Pattern Recognition 44 (2011) 2375–2383

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Active learning with adaptive regularization

Zheng Wang [a,*], Shuicheng Yan [b], Changshui Zhang [a]

[a] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing 100084, China

[b] Department of Electrical & Computer Engineering, National University of Singapore, Singapore 117576, Singapore

ARTICLE INFO

*Article history:*
Received 15 September 2010
Received in revised form
18 January 2011
Accepted 7 March 2011
Available online 15 March 2011

*Keywords:*
Active learning
Adaptive regularization
SVM
TSVM

ABSTRACT

In classification problems, active learning is often adopted to alleviate the laborious human labeling efforts, by finding the most informative samples to query the labels. One of the most popular query strategy is selecting the most uncertain samples for the current classifier. The performance of such an active learning process heavily relies on the learned classifier before each query. Thus, stepwise classifier model/parameter selection is quite critical, which is, however, rarely studied in the literature. In this paper, we propose a novel active learning support vector machine algorithm with adaptive model selection. In this algorithm, before each new query, we trace the full solution path of the base classifier, and then perform efficient model selection using the unlabeled samples. This strategy significantly improves the active learning efficiency with comparatively inexpensive computational cost. Empirical results on both artificial and real world benchmark data sets show the encouraging gains brought by the proposed algorithm in terms of both classification accuracy and computational cost.



## 1. Introduction

Nowadays, with the fast development of Internet techniques and explosive increasing of data warehouses, the unlabeled data is becoming abundant or easily obtained in most cases. On the other hand, annotating the unlabeled samples is costly work and very time consuming. The label information, however, is very important for training a satisfactory learner in machine learning and other related problems. Therefore, more and more attention has been paid to finding a good classifier with minimum labeling efforts in recent years.

*Active learning* is one of the most popular techniques to save human labeling efforts. It has been widely adopted into the sophisticated supervised and semi-supervised tasks [12]. Active learning support vector machine (SVM) [14] is one of the most representative and practical approaches for pool based active learning in machine learning literature. The pioneering investigation is based on hard margin SVM. Furthermore, Campbell et al. [2] introduce active learning into the soft margin SVM, which is much more powerful and practical to deal with the nonseparable classes [15]. Both of these two algorithms prefer to choose the most uncertain sample for current classifier and query its label. This is a typical query criterion for the discriminant models, owing to its simplicity and efficiency [12]. In such active learning scenarios, the label query and classifier modeling are highly correlated. As the query heavily relies on the current classifier, an inappropriate model may lead to very poor active learning performance [9], which behaves as unsatisfactory learning accuracy and inefficient queries. Though active learning is well known for its benefit of saving labeling efforts, it is often less efficient than random query in the initial stages, with very few labeled data. This phenomena can be observed empirically in many popular active learning methods [1,14,17]. This might further prejudice the query efficiency of the whole active learning process, using such an unsatisfactory initialization. As a result, model selection is critical for active learning [13,1].

Choosing the best model is a very difficult problem in both machine learning and statistics fields [15]. It is often reduced and formulated as the parameter selection problem. In soft margin SVM, finding the best classifier can be formulated as a regularized optimization problem. A tunable parameter is used to control the regularization quantity. Given the loss function and the penalty, selection of a good value for the tunable parameter is the model selection problem [7]. In conventional supervised learning methods, the training data is often given beforehand and fixed. In this situation, once a satisfactory parameter is found, it will be fixed as a constant and used all through the following learning process. However, in active learning scenario, the number of the labeled data continually increases with the machine queries. During this process, the training data compose a dynamic set. Correspondingly, the

* Corresponding author.
*E-mail address:* wangzheng04@gmail.com (Z. Wang).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.03.008

learning model should be changed with respect to the data set. In this work, we will show that using fixed regularization parameters is not a very good choice for active learning problems. When the number of labeled data increases, a dynamic parameter is desirable to guarantee a satisfactory learning result.

In this paper, we propose a very efficient active learning method for soft margin SVM with model selection. In supervised learning scenario, cross-validation is one of the most popular ways to find the proper parameter for the SVM model [15,7]. It is to split the available labeled data into a training and a validation sets. The training data is used to construct the SVM classifiers for different parameters, and the validation set is then used to select the most proper one. However, this setting cannot be easily adopted into the active learning framework. The most important reason is that the queried samples are not independently and identically distributed when sampled from the original data distribution. Using the queried data as the validation set may get severe overfitting and thus mislead the following query process. The problem becomes even worse when the original available labeled samples are scarce, which is commonly seen in the active learning. This is also one probable reason why conventional active learning is often less efficient than random query in the initial stages. To tackle this issue, in this work we use the unlabeled samples to compose a pseudo-validation set, and we prove that it works well both theoretically and empirically. To make the parameter selection process more efficient, we introduce the regularization path method [7] into the active learning process to efficiently compute the models based on different regularization values.

The rest of this paper is organized as follows. In Section 2, we introduce the model selection step into active learning framework based on SVM. In Section 3, we present a practical active learning algorithm with an adaptive model, which is called active learning SVM path. In Section 4, we discuss the relationship and difference between our proposed method and the conventional transductive SVM (TSVM) method. The experiments for empirical analysis are given in Section 5. Finally, we conclude in Section 6.

## 2. Active learning with dynamic SVM

Suppose initially there are $l$ labeled points $\mathbf{X}_L=\{\mathbf{x}_1,\ldots,\mathbf{x}_l\}$, $\mathbf{x}_i\in\mathcal{X}\subseteq\mathcal{R}^m$, with labels $\mathbf{y}_L=\{y_1,\ldots,y_l\}$, where $y_i\in\mathcal{Y}=\{1,-1\}$ as we focus on binary problems here. There is also a pool of $u$ unlabeled points $\mathbf{X}_U=\{\mathbf{x}_{l+1},\ldots,\mathbf{x}_{l+u}\}$, $\mathbf{D}_L=\{\mathbf{X}_L,\ \mathbf{y}_L\}$ are randomly generated according to some unknown probability $P(\mathbf{x},y)$. $\mathbf{D}_U=\mathbf{X}_U$ are randomly generated from the marginal probability $P(\mathbf{x})$.

