Active Learning for Support Vector Machines with Maximum Model Change

Wenbin Cai[1], Ya Zhang[1], Siyuan Zhou[1], Wenquan Wang[1], Chris Ding[2], and Xiao Gu[1]

[1] Shanghai Key Laboratory of Multimedia Processing & Transmissions
Shanghai Jiao Tong University, Shanghai, China
{cai-wenbin,ya_zhang,zhousiyuan,wangwenquan,gugu97}@sjtu.edu.cn
[2] University of Texas at Arlington, Arlington, Texas, USA
chqding@uta.edu

Abstract. Margin-based strategies and model change based strategies represent two important types of strategies for active learning. While margin-based strategies have been dominant for Support Vector Machines (SVMs), most methods are based on heuristics and lack a solid theoretical support. In this paper, we propose an active learning strategy for SVMs based on Maximum Model Change (MMC). The model change is defined as the difference between the current model parameters and the updated parameters obtained with the enlarged training set. Inspired by Stochastic Gradient Descent (SGD) update rule, we measure the change as the gradient of the loss at a candidate point. We analyze the convergence property of the proposed method, and show that the upper bound of label requests made by MMC is smaller than passive learning. Moreover, we connect the proposed MMC algorithm with the widely used *simple margin* method in order to provide a theoretical justification for margin-based strategies. Extensive experimental results on various benchmark data sets from UCI machine learning repository have demonstrated the effectiveness and efficiency of the proposed method.

Keywords: Active Learning, Maximum Model Change, SVMs.

1 Introduction

In supervised learning, a large amount of labeled data is usually required to obtain a high quality model. A widely used method for data collection is passive learning, where the training examples are randomly selected according to a certain underlying distribution and annotated by human editors. However, in many practical applications, there might not be sufficient labeled data examples due to the high cost associated with data annotation. To solve this problem, active learning aims at selectively labeling the most informative instances with the goal of maximizing the accuracy of the model trained. In a typical active learning framework, the learner iteratively chooses informative data examples from a large unlabeled set (denoted as pool $\mathcal{U}$) with a predefined sampling function,

T. Calders et al. (Eds.): ECML PKDD 2014, Part I, LNCS 8724, pp. 211–226, 2014.

and then labels them. This data sampling process is repeated until a certain performance expectation is achieved or a certain labeling budget is used up. Active learning is well-motivated in many supervised learning tasks where unlabeled data may be abundant but labeled data examples are expensive to obtain [8,9].

Support vector machines (SVMs), which have arisen from statistical learning theory, play a significant role in the machine learning community with solid mathematical and statistical foundation [11,12]. Due to many desired properties including excellent generalization performance, robustness to the noise, and capability to deal with high dimensional data, SVMs have been successfully applied to many learning applications. As a result, active learning for SVMs has recently drawn a great deal of attention. In previous studies, several active learning algorithms have been specifically proposed for SVMs [5,10,13,14,17,18]. Most of them are derived with the notion of margin, i.e. preferring the points located in the margin. For example, *simple margin* [18], the most widely adopted strategy for SVMs, selects the examples that are closest to the decision boundary as the most informative ones. Although the margin-based active learning heuristics are fairly straightforward and natural for SVMs, these popular approaches lack a solid theoretical justification, i.e. how can we guarantee that margin-based active sampling performs better than passive learning.

In this paper, we introduce a new interpretation for the margin-based active learning by bridging it with the idea of model change. In particular, we attempt to provide theoretical justifications for the margin-based methods. We consider the capability of examples to change the model, and accordingly propose a novel margin-based active learning strategy for SVMs called Maximum Model Change (MMC), which is to choose the examples leading to the maximal change to the current model. The change is quantified as the difference between the current model parameters and the new parameters obtained with the expanded training set. Inspired by the well-studied work on the Stochastic Gradient Descent (SGD) update rule [15,16,19], where the parameters are updated repeatedly according to the negative gradient of the objective function at each single training example, we use the gradient of the loss at a candidate example to approximate the model change. Under the model change principle, the instances lying in the margin are proven to be the ones having the capability to change the model. We further analyze the convergence property of the proposed MMC method, and show that 1) MMC is guaranteed to converge, and 2) the upper bound of label requests made by MMC is smaller than that of passive learning. We further connect MMC with simple margin to provide a uniform view to these two methods. The property holds for other well-known SVMs active learning methods as well. We validate our algorithm with various benchmark data sets from UCI machine learning repository. Extensive experimental results have demonstrated the effectiveness and efficiency of the proposed active learning approach.

The main contributions of this paper are summarized as follows.

– Focusing on SVMs as the base learner, we introduce a novel interpretation for margin-based active learning with model change, and propose a new sampling algorithm called Maximum Model Change (MMC).

- We theoretically analyze the convergence property of the proposed approach, and compare the sampling bound against passive learning.
- We connect MMC with the widely adopted simple margin heuristic in order to provide a uniform view to these two active learning methods.

The rest of this paper is structured as follows: Section 2 briefly reviews the related work. Our active learning approach for SVMs, Maximum Model Change (MMC), is presented in Section 3. Section 4 provides the theoretical justification of the convergence property for the proposed approach, and compare the sampling bound with that of passive learning. Section 5 explores the relationship between MMC and simple margin. Section 6 presents the experimental results. Finally, we conclude the paper in Section 7.

