	15.17%	17.21%
	Acc_{sc}	97.46%	95.87%	**97.14%**	**97.14%**	96.83%	96.83%	96.83%
U-vs-V	ψ_{sc}	10.41%	**6.02%**	7.04%	14.14%	10.30%	26.21%	26.21%
	Acc_{sc}	99.68%	97.15%	97.78%	**99.37%**	98.42%	**99.37%**	**99.37%**

Table III
EXPERIMENTAL RESULTS ON UCI BENCHMARKS FOR SVMS.

Dataset		BST	MS($\lambda = 2$)	MS($\lambda = 1$)	MS($\lambda = 0.5$)	MC	ME	OU
Biodeg	ψ_{sc}	18.67%	**6.07%**	37.35%	38.39%	62.37%	32.13%	45.69%
	Acc_{sc}	88.63%	84.36%	**86.73%**	**86.73%**	86.26%	85.31%	85.78%
Ionosphere	ψ_{sc}	56.13%	29.91%	29.91%	36.61%	69.80%	**18.52%**	50.43%
	Acc_{sc}	90.00%	**88.57%**	**88.57%**	**88.57%**	**88.57%**	80.00%	**88.57%**
WDBC	ψ_{sc}	6.33%	**6.15%**	19.86%	19.86%	17.75%	14.59%	13.53%
	Acc_{sc}	96.46%	**95.58%**	94.69%	94.69%	94.69%	94.69%	95.58%
Parkinsons	ψ_{sc}	14.36%	**5.64%**	25.13%	26.15%	72.31%	6.67%	29.23%
	Acc_{sc}	92.31%	74.36%	**92.31%**	**92.31%**	87.18%	74.36%	**92.31%**
D-vs-P	ψ_{sc}	8.33%	10.95%	10.95%	16.92%	16.92%	12.94%	**8.96%**
	Acc_{sc}	99.07%	**98.45%**	**98.45%**	97.83%	97.83%	97.83%	**98.45%**
E-vs-F	ψ_{sc}	9.40%	**13.74%**	18.60%	18.60%	20.54%	16.86%	15.69%
	Acc_{sc}	98.71%	**98.38%**	97.73%	97.73%	97.73%	97.73%	**98.38%**
M-vs-N	ψ_{sc}	8.44%	**11.11%**	20.25%	20.25%	26.35%	15.17%	13.14%
	Acc_{sc}	98.10%	**97.46%**	96.51%	96.51%	96.83%	97.46%	**97.46%**
U-vs-V	ψ_{sc}	7.74%	15.16%	15.16%	15.16%	15.16%	12.11%	**7.04%**
	Acc_{sc}	99.37%	**99.37%**	**99.37%**	**99.37%**	**99.37%**	**99.37%**	98.10%

In the following, we compare our MS method against these three competitors with evaluation metrics defined above (i.e., Δ_{point}, Δ_{Acc}, and Average improvement). For simplicity, we present the results with the $\lambda = 0.5$ for logistic regression and $\lambda = 2$ for SVM, which have shown to work well in our empirical studies.

1) Δ_{point} and Δ_{Acc}: Similar to previous work [8], we count the number of data sets on which the stopping criterion reaches the best (smallest) value of Δ_{point} and Δ_{Acc}. Table IV shows the results performed on the eight UCI data sets. MS performs better on most data sets for both logistic regression and SVM either in terms of Δ_{point} or Δ_{Acc}, indicating the effectiveness of this method. MC performs poorly in terms of Δ_{Acc} and Δ_{point}, indicating that OU tends to stop active learning at the latter iteration with the goal of achieving a high performance. The performance of ME is inconsistent. It works well on some data sets, but performs poorly on the others. This is likely due to the reason that the expected error estimation is highly correlated with the characteristics of data sets. OU works well in SVMs but poorly in logistic regression, which shows that OU more likely to be a model-specific method that works for margin-based models.

2) Average improvement: To better test the effectiveness of the proposed MS strategy, we adopt the average improvement as the third metric by simultaneously considering the number of selected examples and the accuracy improvement. In order to make it clear and cross the data sets, we further normalize the average improvement in $[50\%, 100\%]$. Figure 2 and Figure 3 show the comparison results on these eight UCI benchmarks. MS is observed to perform the best among these four methods in most cases, demonstrating that MS stops active learning with high performance and less training set. Generally, OU works well for SVM but poorly for logistic regression, which indicates that the uncertainty-based OU might be more appropriate for the margin-based classifier. MC and ME performs poorly most of times. A possible reason can be explained as follows. Both MC and ME put more weights to achieve the high performance, and hence a relatively large training set is required.