### 2.1. Soft margin SVM formulation

The soft margin SVM searches for an optimal hyperplane $f(\mathbf{x})=\mathbf{w}^T\Phi(\mathbf{x})+b$ by solving the following optimization problem[1]:

$$\min_{\mathbf{w},b}\quad \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{l}\xi_i^s,$$

$$\text{s.t.}\quad y_i(\mathbf{w}^T\Phi(\mathbf{x}_i)+b)\geq 1-\xi_i,\ \ \xi_i\geq 0,\ \ 1\leq i\leq l,$$

where $C$ is a trade-off parameter and $\Phi(\mathbf{x}_i)$ is a function mapping the input data into a feature space where the data is better discriminated or represented. For linear case, $\Phi(\mathbf{x})=\mathbf{x}$.

In statistical learning theory, based on structural risk minimization, the soft margin SVM can also be formulated within the regularized unconstrained optimization framework as follows:

$$\hat{f}=\operatorname*{argmin}_{f\in\mathcal{F}}\sum_{i=1}^{l}L(y_i,f(\mathbf{x}_i))+\lambda\Omega(f), \tag{1}$$

where $L(y_i,f(\mathbf{x}_i))=\max(0,1-y_if(\mathbf{x}_i))$ is the hinge loss function, $\Omega(f)=\|\mathbf{w}\|^2$ is the regularization term, which describes the model complexity, and $\lambda$ is the regularization parameter, which controls the regularization quantity. Selection of a good value of $\lambda$ is a so-called model selection problem.

The optimal solution for the soft margin SVM can be explicitly expressed as

$$\hat{f}(\mathbf{x})=\frac{1}{\lambda}\sum_{i=1}^{l}\alpha_iy_iK(\mathbf{x},\mathbf{x}_i)+b, \tag{2}$$

where $K(\mathbf{x},\mathbf{x}_i)=\Phi(\mathbf{x})\cdot\Phi(\mathbf{x}_i)$ is the kernel function. In this formulation, there are only a part of the data involved in the expression of the classifier, for which $0<\alpha_i\leq 1$.

### 2.2. Active learning SVM

In standard active learning SVM, the optimal classifier is among the hypotheses correctly classifying current labeled data. These consistent hypotheses compose the version space [14]. The active learner tries to find the new sample, which expectedly reduces the version space optimally, to query [14]. Though there is no consistent hypothesis to compose the version space in inseparable case for soft margin active learning SVM, the same query scenario is still found to be the most reasonable one [2]. This is best known as the most uncertainty principle, which is used in many active learning methods [12]. It can be expressed as

$$i=\operatorname*{argmin}_{i\in U}|\hat{f}(\mathbf{x}_i)|. \tag{3}$$

The premise of query efficiency under this strategy is a well trained classifier $\hat{f}(\mathbf{x})$ in (2), which gives accurate predictions for unlabeled samples. It is constructed by the soft margin SVM, which is the solution of the regularized optimization, expressed by (1).

### 2.3. Active learning with dynamic regularization

The target of the learning problem is to find the optimal classifier, which is expectable to best approximate the Bayesian rule. Under this circumstance, active learning and model selection are two complementary and heavily correlated problems. However, these two parts have been studied separately as two independent problems, and little research has been done to consider them together. So far as we know, [13] is the only work to analyze such a problem. In [13], an ensemble active learning method is presented for linear regression, which averages all available models for active learning and is computationally expensive. Besides, the proposed algorithm in [13] cannot deal with the classification problem, which is thus still an open problem.

It is obvious that in soft margin SVM and other similar regularized optimization problems, the solution is decided by two factors. One is the current training set $\mathbf{D}_L=\{\mathbf{X}_L,\ \mathbf{y}_L\}$ and the other one is the regularization parameter $\lambda$. The optimum solution can be expressed as a function of these two factors, $\hat{f}(\cdot)=\hat{f}(\cdot|\mathbf{D}_L,\lambda)$. The target is to find the $\hat{f}$ to best approximate the Bayesian classifier, using all currently available knowledge. Intuitively, it is reasonable to use different amounts of regularization when training different numbers of labeled data. Proposition 1 preliminarily analyzes the behavior of the regularization parameter changing with respect to the increase of

[1] In this paper we focus on the 1-norm soft margin SVM with $s=1$.

the training set, and how this affects the learning accuracy and speed. It is originally given by [5], which is applicable to general regularization problem, and [10] specifies it for SVM. It reflects that proper update of the regularization parameter helps to control the learning performance in incremental learning for SVM.[2]

**Proposition 1.** *Assume the* $\mu(\mathbf{x})=Pr(y=1|\mathbf{x})$ *is the conditional probability in a Hilbert space of functions on* $[0,1]^d$, *which are m times differentiable,* $m\geq 2$. $g(\mathbf{x})=\mathrm{sign}(2\mu(\mathbf{x})-1)$ *is the corresponding Bayesian classifier.* $\hat{f}(\mathbf{x})$ *is the optimal classifier for regularized problem* (1). *Then with the increase of the number of the labeled training samples, if* $\lambda/l\to 0$, *and* $l^{-1}(\lambda/l)^{-((3/2m)+\varepsilon)}\to 0$, *for* $\varepsilon>0$, $\hat{f}(\mathbf{x})$ *will converge to the Bayesian classifier in the following manner*:

$$\int_{\mathbf{x}}(\hat{f}(\mathbf{x})-g(\mathbf{x}))^2=O(\lambda/l)+O_\mu(l^{-1}(\lambda/l)^{-(1/2m)}(\log(l/\lambda))^{d-1}).$$

$$\sup_{\mathbf{x}}|\hat{f}(\mathbf{x})-g(\mathbf{x})|=O((\lambda/l)^{(1/2)-(1/4m)-(\varepsilon/4)})+O_\mu(l^{-(1/2)}(\lambda/l)^{-((1/2m)+(\varepsilon/4))}(\log(l/\lambda))^{d-1}).$$

In this proposition, the expressions describe the asymptotic relationship between the best learnt classifier $\hat{f}$ and the optimal Bayesian classifier $g$. They show that the convergence rate of $\hat{f}$ to $g$ is dominated by two factors, the regularization weight $\lambda$ and the number of the labeled data $l$. With the increase of labeled samples, $\lambda/l$ should be chosen in specific smoothing manner under the corresponding measure [10]. In general supervised learning problems, it is reasonable to select the regularization parameter and fix it for the corresponding training task, as the training set is given beforehand and fixed. However, it should not be the case for incremental learning. Accordingly, it is neither appropriate for active learning, and may lead to unsatisfactory suboptimal results. An improperly designed regularization may induce active learning to query useless or even harmful samples, and further mislead the following update of the classifier and the query of new samples.