2 Related Work

The goal of active learning is to train a high quality model using as few labeled training set as possible, therefore minimizing the labeling cost. In this section, we first briefly review several general active learning strategies, and then summarize existing margin-based active learning methods for SVMs.

2.1 Active Learning

Various active learning strategies have been proposed in the literature. Here we briefly review the typical active learning strategies:

1. **Uncertainty Sampling (US)**: The US approach aims to choose the examples whose labels the current classifier is most uncertain about. This strategy is usually straightforward to implement for probabilistic models. Take binary classification as an instance, US aims to query the data point whose posterior probability is most close to 0.5 [22]. For multi-class classification problems, examples with the smallest margin between the first and second most probable class labels are selected [1].
2. **Query By Committee (QBC)**: The QBC strategy generates a committee of model members and select unlabeled instances about which the models disagree the most [4]. A popular function to quantify the disagreement is vote entropy. To efficiently generate the committee, popular ensemble learning methods, such as Bagging and Boosting, have been employed [2].
3. **Expected Error Reduction (EER)**: The EER strategy aims to minimize the generalization error of the model. Roy et al. [20] proposed an optimal active sampling method to choose the example that leads to the lowest generalization error on the future test set once labeled and added to the training set. The weakness is that the computational cost of this method is extremely high. Instead of choosing the example yielding the smallest generalization error, Nguyen et al. [7] suggested to query the instance that has the largest contribution to the current error. Cohn et al. [21] proposed a statistically optimal active learning approach, which aims to choose the examples minimizing the output variance to reduce the generalization error.

4. **Expected Model Change (EMC)**: This strategy is to select data points that are expected to incur a large model change once added to the training set. Settles et al. [23] proposed an algorithm for logistic regression, and the change is quantified as the gradient length of the objective function obtained by the enlarged training set. Donmez et al. [3] presented a sampling approach for ranking tasks, which measures the change as the difference between the current model and the additional model trained with the selected examples. Recently, Cai et al. [24] applied this strategy to regression tasks.

There are several other active learning strategies proposed. A comprehensive active learning survey can be found in [6].

2.2 Active Learning for SVMs

Support vector machines (SVMs), built on solid mathematical and statistical foundation, play an important role in supervised learning. Many active learning algorithms, especially margin-based algorithms, have been specifically proposed for SVMs. We summarize existing margin-based active learning for SVMs as follows:

1. **Simple Margin** [18]: The simple margin algorithm is one of the most widely adopted active learning strategy when employing SVMs as the base learner, which chooses the examples that are closest to the separating hyperplane.
2. **MaxMin Margin** [18]: This active learning method aims to select the data instances that equally split the version space once labeled and added to the training set.
3. **Ratio Margin** [18]: This sampling approach is an extension of MaxMin Margin by taking particular consideration of the shape of version space.
4. **Representative Sampling** [5]: This sampling algorithm selects the most representative points within the margin using the clustering techniques.
5. **Multi-criteria-based Sampling** [14]: This approach simultaneously considers multiple criteria for sampling, and queries the data examples that are both informative and representative.
6. **Diversity-based Sampling** [13]: This strategy extends the simple margin to batch mode active learning by incorporating a diversity measure, which is calculated by their angles, to enforce the selected points to be diverse.
7. **Confidence-based Sampling** [10]: This active sampling algorithm can be regarded as a variant of the simple margin, which measures the uncertainty value of each sample as its conditional error.

As listed above, a common feature among the margin-based active learning methods is that they tend to pick the data examples located in the margin. Although existing margin-based active learning strategies are quite straightforward for SVMs, one limitation is that they lack solid theoretical support. In the next sections, we propose a new active learning algorithm for SVMs, together with theoretical justification.

3 Maximum Model Change for SVMs

In this section, we first provide a brief introduction to SVMs, focusing on the model fitting with the Stochastic Gradient Descent (SGD) rule. Then, the details of the proposed active learning algorithm, Maximum Model Change (MMC), for SVMs are provided. Finally, we analyze the computational complexity of the proposed algorithm.

3.1 Training SVMs with Stochastic Gradient Descent

For simplicity, we concentrate on the binary classification problem in this paper. It is straightforward to generalize the proposed method to multi-class problems. Given a training set $\mathcal{L} = \{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $y_i \in \{1, -1\}$ is a class label, the separation hyperplane of linear SVM model is represented as:

$$f(x) = w^{\mathrm{T}} x + b = 0, \tag{1}$$

where w denotes the weight vector parameterizing the model. For simplicity, we omit the bias term b throughout this study, which is commonly used in practice. In fact, it is easy to employ the bias by padding extra dimension of all 1's.

Building a SVM classifier is to solve the following Quadratic Programming (QP) problem:

$$\begin{aligned} \min_{w} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i w^{\mathrm{T}} x_i \geq 1 - \xi_i, \\ & \xi_i \geq 0, \qquad i = 1, ..., n, \end{aligned} \tag{2}$$

where ξ_i is a slack variable. The above QP problem can be equivalently rewritten as an unconstrained problem by re-arranging the constraints and substituting the parameter C with $C = \frac{1}{\lambda}$ as follows:

$$\min_{w} \ \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n}[1 - y_i w^{\mathrm{T}} x_i]_+, \tag{3}$$

where the subscript indicates the positive part. The first term is the regularizer, and the second term represents the standard Hinge loss. More generally, the soft margin loss is adopted with a margin parameter $\gamma \geq 0$, which treats the margin as a variable [19]. Thus, the SVM optimization problem can be reformulated as:

$$\min_{w} \ \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n}[\gamma - y_i w^{\mathrm{T}} x_i]_+, \tag{4}$$

where the second term [...]+ is the so-called soft margin loss. When $\gamma = 1$, the second term is the Hinge loss.