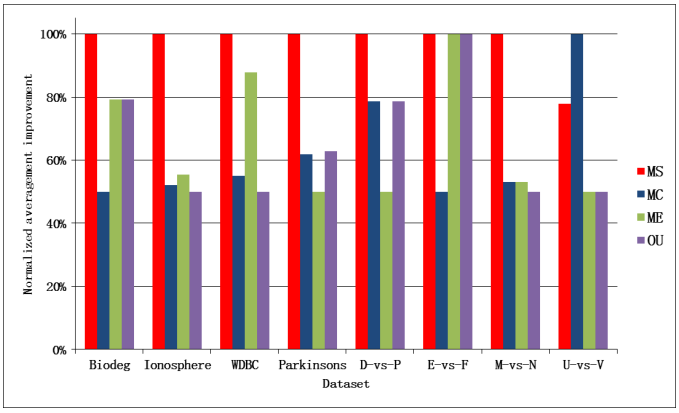

Figure 1. Normalized average improvement on UCI data sets in logistic regression.

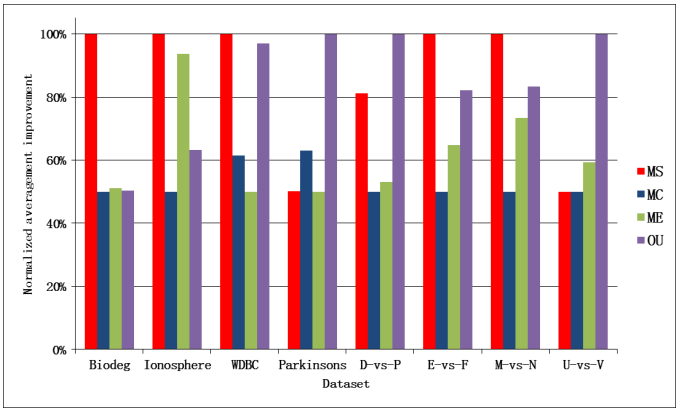

Figure 2. Normalized average improvement on UCI data sets in SVMs.

VI. Conclusions

In this paper, we propose a novel model stability based stopping criterion, which considers the potential capability of each unlabeled examples to change the model once added to the training set. Inspired by the widely used stochastic gradient update rule, we use the gradient of the loss at each candidate example to measure its capability to change the classifier. We apply the stability-based stopping criterion to two popular classifiers: logistic regression and support vector machines (SVMs). Substantial experimental results on various UCI benchmark data sets have demonstrated that the proposed approach outperforms state-of-art methods in most cases.

Acknowledgement

This research was supported by the High Technology Research and Development Program of China (2012AA011702), STCSM # 14511107500, and National Natural Science Foundation of China (No. 61003107 & No. 61221001).

References

[1] Tong, Simon, and Edward Chang. "Support vector machine active learning for image retrieval." *Proceedings of the ninth ACM international conference on Multimedia.* ACM, pp.107-118, 2001.

[2] Tang, Min, Xiaoqiang Luo, and Salim Roukos. "Active learning for statistical natural language parsing." *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 2002.

[3] Li, Lianghao, et al. "Multi-domain active learning for text classification." *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2012.

[4] D. Lewis and W. Gale. , "A sequential algorithm for training text classifiers." *ACM SIGIR Conference on Research and Development in Information Retrieval* , pp. 3-12. ACM/Springer, 1994.

[5] H.S. Seung, M. Opper, and H. Sompolinsky, "Query by committee." *ACMWorkshop on Computational Learning Theory*, pp. 287-294, 1992.

[6] Schohn, G., Cohn, D., "Less is more: active learning with support vector machines." *Proceedings of the Seventeenth International Conference on Machine Learning.* Morgan Kaufmann, San Francisco, CA, pp. 839-846,2000.

[7] Tomanek, K., Wermter, J., Hahn, U., "An approach to text corpus construction which cuts annotation costs and maintains reusability of annotated data." *Proceedings of the Joint Meeting of the Conference on Empirical Methods on Natural Language Processing and the Conference on Natural Language Learning*, Prague, Czech Republic, pp.2007.

[8] Zhu J, Wang H, Hovy E, "Confidence-based stopping criteria for active learning for data annotation[J]." *ACM Transactions on Speech and Language Processing (TSLP)*, vol 6(3),pp. 3, 2010.

[9] Li, M., Sethi, I.K., "Confidence-based active learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 28 (8), pp. 1251-1261, 2006.

[10] Laws, F. and Schutze, H., "Stopping criteria for active learning of named entity recognition." In *Proceedings of the 22nd International Conference on Computational Linguistics.*, pp. 465-472,2008.

[11] Fushiki T. "Bootstrap prediction and Bayesian prediction under misspecified models[J]," *Bernoulli*, pp. 747-758, 2005.

[12] Burges, C.: A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 121-167, 1998.

[13] Andrew I. Schein, Lyle H. Ungar, "Active learning for logistic regression: an evaluation", *Machine Learning*, Volume 68, Issue 3, pp 235-265

[14] Tong S, Koller D., "Support vector machine active learning with applications to text classification[J]," *The Journal of Machine Learning Research*, vol 2, pp. 45-66, 2002.

[15] Platt, John C. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods." *Advances in large margin classifiers.* 1999.