In active learning, the situation is much more complex than the standard incremental learning problem. Thus we need to carefully revise the parameter during the whole active learning process to keep high learning accuracy and query efficiency.

### 2.4. Model selection using unlabeled data

In conventional regularized learning problems, prior given a set of candidate parameters, the most suitable one is selected to deduce the satisfactory classifier. It is a very popular choice to use the one obtaining the highest prediction accuracy on extra validation data. Besides this validation method, Lin [10] reveals that the margin of the data can properly control the error rate, therefore can also be used as the parameter selection measure. The following theorem describes this characteristic.

**Proposition 2** (*Lin [10]*). *Let* $E[\max(0,1-yf(\mathbf{x}))]$ *stand for the expectation of the hinge loss. The minimizer* $\hat{f}(\mathbf{x})$ *of* $E[\max(0,1-yf(\mathbf{x}))]$ *is the Bayesian rule.*

As a result, the empirical approximation of the hinge loss can be used as the validation criterion to find the best parameter, namely,

$$\hat{\lambda}=\operatorname{argmax}_{\lambda\in\Lambda}\sum_i\max(0,1-y_if(\mathbf{x}_i)), \tag{4}$$

where $\Lambda$ is the candidate set of the tunable parameter.

In active learning, the labeled data is very scarce and also has severe bias from the original data distribution. Focusing too much on the queried data takes a high risk of overfitting, which then leads to ineffective active learning. Though the queried data cannot be used as the validation set, there are plenty of unlabeled data to use. Thus we propose a model selection criterion based on the margin of the unlabeled data, which is a pseudo-margin [6] based on the prediction label given by current classifier:

$$\hat{\lambda}=\operatorname{argmax}_{\lambda\in\Lambda}\sum_{i=1}^{u}\max(0,1-|f(\mathbf{x}_i)|). \tag{5}$$

$|f(\mathbf{x}_i)|$ is the pseudo-margin with pseudo-label $\hat{y}(\mathbf{x}_i)=\mathrm{sign}(f(\mathbf{x}_i))$.

It is easy to see that this large pseudo-margin criterion is a good backup, and we present Proposition 3 to show that the pseudo-margin is an acceptable approximation of the true margin. It is a direct deduction from Proposition 2, so we omit the proof here.

**Proposition 3.** *The pseudo-margin of the unlabeled validation set will control the true margin of these samples, from both upper and lower sides, when u is sufficient enough*:

$$\begin{aligned}&\lim_{u\to\infty}\frac{1}{u}\sum_{i=1}^{u}\max(0,1-|f(\mathbf{x}_i)|)\\&\leq\lim_{u\to\infty}\frac{1}{u}\sum_{i=1}^{u}\max(0,1-y_if(\mathbf{x}_i))\\&\leq\lim_{u\to\infty}\frac{1}{u}\sum_{i=1}^{u}\max(0,1-|f(\mathbf{x}_i)|)+2P_e(f),\end{aligned}$$

*where* $P_e(f)$ *is the error rate of the classification function f.*

### 2.5. Solution path for SVM

Based on the above analysis, we should use dynamic regularization in the active learning process. In practical problems, extensive exploration of the optimal regularization parameter is seldom pursued, since this requires retraining the model many times corresponding to different parameter settings through a candidate set $\Lambda$. The computational burden is extremely heavy and intractable in active learning problems, as the parameter should be updated all through the query process. Fortunately, the solution of SVM for all regularization parameter values can be explored efficiently along a solution path, without having to retrain the model multiple times [7]. Moreover, the solution path for SVM is piecewise linear with respect to the regularization parameter $\lambda$. When other factors are fixed, there exists a strictly decreasing sequence $\lambda^t\geq\lambda^{t+1}\geq 0$, $t=1,\ldots,N$, such that the SVM solution $\hat{f}_\lambda$ can be represented as a piecewise linear function of the regularization parameter $\lambda$ as

$$\hat{f}_\lambda=\hat{f}_{\lambda^t}+(\lambda-\lambda^t)h^t,\quad\forall\lambda\in[\lambda^{t+1},\lambda^t],$$

where $h^t$, $t=1,\ldots,N$ denotes a sequence of functions in $\mathcal{F}$. Refer to [7] for more precise details.

The SVM solution path algorithm starts from a large $\lambda^*\approx\infty$. The initial solution is obtained by solving a linear programming problem, and then the algorithm iteratively computes the grid point $\lambda^{t+1}$ and sensitive weight $h^{t+1}$. As $\lambda$ decreases, the algorithm computes the solution for every value of $\lambda$. Following this path, without introducing much extra computational cost, the optimum classifier $\hat{f}_\lambda$ for different regularization values in the candidate set $\Lambda$ is efficiently traced and constructed.

[2] Though this characteristic is derived under the 2-norm loss, as 1-norm hinge loss function is not differentiable and the 1-norm soft margin SVM should be in a similar manner, though it is difficult to explicitly express.

## 3. Active learning SVM path

**Algorithm 1.** ASVMpath.

**Input:** $l$ labeled data $\mathbf{D}_L$; $u$ unlabeled data $\mathbf{X}_U$.
**Repeat**
**Step 1:** Construct $\hat{f}_\lambda$ for every $\lambda \in \Lambda$.
**Step 2:** Find the best $\hat{\lambda}$ for Eq. (5).
**Step 3:** Label the most uncertain data in $\mathbf{X}_U$ based on $\hat{f}_{\hat{\lambda}}$, move it to $\mathbf{D}_L$.
**until** *stop criterion* is *true*

Based on the analysis given in the above section, we summarize the active learning SVM path (ASVMpath) algorithm, and the pseudo-codes are listed in Algorithm 1.

Initially we are given $l$ labeled points $\mathbf{D}_L=\{\mathbf{X}_L, \mathbf{y}_L\}$, and a pool of $u$ unlabeled points $\mathbf{X}_U$, and set the upper bound of the regularization parameter sufficiently large as $\lambda^*$. Then we run the following three steps iteratively, and our active learning SVM path algorithm efficiently finds the most informative samples to label and consistently construct satisfactory classifiers.