To find w minimizing the objective function, a widely used search approach is Stochastic Gradient Descent (SGD), which updates the parameter w repeatedly according to the negative gradient of the objective function with respect to each training example:

$$w \leftarrow w - \alpha \frac{\partial \mathcal{O}_w(x_i)}{\partial w}, \quad i = 1, 2, ..., n, \tag{5}$$

where $\mathcal{O}_w(x_i)$ and α are the objective function and the learning rate, respectively. With the particular of objective function (4), the update rule can be written as:

$$w \leftarrow \begin{cases} (1 - \alpha\lambda)w + \alpha y_i x_i, & \text{if } y_i w^{\mathrm{T}} x_i < \gamma, \\ (1 - \alpha\lambda)w, & \text{otherwise.} \end{cases} \tag{6}$$

In the literatures, several SGD-based learning algorithms have been well studied for solving the SVMs optimization problems [15,19]. They share the same update rule (6) with different scheduling of the learning rate.

3.2 Model Change Computation

Here, we consider the SGD rule in the active learning cases. Suppose a candidate example x^+ is added to the training set with a given class label y^+, the objective function on the expanded training set $\mathcal{L}^+ = \mathcal{L} \cup (x^+, y^+)$ then becomes:

$$\min_w \ \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{n} [\gamma - y_i w^{\mathrm{T}} x_i]_+ + \underbrace{[\gamma - y^+ w^{\mathrm{T}} x^+]_+}_{:=\ell_w(x^+)}. \tag{7}$$

As a result, the parameter w is changed due to the inclusion of the new example (x^+, y^+). We estimate the effect of adding the new point on the training loss to approximate the change, and hence the model change can be approximated with the gradient of the loss function at the example (x^+, y^+):

$$\mathcal{C}_w(x^+) = \triangle w \approx \alpha \frac{\partial \ell_w(x^+)}{\partial w}. \tag{8}$$

The derivative of the loss at the candidate point (x^+, y^+) is calculated as:

$$\frac{\partial \ell_w(x^+)}{\partial w} = \begin{cases} -y^+ x^+, & \text{if } y^+ w^{\mathrm{T}} x^+ < \gamma, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Clearly, the model updates its weight based on solely those examples that satisfy the inequality $y^+ w^{\mathrm{T}} x^+ < \gamma$, which is straightforward for SVMs.

The goal of MMC is to query the example that maximally changes the current model. According to (8) and (9), only the set $\Psi = \{x : y^+ w^{\mathrm{T}} x^+ < \gamma\} \subseteq \mathcal{U}$ has the ability to change the model, and hence only this set needs to be considered in active learning. The sampling criteria can be expressed as:

$$x^*_{\mathrm{MMC}} = \arg\max_{x^+ \in \Psi} ||\mathcal{C}_w(x^+)||. \tag{10}$$

Algorithm 1. MMC active learning for SVMs

Input: The labeled data set $\mathcal{L}=\{(x_i, y_i)\}_{i=1}^{n}$, the unlabeled pool set $\mathcal{U}$, the parameter γ, the SVM classifier $f(x)$ trained with $\mathcal{L}$.

1: **for** each x^+ in $\mathcal{U}$ **do**
2: **if** $|w^{\mathrm{T}}x^+| < \gamma$ **then**
3: $\mathbb{E}_{y^+}\{||\mathcal{C}_w(x^+)||\} \leftarrow ||x^+||$.
4: **end if**
5: **end for**

Output: $x^* \leftarrow \arg\max_{x^+} \mathbb{E}_{y^+}\{||\mathcal{C}_w(x^+)||\}$

In practice, the true label y^+ of the example x^+ is unknown in advance. With $y \in \{1, -1\}$, we have

$$\Omega = \{x : |w^{\mathrm{T}}x^+| < \gamma\} \subseteq \{x : y^+ w^{\mathrm{T}}x^+ < \gamma,\ y \in \{1, -1\}\}.$$

We hence rewrite the inequality constraint $y^+ w^{\mathrm{T}}x^+ < \gamma$ as $|w^{\mathrm{T}}x^+| < \gamma$. Meanwhile, we take the expected model change over each possible class labels $y^+ \in Y = \{1, -1\}$ to approximate the true change. Suppose that the learning rate α for each candidate point is identical, the final sampling criteria can be reformulated as:

$$\begin{aligned} x^*_{\mathrm{MMC}} =& \arg\max_{x^+\in\Omega} \mathbb{E}_{y^+}\{||\mathcal{C}_w(x^+)||\} \\ =& \arg\max_{x^+\in\Omega} \sum_{y^+\in Y} \hat{P}(y^+|x^+)||-y^+x^+|| \\ =& \arg\max_{x^+\in\Omega} \sum_{y^+\in Y} \hat{P}(y^+|x^+)||x^+|| \\ =& \arg\max_{x^+\in\Omega} ||x^+||, \end{aligned} \tag{11}$$