**Step 1:** Use labeled data to train the soft margin SVM classifier, and construct the regularization path for a candidate parameter set $\Lambda \subseteq [0,\lambda^*]$.
**Step 2:** Use the unlabeled samples to find the most suitable regularization parameter using Eq. (5), and construct current best classifier.
**Step 3:** Find the most uncertain sample, based on current classifier and query its label, as stated in Eq. (3). Then put it into the training set.

## 4. Discussion

In our proposed active learning SVM path algorithm, both labeled and unlabeled data are used to find the satisfactory model and the most demanded labels. Transductive SVM (TSVM) is an off-the-shelf method to incorporate both labeled and unlabeled data for training a better classifier [8]. It is a natural consideration to introduce active learning directly into the TSVM framework. However, we will analyze the difference of these two scenarios and discuss the present shortage for active learning. In the meantime, we will show that our method is much more suitable and practical for active learning literature.

### 4.1. Less is more

TSVM combines both labeled and unlabeled data for training, and aims at finding suitable labels for the unlabeled data and a satisfactory classifier simultaneously, as

$$\begin{aligned}\min_{\mathbf{w},b}\quad & \frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{l}\xi_i+\tilde{C}\sum_{i=l+1}^{n}\xi_i,\\ \text{s.t.}\quad & y_i(\mathbf{w}^T\Phi(\mathbf{x}_i)+b)\geq 1-\xi_i,\ \xi_i\geq 0,\ 1\geq i\geq l;\\ & |\mathbf{w}^T\Phi(\mathbf{x}_i)+b|\geq 1-\xi_i,\ \xi_i\geq 0,\ l+1\geq i\geq n.\end{aligned}$$

Extra objective term and constraints are added into the standard SVM problem to embed the unlabel information. It can also be reformulated into a regularized optimization problem as

$$\hat{f}=\operatorname*{argmin}_{f\in\mathcal{F}}\sum_{i=1}^{l}L(y_i,f(\mathbf{x}_i))+\gamma\sum_{i=l+1}^{n}L(|f(\mathbf{x}_i)|)+\lambda\Omega(f).$$

It introduces the unlabeled samples directly into the training routine. The price is to introduce one more parameter to tune. The original TSVM [8] iteratively updates the classifier and the combination of $\hat{y}_i$'s, which is computationally very expensive. Collobrt et al. [4] have reviewed recent advances in solving TSVMs. Some methods run much faster than the original solution, while they are still far from satisfactory.

#### 4.1.1. Domain adaption

The success of TSVM heavily relies on the assumption that the labeled and unlabeled data are independently and identically distributed with respect to the same data distribution. However, in active learning, this condition is no longer guaranteed. This problem is also considered as domain adaption or transfer learning [13,11], where the training and test distributions may be different. Using TSVM for active learning may easily get overfitting to the already queried data and stuck in a biased model, while adaptively changing the TSVM model during the active learning process is difficult and time consuming. On the other hand, there are efficient ways to trace the solution path according to different regularization weights for soft margin SVM, so that the model can be changed adaptively corresponding to the change of training set after each query. This adaptive learning mechanism achieves much better prediction ability on the unlabeled samples, and is more suitable for active learning.

#### 4.1.2. Description vs. substantiation

In both SVM and TSVM, the solutions can be explicitly expressed using the training data by representor theorem [15] as in Eq. (2). TSVM uses the unlabeled data directly in the representation of the optimal classifier. It enhances the description ability of labeled data and increase the freedom of the classifier. However, this benefit can only be gained with well set regularization parameters. Though some heuristic parameter tuning methods are proposed for TSVM [4,16], they are only tractable for a fixed training set and only focus on one of the two trade-off parameters. To properly trace the best model among its feasible region is intractable, especially in active learning problems. As a result, though the description ability of TSVM is better, it is unlikely to find the most suitable solution under this representation.

Soft margin SVM only uses labeled data to construct the optimum classifier, which limits its description ability. However, the classifier with different regularization parameters can be efficiently computed and validated. In this situation, the most satisfactory solution under this representation is more probable to be obtained. Using a relatively simple representation, the adaptive regularization can be easily achieved during the dynamic learning process.

In practical learning problems, the available data should be properly used to train and validate the classifier so as to exert their optimum efficiency. When there are enough available information to ascertain all parameters with high probability, TSVM would be a better choice. On the contrary, in the situation of active learning, it always has limited and expensive resource. At this time, the adaptive regularized SVM is preferred.

In the machine learning community, researchers used to put much effort into finding the new powerful representation or hypotheses for the target concept in the learning problem, however, neglect the difficulty of finding a satisfactory suboptimum among these hypotheses with tractable validation procedure. This problem is much more critical in active learning scenarios. The substantiation of the result is as important as the result itself. The analysis and empirical results in this paper

suggest to sacrifice some description ability, while keeping a simpler model with good "validated ability".

### 4.2. Computational complexity

The model training complexity is very important for the efficiency of the active learning methods, as for each query, at least one classifier should be reconstructed. In standard TSVM, the training complexity depends on all available data, which is with the worst case $O((l+2u)^3)$. As far as we know, the CCCP TSVM [4] is the currently fastest TSVM solver which is in the order of $O((l+2u)^2)$. However, the magnitude of the polynomial is still non-neglectable when there are too many unlabeled samples. In our method, each classifier training process only cost $O(l^2)$ computational complexity, and the validation process need $O(u)$ complexity, which is extremely smaller than conventional methods. Thus it is especially practical and suitable in active learning, where $l \ll u$.

## 5. Experiments

In this section, we systematically compare the active learning SVM (ASVM), active learning TSVM (ATSVM) with our active learning SVM path (ASVMpath) algorithm. All three active learning methods use the most uncertain query principle. We also give the results based on random query (RSVMpath) as baselines.

### 5.1. Experiment setting

We use six data sets from the UCI benchmarks, and two artificial data sets from [3], which we call semi-supervised learning (SSL) data sets, to evaluate the above mentioned algorithms. The details of the data sets are shown in Table 1 for UCI benchmarks and Table 2 for SSL data sets. As the original TSVM is very time consuming, and it is not affordable in active learning scenario, we use the CCCP TSVM[3] as the TSVM solver. We also use the SVM solver in the same toolbox. We repeat all the algorithms 20 times for each data set, and the average results are reported.

In conventional active learning problems for classification, a common setting is that there should be sufficient labeled samples to initially train an acceptable classifier[4] [12]. However, in a practical problem where active learning is really needed, there is usually very limited initial label information. In our experiments, we prefer to simulate the latter more practical situation. In each run, we randomly sample four labeled data (two for each class) as the initial labeled data, and other samples then compose the unlabeled pool. In the experiments, each query selects one sample and labels it based on the ground-truth. The process stops when there are 80% or 400 samples labeled.

There is no free parameter in our ASVMpath in the linear case. For the nonlinear case, the only tunable factor in our method is the kernel setting, and we use the radial basis function (RBF) as the kernel function with the kernel width parameter being the median distance among the training data.

In ASVM and ATSVM, we need to choose proper values for the regularization parameters. Therefore, we conduct the methods for a candidate parameter set $C \in \{0.01,1,5,10,50,100\}$,[5] and the corresponding result is marked by $\text{ASVM}_C$ or $\text{ATSVM}_C$. As there is short of extra data to serve as the validation set in active learning, all the learning results are recorded for comparisons. For nonlinear case, the kernel setting of ASVM and ATSVM is the same as that for ASVMpath.

Though we record the results for all the candidate parameters in the following comparisons, it does not mean that ASVM and ATSVM can obtain the best result among them, as parameter selection should be conducted before the testing process and thus is easily biased to a suboptimum solution. Our ASVMpath algorithm suffers great injustice to be compared with the best of those results. However, our method still performs the best in almost all cases, which is a very challenging task and shown in the following experiments.

**Table 1**
UCI benchmarks used in our experiments.

| Data set | # Dimension | # Samples |
|---|---|---|
| Australian | 14 | 690 |
| Chess | 36 | 3196 |
| Crx | 15 | 690 |
| Breast | 10 | 699 |
| Heart | 13 | 270 |
| Vote | 16 | 435 |

**Table 2**
SSL data sets used in our experiments.

| Data set | # Dimension | # Samples |
|---|---|---|
| G50C | 10 | 550 |
| G10N | 10 | 550 |

### 5.2. Comparisons with ASVM on UCI benchmarks

In these experiments, we compare the methods of active learning SVM with fixed regularization (ASVM) and adaptive regularization (ASVMpath) on UCI benchmarks. As most of the data sets are collected from the real world problems, and have complex distributions, we compare all algorithms for their nonlinear versions. The learning accuracy curves during the whole active learning process are shown in Fig. 1.[6] We can see our proposed ASVMpath algorithm consistently performs better in most cases. These results show that our adaptive regularization method is indeed helpful for active learning, and the parameter selection strategy based on unlabeled data works properly. The improvement introduced by our method is expectable, as our algorithm automatically selects the relatively better models during the active learning process. To give a comprehensive view, we will analyze all related methods together after showing the ATSVM learning results. Note that some curves cannot be fully observed in the figure. There are two situations. One is that the curves are covered by other curves, the other one is that the performance is too bad to reach the current scope.

### 5.3. Comparisons with ATSVM on UCI benchmarks

In these experiments, we compare our method with ATSVM on UCI benchmarks. We still compare their nonlinear versions. The learning accuracy curves during the whole active learning process are shown in Fig. 2. It tells us that our proposed ASVMpath

[3] The code is from http://www.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html.

[4] There can be no such requirement for regression problem, such as in [17]. Though we can use regression model to solve the classification problem, it is not the optimum choice. We focus on directly solve the active learning classification problem with very limited initial labeled data.

[5] The learning curves of $C=0.1$ and 1 are very similar, and they are covered with each other in most cases in the plots, so we only give $C=1$ here.

[6] The starting accuracy is different for the curves, as the SVMs for different curves use different regularization parameters.