where $\hat{P}(y^+|x^+)$ represents the conditional probability of label y^+ given example x^+ estimated by the current classifier. The last step above follows from the fact that $\hat{P}(y^+ = +1|x^+) + \hat{P}(y^+ = -1|x^+) = 1$. An intuitive explanation for MMC is that the data examples maximally changing the current classifier are expected to result in faster convergence to the optimal model. The corresponding pseudo-code is given in Algorithm 1.. Based on the above derivation, MMC can be deemed as a margin-based active learning strategy as well because it shares the common feature of preferring examples within the margin, i.e. $\{x : |w^{\mathrm{T}}x^+| < \gamma\}$.

3.3 Computational Complexity

Assume that there are n labeled examples in the training set, and m unlabeled instances in the pool set. There are three main operations in the MMC method: SVM training, sample filtering, and sample selection.

SVM training typically needs $O(n^2)$ calculation. For the sample filtering, the main operation is to calculate the inner product, which has a time complexity of

$O(d)$. Therefore, the total time complexity is $O(md)$ at this step. For the sample selection, most time is spent on computing the norm with a complexity of $O(d)$, and hence the total time complexity is $O(kd)$) if there are k eligible examples. Summing up, the total time complexity for MMC is $O(n^2 + (m+k)d)$, which is promising for real-world tasks.

4 Theoretical Analysis

The goal of a learning model is to minimize the generalization error on future data. Clearly, the generalization error is changed if and only if the model is changed. Thus, active learning only needs to select the samples that change the current model, which is support vectors for SVMs. A nice feature of SVMs is that support vectors usually represent a tiny portion of the training data.

We have shown that points within the margin are the ones having the ability to change the current model. In this section, we attempt to provide a theoretical backup behind our strategy by analyzing the convergence property. Assume that $\{\exists\epsilon,\, x : y^+ w^{\mathrm{T}} x^+ \leq \gamma - \epsilon\} = \{x : y^+ w^{\mathrm{T}} x^+ < \gamma\}$. Since the scaling factors is to scale the derived bound by some fixed constant, which does not affect the convergence property, for clarity, we drop the scaling factors in the update rule:

$$w \leftarrow \begin{cases} w + y_i x_i, & \text{if } y_i w^{\mathrm{T}} x_i < \gamma, \\ w, & \text{otherwise.} \end{cases} \tag{12}$$

Theorem. (Convergence property) *Suppose that* $||x_j|| \leq R$ *for all* $x_j \in \mathcal{L} \cup \mathcal{U}$. *Let the current solution be* w^c, *and further suppose that there exists an optimal solution* w^* *such that* $y_j (w^*)^T x_j \geq \gamma$ *for all examples* x_j. *Let* $||w^c|| = M$ *and* $||w^*|| = N$. *Then, the total number of label requests* $\mathcal{A}$ *made by MMC is at most*

$$O\left(\frac{N}{\gamma}\left(M + N + \frac{N(R^2 - \epsilon)}{\gamma}\right)\right).$$

Proof. The proposed MMC algorithm chooses the data points only that change the current model, which implies that

$$y^+ w^{\mathrm{T}} x^+ < \gamma \Leftrightarrow y^+ w^{\mathrm{T}} x^+ \leq \gamma - \epsilon. \tag{13}$$

According to the SGD update rule in Eq. (12), we have

$$w^{(t+1)} \leftarrow w^{(t)} + y^+ x^+, \qquad t = 1, 2, ..., \mathcal{A}. \tag{14}$$

where $w^{(t=1)}$ stands for the current solution, i.e. $w^{(t=1)} = w^c$. According to the above update rule in Eq. (14), we get:

$$\begin{aligned} ||w^{(t+1)}||^2 =& ||w^{(t)} + y^+ x^+||^2 \\ =& ||w^{(t)}||^2 + ||x^+||^2 + 2y^+ (w^{(t)})^{\mathrm{T}} x^+ \\ \leq& ||w^{(t)}||^2 + ||x^+||^2 + 2(\gamma - \epsilon) \\ \leq& ||w^{(t)}||^2 + R^2 + 2(\gamma - \epsilon). \end{aligned} \tag{15}$$

and

$$\begin{aligned}(w^{(t+1)})^{\mathrm{T}}w^* =&(w^t)^{\mathrm{T}}w^* + y^+(x^+)^{\mathrm{T}}w^*\\ \geq&(w^t)^{\mathrm{T}}w^* + \gamma. \end{aligned} \tag{16}$$

Through iterative deduction of the above two equations, we have

$$\begin{aligned}||w^{(\mathcal{A}+1)}||^2 \leq&||w^c||^2 + \mathcal{A}R^2 + 2\mathcal{A}(\gamma - \epsilon)\\ =&M^2 + \mathcal{A}R^2 + 2\mathcal{A}(\gamma - \epsilon). \end{aligned} \tag{17}$$

and

$$(w^{(\mathcal{A}+1)})^{\mathrm{T}}w^* \geq (w^c)^{\mathrm{T}}w^* + \mathcal{A}\gamma. \tag{18}$$