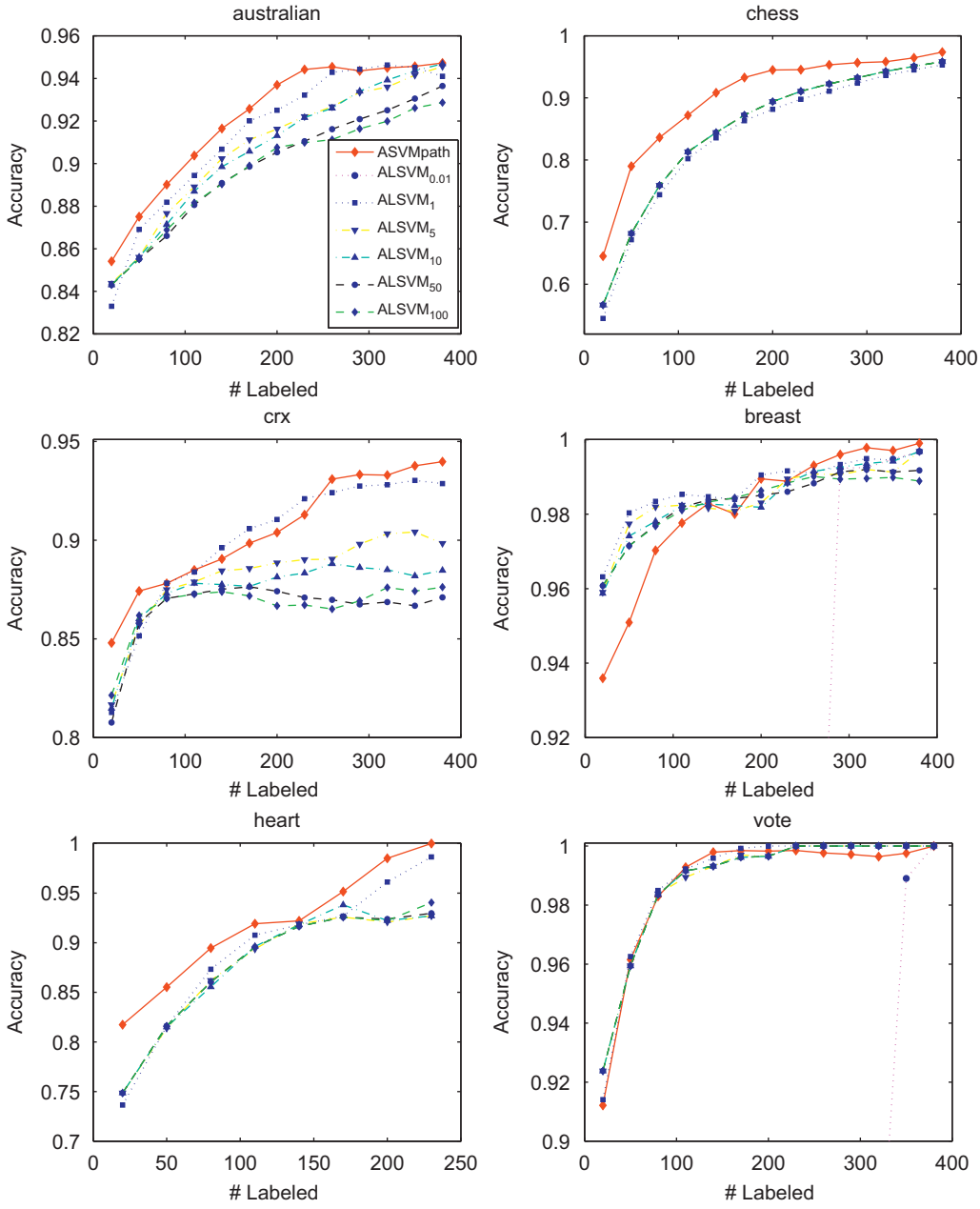


**Fig. 1.** Active learning SVM comparison results on UCI benchmarks. The compared methods are the same for all data sets. We indicate the correspondence between the curve and the method in the first plot. Each curve represents the average result of 20 runs.

algorithm consistently performs better in most cases. Combined with the results shown in Fig. 1 in last experiments, we see that both ASVM and ATSVM under the fixed parameter do not have consistent results in two aspects. One phenomenon is that different data sets favor different parameters. Another phenomenon is that even for the same data set, it is hard to tell which parameter is always better, as for most of the cases, the curves for different parameters intercross with each other many times. This phenomenon validates our analysis that using fixed regularization parameter is not a good choice for active learning. Although our method automatically selects its regularization parameters, it outperforms ASVM and ATSVM for any parameter in most cases. At the worst case, it is still comparable with the best result obtained by the other two methods. In our experiments, the compared methods yield similar results for the vote data set. This may be due to the reason that the SVM methods are not very sensitive to the regularization parameter in this case, as this data set can be easily classified. We can see in the corresponding figure that all methods get high accuracy with a small part of samples labeled. For the breast data set, both ASVMpath and ATSVM perform worse than ASVM. The reason is probably that the data distribution is not suitable for using the unlabeled data to boost the learning performance. Using the unlabeled data is not a good choice in this situation, if the unlabeled information is not properly consistent with the available labeled samples. However, in all the experiments, the best results for ASVM and ATSVM cannot be guaranteed without appropriate validation sets.

We also show the variance on some data sets for different methods in Fig. 3, which tells that ASVMpath is much more stable. Combined with previous results for learning accuracy, they indicate ASVMpath performs significantly better.

Among all the compared methods, the computational cost of ATSVM is the highest. The ASVM and our ASVMpath are much cheaper. We record the running time for the whole active learning

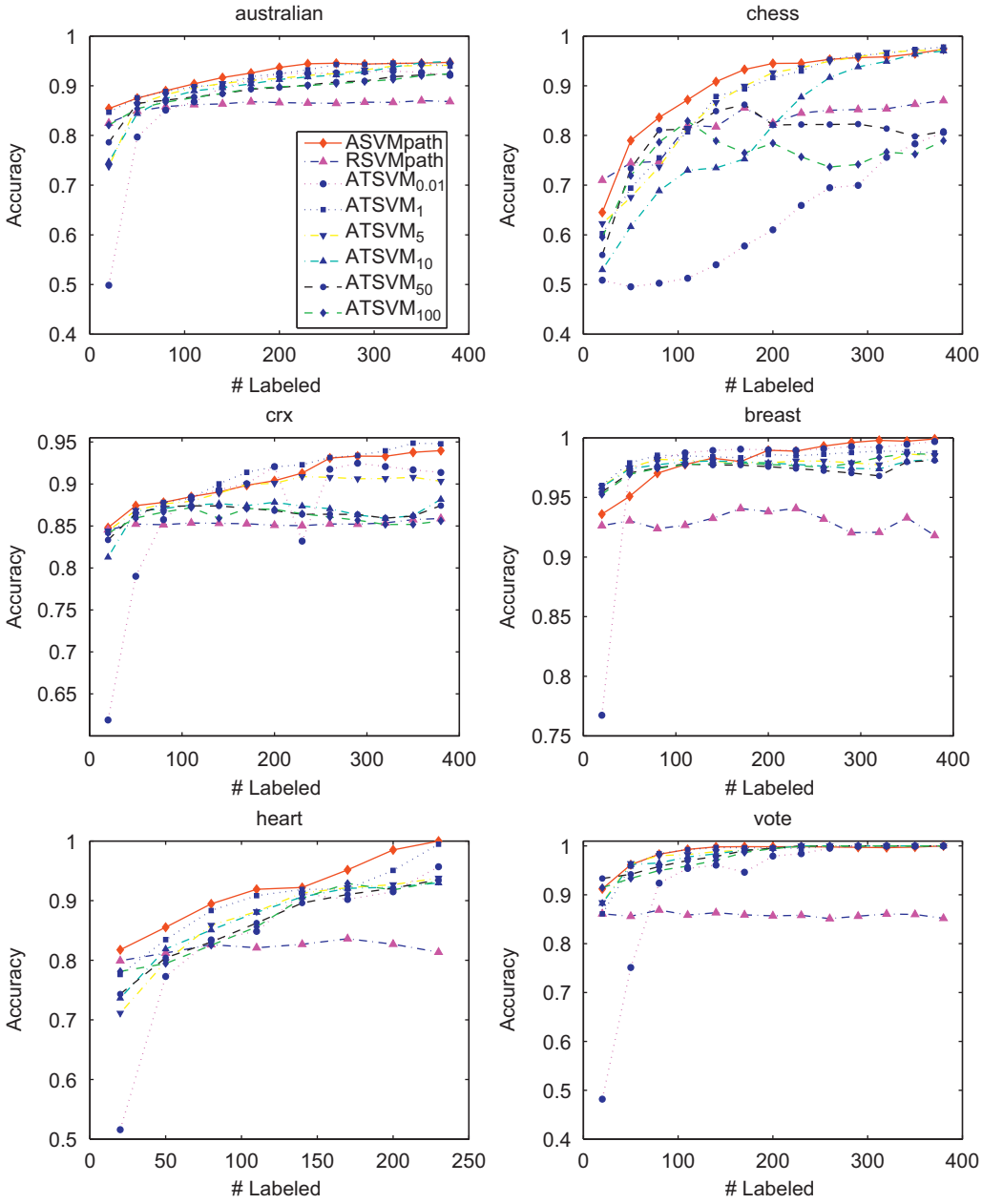


**Fig. 2.** Comparison results on UCI benchmarks for ASVMpath, RSVMpath and ATSVMs. The compared methods are the same for all data sets. We indicate the correspondence between the curve and the method in the first plot. Each curve represents the average result of 20 runs.

process for each method, based on computer simulated oracle. On the Australian data set, the running time is 8936 s for ATSVM, 84 s for ASVM, and 426 s for ASVMpath. We get these results under Matlab 2009a[7] on an Intel Core 2 Quad 2.50 GHz CPU with 8G RAM.

It should be noted that the compared experiments are unjust for ASVMpath. In fair cases, it is not possible to use the unseen labels to select the best parameter for either ASVM or ATSVM. However, we should introduce the validation method to do this job. In that situation, their current best test result can hardly be achieved. Besides, the validation will cost much longer running time for ASVM and ATSVM, by repeating the training procedure for each candidate parameter. For instance, if we use six candidate parameters for ASVM and ATSVM, their running time for the Australian data set should increase to about 500 and 53,600 s, respectively. In this sense, the parameter free algorithm ASVMpath becomes the most economic method.

### 5.4. Comparisons with ATSVM on SSL data

In these experiments, we use the artificial data sets which are proposed to demonstrate the performance of semi-supervised learning methods. They should be more suitable for TSVM. We only use linear version ASVMpath and ATSVM to conduct the following experiments. All other settings are the same as in the above comparisons. For better viewing, we only show the learning curves of ASVMpath and the ATSVM using best parameter.

Figs. 4 and 5 show the most representative comparison results. In Fig. 4, we can see the learning result is not optimal for ATSVM with fixed parameter, compared with the result reported in [3] for

[7] In previous experiments, ASVMpath is running in Matlab, while the SVM and TSVM for ASVM and ATSVM are implemented in C++ codes. Here, we implement all of them in Matlab for a fair comparison.

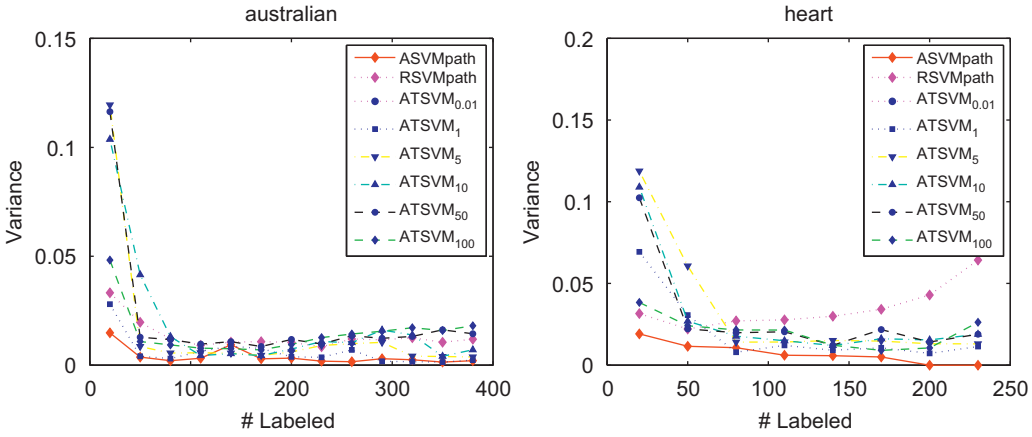


**Fig. 3.** Variance of the learning accuracy for different methods on UCI benchmarks. The variance of ASVM is very similar with that of ATSVM, so we only present it for ATSVM.

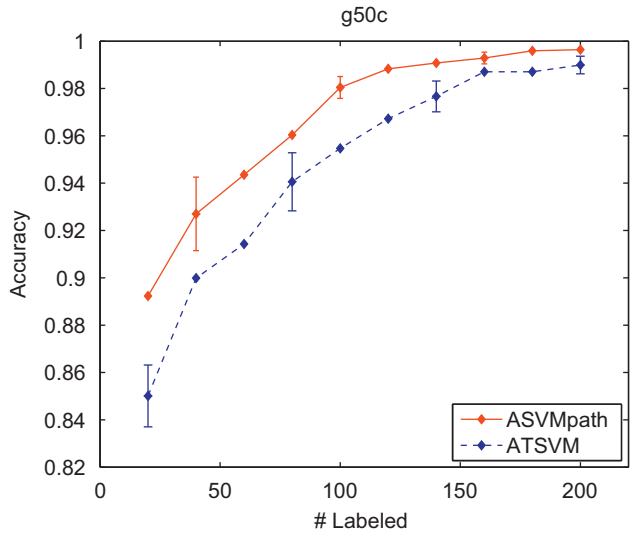


**Fig. 4.** The situation that ASVMpath outperforms ATSVM. The average result with error bars is given for each methods.

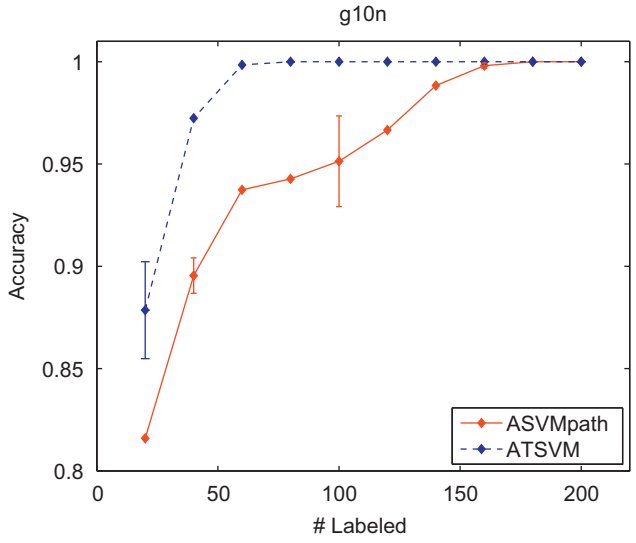


**Fig. 5.** The situation that ATSVM outperforms ASVMpath. The average result with error bars is given for each methods.

the same data set with the same number of labeled data. It means in this data set, the ATSVM using current query scenario performs worse than random queries. ASVMpath instead uses a simple and faster method but gets better results.