Because $(w^c)^{\mathrm{T}}w^* = ||w^c|| \cdot ||w^*|| \cos\phi$, where ϕ is the angle between w^c and w^*, we have:

$$\begin{aligned}(w^{(\mathcal{A}+1)})^{\mathrm{T}}w^* \geq&\mathcal{A}\gamma - ||w^c|| \cdot ||w^*||\\ =&\mathcal{A}\gamma - MN. \end{aligned} \tag{19}$$

According to the Cauchy-Schwartz inequality, we see that

$$(w^{(\mathcal{A}+1)})^{\mathrm{T}}w^* \leq ||(w^{(\mathcal{A}+1)})|| \cdot ||w^*||. \tag{20}$$

Putting together Eq. (17) and Eq. (19) we get

$$\mathcal{A}\gamma - MN \leq \sqrt{M^2 + \mathcal{A}R^2 + 2\mathcal{A}(\gamma - \epsilon)}N. \tag{21}$$

Hence, we get

$$\begin{aligned}\mathcal{A} \leq&\frac{N}{\gamma}\left(2(M+N) + \frac{N(R^2 - 2\epsilon)}{\gamma}\right)\\ =&O\left(\frac{N}{\gamma}\left(M + N + \frac{N(R^2 - \epsilon)}{\gamma}\right)\right). \end{aligned} \tag{22}$$

□

Corollary. *Suppose that $||x_j|| \leq R$ for all $x_j \in \mathcal{L} \cup \mathcal{U}$. Let the current solution be w^c, and further suppose that there exists an optimal solution w^* such that $y_j(w^*)^T x_j \geq \gamma$ for all examples x_j. Let $||w^c|| = M$ and $||w^*|| = N$. Suppose the probability of selecting the points satisfying the inequality $y^+ w^T x^+ < \gamma$ is P_a. Then the total number of label requests made by passive learning is at most*

$$O\left(\frac{N}{\gamma P_a}\left(M + N + \frac{N(R^2 - \epsilon)}{\gamma}\right)\right).$$

Proof. This corollary can be directly derived from the above theorem, and hence we skip the proof and only present the result. □

According to the theoretical justifications provided by the above convergence theorem and corollary, we get the following conclusions: (1) because $0 < P_a < 1$, the upper bound of label requests made by MMC is proven to be smaller than random selection, demonstrating that the margin-based strategy is expected to outperform passive learning, and (2) the convergence property guarantees that MMC converges with the maximal label requests derived above.

5 Linkage between MMC and Simple Margin

As discussed before, one of the most widely used SVM active learning solution is simple margin, which chooses the points that are closet to the decision boundary. The distance between a point x and the boundary $w^{\mathrm{T}}x = 0$ is computed as:

$$\mathrm{Dist}(w, x) = \frac{|w^{\mathrm{T}}x|}{||w||}, \tag{23}$$

and the sampling function can be written as:

$$x^*_{\mathrm{SM}} = \arg\min_{x^+ \in \mathcal{U}} \mathrm{Dist}(w, x^+) = \arg\min_{x^+ \in \mathcal{U}} |w^{\mathrm{T}}x^+|. \tag{24}$$

Although it achieves good practical performance, it still lacks of reasonable theoretical justifications.

Here, we attempt to explore the connection between MMC and simple margin to provide a potentially theoretical justification. Let $x^+_{(j)}$ be the j-th close-to-boundary example in the pool, e.g. $x^+_{(1)} = x^*_{\mathrm{SM}}$. Assume there are m unlabeled examples in the pool. According to Eq. (24), we have

$$|w^{\mathrm{T}}x^*_{\mathrm{SM}}| = |w^{\mathrm{T}}x^+_{(1)}| < |w^{\mathrm{T}}x^+_{(2)}| < \cdots < |w^{\mathrm{T}}x^+_{(m)}|. \tag{25}$$

Now, let us consider the inequality $|w^{\mathrm{T}}x^+| < \gamma$ used for sample filtering. If we restrict the margin parameter γ as:

$$|w^{\mathrm{T}}x^*_{\mathrm{SM}}| < \gamma \leq |w^{\mathrm{T}}x^+_{(2)}|, \tag{26}$$

it is clear to see that there will be only one point, i.e. the one most close to boundary, satisfying this inequality. Hence we have

$$x^*_{\mathrm{SM}} = \Omega = \{x : |w^{\mathrm{T}}x^+| < \gamma\} \Rightarrow x^*_{\mathrm{SM}} = x^*_{\mathrm{MMC}}. \tag{27}$$

Thus, simple margin can be viewed as a special case of MMC, and the theoretical results derived above is applicable to this popular method as well.

6 Experiments

6.1 Data Sets and Experimental Settings

To validate the performance of the proposed algorithm, we use eight benchmark data sets of various sizes from the UCI machine learning repository[1]: **Biodeg**,

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

Table 1. The information of the eight binary-class data sets from UCI repository

Data set		# Examples	# Features	Class distribution
Biodeg		1055	41	356/699
Ionosphere		351	34	225/126
Parkinsons		195	22	147/48
WDBC		569	30	357/212
Letter	**D-vs-P**	1608	16	805/803
	E-vs-F	1543	16	768/775
	M-vs-N	1575	16	792/783
	U-vs-V	1577	16	813/764

Ionosphere, **Parkinsons**, **WDBC**, **Letter**. For **Letter**, 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 construct a binary-class data set for each pair. Table 1 shows the information of the eight binary-class data sets.

Each data set is randomly divided into three disjoint subsets: the base labeled training set (denoted as $\mathcal{L}$), the unlabeled pool set (denoted as $\mathcal{U}$), and the test set (denoted as $\mathcal{T}$). We use the base labeled set $\mathcal{L}$ as the small labeled data set to train the initial SVM 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 active learning algorithms. More specifically, the active learning scenario for each data set is constructed as: $\mathcal{L}(5\%)+\mathcal{U}(75\%)+\mathcal{T}(20\%)$. We normalize the features with the function below:

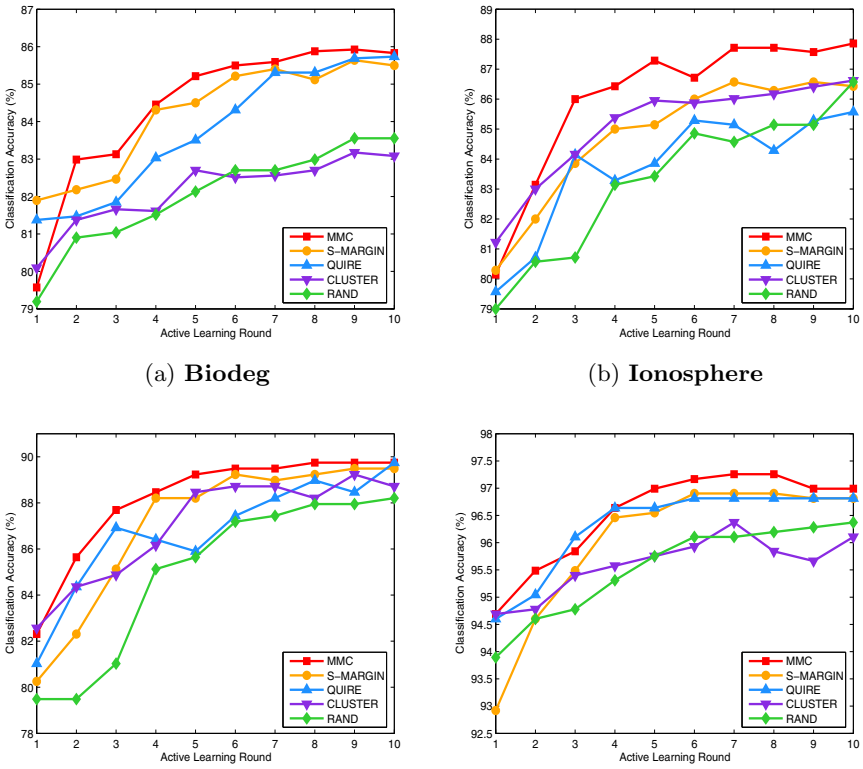
$$f^N_{(i,j)} = \frac{f_{(i,j)} - \min_{i\in n}\{f_{(i,j)}\}}{\max_{i\in n}\{f_{(i,j)}\} - \min_{i\in n}\{f_{(i,j)}\}}, \tag{28}$$

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

The optimal margin parameter γ is determined by the standard 5-fold cross validation. In this study, the active learning process iterates 10 rounds. In each round of data selection, 3% of the whole examples are selected from $\mathcal{U}$. These examples are then added to the training set, and SVM classifiers are re-trained and tested on the separate test set $\mathcal{T}$.

6.2 Comparison Methods and Evaluation Metric

To test the effectiveness of the proposed active learning algorithm, we compare it against the following four competitors including three state-of-art active learning for SVMs methods, and one baseline random selection: (1) S-MARGIN [18]: the simple margin algorithm, (2) CLUSTER [5]: the clustering-based representative sampling approach, (3) QUIRE [14]: the multi-criteria-based sampling, and (4) RAND: the random sampling. A detailed description of each of these algorithms



(a) **Biodeg** (b) **Ionosphere**

(c) **Parkinsons** (d) **WDBC**

Fig. 1. Comparison results of different active learning algorithms on the **Biodeg**, **Ionosphere**, **Parkinsons**, and **WDBC** data sets