There is also the situation in Fig. 5, where ATSVM outperforms our ASVMpath algorithm. We observe that in this situation ATSVM performs much better from the beginning of the query process. In this situation, the TSVM model right fits the data, so it becomes a better choice. However, the active learning may not be in urgent need at this time, as the TSVM based on random sampling is good enough. These two experiments also show the consequence for choosing the improper method. Misuse of ATSVM will reduce the final learning accuracy, while incorrect use of our ASVMpath method will increase the label complexity. These two methods have different effects, as they use the unlabeled data in very different ways, which has been discussed in Section 4. If it is known beforehand the TSVM model fits the problem perfectly, TSVM is the better choice, and in this situation, active learning may also have less superiority. Otherwise, if the user is not sure which model the problem prefers, the ASVMpath method should be used.

## 6. Conclusions

In this paper, we studied the active learning problem based on soft margin SVM, and an adaptive regularization method was proposed to get consistently satisfactory learning results during the whole active learning process. To efficiently achieve this target, we used the regularization path algorithm to generate the candidate set and used the pseudo-margin on the unlabeled data as the measurement to validate the effect of different parameter values and continually choose the best one along the query process. We compared the proposed method with the active learning SVM and TSVM. The empirical results showed that our proposed algorithm outperformed those methods under the active learning scenarios. Moreover, it runs much faster than TSVM. These results suggested that more attention should be put on how to make the optimal use of the available information to find the best attainable learning result.

## Acknowledgements

This research was supported by National Natural Science Foundation of China (NSFC Grant No. 60835002, No. 61021063 and No. 61075004).

## References


[1] Y. Baram, R. El-Yaniv, K. Luz, Online choice of active learning algorithms, Journal of Machine Learning Research 5 (March) (2004) 255–291.

[2] C. Campbell, N. Cristianini, A. Smola, Alex: query learning with large margin classifiers, in: Proceedings of the International Conference of Machine Learning (ICML), 2000, pp. 111–118.

[3] O. Chapelle, A. Zien, Semi-supervised classification by low density separation, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2005.

[4] R. Collobert, F. Sinz, J. Weston, L. Bottou, T. Joachims, Large scale transductive SVMS, Journal of Machine Learning Research 7 (8) (2006) 1687–1712.

[5] D. Cox, F. OSullivan, Asymptotic analysis of penalized likelihood and related estimates, The Annals of Statistics 18 (4) (1990) 1676–1695.

[6] F. d'Alche Buc, Y. Grandvalet, C. Ambroise, Semi-supervised MarginBoost, in: Advances in Neural Information Processing Systems 14 (NIPS), MIT Press, Cambridge, MA, USA, 2002, pp. 553–560.

[7] T. Hastie, S. Rosset, R. Tibshirani, J. Zhu, The entire regularization path for the support vector machine, Journal of Machine Learning Research 5 (2004) 1391–1415.

[8] T. Joachims, Transductive inference for text classification using support vector machines, in: Proceedings of the International Conference of Machine Learning (ICML), 1999, pp. 200–209.

[9] T. Luo, K. Kramer, D. Goldgof, L.O. Hall, S. Samson, A. Remsen, T. Hopkins, Active learning to recognize multiple types of plankton, Conference of the International Association for Pattern Recognition (ICPR), vol. 3, 2004, pp. 478–481.

[10] Y. Lin, Support vector machines and the Bayes rule in classification, Data Minning and Knowledge Discovery 6 (3) (2002) 259–275.

[11] S.J. Pan, Q. Yang, A survey on transfer learning, Technical Report HKUST-CS08-08, Hong Kong University of Science and Technology, 2008.

[12] B. Settles, Active learning literature survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[13] M. Sugiyama, N. Rubens, Active learning with model selection in linear regression, in: Proceedings of the SIAM International Conference on Data Mining (SDM), 2008, pp. 518–529.

[14] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of Machine Learning Research 2 (2000) 999–1006.

[15] V. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.

[16] Z. Xu, R. Jin, J. Zhu, I. King, M. Lyu, Z. Yang, Adaptive regularization for transductive support vector machine, in: Advances in Neural Information Processing Systems (NIPS), 2009, pp. 2125–2133.

[17] K. Yu, J. Bi, T. Volker, Active learning via transductive experimental design, in: Proceedings of the International Conference of Machine Learning (ICML), 2006, pp. 1081–1088.



**Zheng Wang** received his BS degree from the Harbin Institute of Technology, China in 2004. He is currently a PhD candidate in the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, Beijing, China. His research interests focus on machine learning and its applications.

**Shuicheng Yan** received his PhD degree from the School of Mathematical Sciences, Peking University, 2004. He spent three years as Postdoctoral Fellow at the Chinese University of Hong Kong and then at the University of Illinois at Urbana-Champaign, Urbana, and he is currently an Assistant Professor in the Department of Electrical and Computer Engineering at the National University of Singapore. In recent years, his research interests have focused on computer vision (biometrics, surveillance, and internet vision), multimedia (video event analysis, image annotation, and media search), machine learning (feature extraction, sparsity/non-negativity analysis, large-scale machine learning), and medical image analysis. He has authored or coauthored over 140 technical papers over a wide range of research topics. Dr. Yan has served on the editorial board of the *International Journal of Computer Mathematics*, as Guest Editor of a special issue of *Pattern Recognition Letters*, and as a Guest Editor of a special issue of *Computer Vision and Image Understanding*. He has served as Co-Chair of the IEEE International Workshop on Video-oriented Object and Event Classification (VOEC'09) held in conjunction with ICCV'09. He is the special session chair of the Pacific-Rim Symposium on Image and Video Technology 2010. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology.

**Changshui Zhang** received his BS degree in mathematics from Peking University, Beijing, 1986 and PhD degree from Tsinghua University, Beijing, 1992. In 1992, he joined the State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Automation, Tsinghua University, where he is currently a professor. His interests include pattern recognition, machine learning, etc. He has authored more than 200 papers. He currently serves on the editorial board of the Pattern Recognition journal.