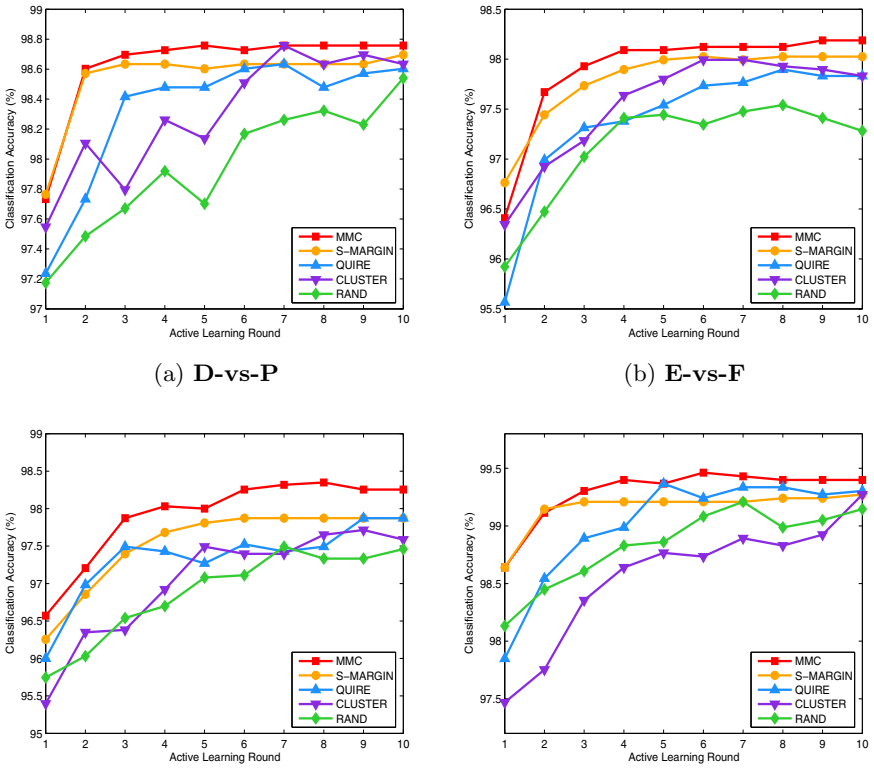
is provided in Section 2. For evaluation, the classification accuracy is adopted to measure the performance on the test set:

$$\text{Accuracy} = \frac{1}{|\mathrm{T}|}\sum_{i=1}^{|\mathrm{T}|}\mathbf{1}\{f(x_i) = y_i\}, \tag{29}$$

where $|\mathrm{T}|$ stands for the size of the test set, and y_i and $f(x_i)$ are the ground truth and prediction of x_i, respectively. $\mathbf{1}\{.\}$ is the indicator function. To avoid random fluctuation, each experiment is repeated 10 times by varying the base-pool-test sets, and the averaged classification accuracy is reported.

6.3 Comparison Results and Discussions

The comparison results of the five data selection algorithms on these eight UCI benchmark data sets are presented in Figure 1 and Figure 2. The X-axis denotes



(a) **D-vs-P** (b) **E-vs-F**

(c) **M-vs-N** (d) **U-vs-V**

Fig. 2. Comparison results of different active learning algorithms on the **D-vs-P**, **E-vs-F**, **M-vs-N**, and **U-vs-V** data sets

the number of iterations for the active learning process, and the Y-axis represents the classification accuracy. Several general observations as shown in these figures are explained as follows.

(1) For all five algorithms, the classification accuracy generally increases with the iterations of active learning, which matches the intuition that model's performance is positively correlated with the amount of training set available.

(2) The proposed MMC algorithm is observed to perform the best among the five approaches in most cases during the entire data selection process, demonstrating that the proposed active learning method is more effective in choosing the most informative examples to improve the model quality. This is likely due to the reason that MMC quantifies the model change as the gradient of the loss, which is highly correlated with the objective function used to evaluate the SVM models. Therefore, the examples selected by MMC are more likely to contribute positively to improve the model. In addition, we observe that MMC converges much faster than the competitors on several data sets (e.g. **D-vs-P**, **E-vs-F**, **M-**

Table 2. The p-value of Wilcoxon signed rank test of MMC versus S-MARGIN, QUIRE, CLUSTER and RAND on the UCI data sets

Data sets	vs. S-MARGIN	vs. QUIRE	vs. CLUSTER	vs. RAND
Biodeg	$\underline{p<0.1}$	$\underline{p<0.1}$	$p<0.05$	$p<0.05$
Ionosphere	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$
Parkinsons	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$
WDBC	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$
D-vs-P	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$
E-vs-F	$\underline{p<0.1}$	$p<0.05$	$p<0.05$	$p<0.05$
M-vs-N	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$
U-vs-V	$p<0.05$	$p<0.05$	$p<0.05$	$p<0.05$

vs-N, **U-vs-V**), i.e. the highest classification accuracy is achieved with much less examples added to the training set. This agrees with the intuitive explanation that the data examples greatly changing the current classifier are expected to produce faster convergence to the optimal model.

(3) We see that the performance of CLUSTER is inconsistent. It works well on some data sets, but performs poorly on the others. This phenomena may be explained as follows. The CLUSTER method utilizes a clustering technique to choose the representative data points, which may fail if there is no clear cluster structure in the data. On the contrary, QUIRE is observed to yield relatively good performance on most data sets. The success of QUIRE may be attributed to the principle of choosing examples that are both informative and representative.

(4) To better validate the effectiveness of the proposed approach, we conduct the significance test on the comparisons. Table 2 presents the results of Wilcoxon signed rank test of MMC versus S-MARGIN, QUIRE, CLUSTER and RAND strategies on the benchmark UCI data sets. The comparison results with $p>0.05$ are underlined. It shows that the proposed method performs statistically better ($p<0.05$) than S-MARGIN, QUIRE, CLUSTER and RAND on most data sets. We also perform the 2-tailed paired T-test to further examine the effectiveness of MMC. Due to the space limitation, the p-values according to the 2-tailed T-test are not reported here, and the results show that MMC significantly outperforms ($p<0.05$) the competitors in most cases during the sample selection process.

6.4 Efficiency Comparison

In this subsection, we compare the CPU running time taken by MMC versus the competitors. All algorithms were implemented using MATLAB on a standard desktop computer with 2.53 GHz CPU and 8 GB of memory.

Table 3 shows the comparison results, together with the information of the pool set. As shown in the table, the time complexity of MMC is slightly higher than that of S-MARGIN, but much more efficient than the other two strategies, i.e. QUIRE and CLUSTER. This is due to the reason that QUIRE involves

Table 3. The CPU running time (seconds), together with the information of pool set

Data sets	# Ex. × Features ($\mathcal{U}$)	MMC	S-MARGIN	QUIRE	CLUSTER
Biodeg	791 × 41	0.04	0.01	100.85	1.12
Ionosphere	263 × 34	0.02	0.01	3.79	0.21
Parkinsons	146 × 22	0.00	0.00	0.84	0.15
WDBC	427 × 30	0.01	0.00	16.38	0.44
D-vs-P	1206 × 16	0.07	0.02	341.94	1.03
E-vs-F	1157 × 16	0.02	0.01	301.87	0.78
M-vs-N	1181 × 16	0.02	0.02	324.70	1.02
U-vs-V	1183 × 16	0.05	0.01	219.10	1.00

calculating the inverse of a large scale matrix, and CLUSTER requires considerable efforts on clustering. In summary, the proposed MMC method is quite efficient in computational complexity, and is promising for real-world applications.

7 Conclusions

In this paper, focusing on SVMs, we introduce a new interpretation for margin-based active learning with the idea of expected model change, and accordingly propose a novel margin-based active learning algorithm named Maximum Model Change (MMC), which is to choose the examples leading to the maximal change in the current classifier. The change is measured as the difference between the current model parameters and the updated parameters trained with the accumulated training set. Inspired by the SGD rule for solving the SVMs optimization problems, the change is approximated as the gradient of the loss at a candidate point. In addition, we provide a theoretical analysis of the convergence property for the proposed algorithm, and compare the derived sampling bound against passive learning. The comparison shows that the upper bound of sample requests made by MMC is smaller than passive learning. We further connect the proposed approach with the widely adopted simple margin approach to provide a theoretical justification for this popular algorithm. Substantial experimental results on various benchmark UCI data sets have demonstrated that the proposed strategy is highly effective in selecting informative examples, and efficient in computation time.

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

References

1. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proc. of EMNLP 2008, pp. 1070–1079 (2008)

2. Abe, N., Mamitsuka, H.: Query learning strategies using boosting and bagging. In: Proc. of ICML 1998, pp. 1–10 (1998)
3. Donmez, P., Carbonell, J.G.: Optimizing estimated loss reduction for active sampling in rank learning. In: Proc. of ICML 2008, pp. 248–255 (2008)
4. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning, 133–168 (1997)
5. Xu, Z., Yu, G., Tresp, V., Xu, X., Wang, J.: Representative sampling for text classification using support vector machines. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 393–407. Springer, Heidelberg (2003)
6. Settles, B.: Active learning. Morgan & Claypool (2012)
7. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proc. of ICML 2004, pp. 623–630 (2004)
8. Tang, M., Luo, X., Roukos, S.: Active learning for statistical natural language parsing. In: Proc. of ACL 2002, pp. 120–127 (2002)
9. Li, L., Jin, X., Pan, S., Sun, J.: Multi-domain active learning for text classification. In: Proc. of KDD 2012, pp. 1086–1094 (2012)
10. Li, M., Sethi, I.K.: Confidence-based acitve learning. IEEE Trans. Pattern Analysis and Machine Intelligence, 1251–1261 (2006)
11. Vapnik, V.: The nature of statistical learning Theory. Springer (1999)
12. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 121–167 (1998)
13. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: Proc. of ICML 2003 (2003)
14. Huang, S., Jin, R., Zhou, Z.: Active learning by querying informative and representative examples. In: Proc. of NIPS 2010, pp. 892–900 (2010)
15. Shwartz, S., Singer, Y., Srebro, N.: Pegasos: primal estimated sub-gradient solver for SVM. In: Proc. of ICML 2007, pp. 807–814 (2007)
16. Wang, Z., Crammer, K., Vucetic, S.: Breaking the curse of kernelization: budgeted stochastic gradient descent for large-scale SVM training. Journal of Machine Learning Research, 3103–3131 (2012)
17. Shen, D., Zhang, J., Su, J., Zhou, G., Tan, C.L.: Multi-criteria-based active learning for named entity recognition. In: Proc. of ACL 2004, pp. 589–596 (2004)
18. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of Machine Learning Research, 45–66 (2001)
19. Kivinen, J., Smola, A., Williamson, R.: Online learning with kernels. IEEE Trans. Signal Processing, 2165–2176 (2004)
20. Roy, N., McCallum, A.: Toward optimal active learning through sampling estimation of error reduction. In: Proc. of ICML 2001, pp. 441–448 (2001)
21. Chon, A., Ghahramani, Z., Jordan, M.I.: Active learning with statisitical models. Journal of Machine Learning Research, 129–145 (1996)
22. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Proc. of SIGIR 1994, pp. 3–12 (1994)
23. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Proc. of NIPS 2008, pp. 1289–1296 (2008)
24. Cai, W., Zhang, Y., Zhou, J.: Maximizing expected model change for active learning in regression. In: Proc. of ICDM 2008, pp. 51–60 (2013)