# A new semi-supervised support vector machine learning algorithm based on active learning

**Abstract:**

Semi-supervised support vector machine is an extension of standard support vector machine in machine learning problem in real life. However, the existing semi-supervised support vector machine algorithm has some drawbacks such as slower training speed, lower accuracy, etc. This paper presents a semi-supervised support vector machine learning algorithm based on active learning, which trains early learner by a spot of labeled-data, selects the best training samples for training and learning by active learning and reduces learning cost by deleting non- support vector. Simulative experiments have shown that the algorithm may get good learning effect at less learning cost.

SECTION I.

## Introduction

Currently, support vector machine (SVM) is a very popular research direction in machine learning, which is an effective learning machine that has been confirmed in experiments and has successful application in many areas. However, the traditional SVM is mainly used to deal with supervised learning problems, namely, which needs to mark sample data to train classifiers, while the large amount of data in real life which is unmarked and it is time-consuming for marking the data. This contributed to the study of machine learning into a new phase. Semi-supervised learning which has been combined with marked and unmarked data is becoming the new hot spot. The broad application prospect which has been led to by Semi-supervised learning in the field of pattern recognition and artificial intelligence systems is the main reason for such research heats up.

This paper attempts to research the characteristics of support vector machine and explores a way for improving the performance of classifiers which is a practical approach for classifying used a small number of labeled samples and a large number of unlabeled samples. The paper is organized as follows. Section II describes the related work and it includes the mathematical model of SVM, the concept of semi-supervised learning and typical algorithms and the concept and significance of active learning; The semi-supervised support vector machine (S3VM) for the status quo, the new algorithm features and implementation steps is discussed in section III; The experimental results of analysis of the algorithm are shown in section IV, followed by conclusions of this paper in Section V.

SECTION II.

## RELATED WORK

### A. Description of SVM

SVM is a kind of method that non-linear problem in low-dimensional space is mapped to a high-dimensional space so that a simple linear classification techniques can be dealt with is suitable for small sample learning [1]. SVM can effectively overcome the traditional method of over-fitting problems and neural network learning problems commonly seen in local minimum, so it has strong generalization ahility.

SVM can ensure that the non-linear problem mapped to high-dimensional space does not increase the computational complexity and also effectively overcomes the curse of dimensionality by introducing the kernel function and the associated dot product operation. SVM can be summed up for solving the quadratic programming problem. For a given labeled training data set:

Consider the general nonlinear situation: we define *ξi* as the relaxation variable, and then the mathematical expressions for the problem is,

The model of SVM has been provided that which is a powerful analytical tool for solving a variety of pattern recognition, function simulation and forecasting models. However, SVM can only be used in analysis of issues for marked sample data, while there are large amounts of data are unmarked in real-life, so that the scope of application of SVM has been limited. At present, an effective way to solve this problem is to use semi-supervised learning.

### B. Description of S3VM

Semi-supervised learning is an integrated way of learning which uses a small amount of labeled data to obtain an initial classifier, and then uses a large number of unlabeled data to further improve the performance of the initial classifier to eventually achieve the precision of learning. Although it is a new field of research, but also has obtained some initial results. The typical semi-supervised learning method will be used in SVM is transductive support vector machine (TSVM) learning method [2]. The principle of TSVM algorithm can be expressed as, given a group of independent and identically distributed labeled data sets:

### C. Active Learning

Active learning is a training process of a machine learning method which dynamically selects samples from the candidates according to a strategy when training [3]. We know from the previous section that the unlabeled training samples are used in TSVM, but we cannot draw the conclusions that more samples are used for training and then the learning performance will certainly be better. Actually, the unlabeled samples for the semi-supervised learning methods used in practical applications may come from different environments, the samples characteristics of distribution is complex and unknown, would most likely has the noise. Active Learning can make use of existing knowledge and take the initiative to select the most likely to solve the problem of the sample, effectively reduce the number of samples required for assessment.

SECTION III.

## The improved algorithm based on active learning

The classifier has a greater performance improvement when use algorithms of TSVM than a single classifier, while TSVM main drawback is that it needs to estimate the training sample distribution ratio of positive and negative and the positive number of samples in advance. And incorrect estimate will result in poor results. On this issue, Chen Yisong et al proposed an improved algorithm PTSVM [4]. The algorithm only is fitted for small number of unlabeled samples, although annotated paired and reset label to improve the performance of the TSVM. When there are more samples, this frequent label and label replacement will result in rapid increase in the complexity of the algorithm, and then the PTSVM algorithm is far more than the general TSVM algorithm. However, the unlabeled samples in most of the situation in practical applications are far more than labeled samples, thus we need to improve the algorithm. Zhong Qingliu et al against the problems that TSVM needs the prior accurate estimate of the distribution of positive and negative rates of the samples and PTSVM algorithm needs repeated calculations and training, so that it results in computational cost, they proposed a semi-supervised learning algorithm based on SVM and by gradual approach. However, this algorithm does not combine with the active learning, so cannot choose the best samples for study [5]. Chen Yaodong et al combined semi-supervised learning and active learning to study (in this paper referred to as the CSSA), but the just labeled samples have been increase into the training set, so the training set is slowly increasing. However, the more training samples, then the slower training speed. Therefore, inspired by the following two papers for its shortcomings, this paper presents a new semi-supervised support vector machine learning algorithm based on active learning (AS3VM).

In semi-supervised learning, since the initial labeled samples are relatively small, the initial classifier trained by the small samples may be classified ineffective, and lead to some unlabeled samples markup annotation errors. If we use these wrong annotations samples for training, it is clear that the classifier training has been ineffective. In addition, because we use SVM classifier for classification, not all of the samples work when training, and only a small amount training samples which is called support vector, and these support vector located at the geometric position around the hyperplane [6] (See in the figure 1.). In order to reduce the number of training samples to improve learning efficiency, we should try to select those who may be the samples of support vector to learn, this is the active learning strategies in this paper for choosing samples, and that is we will select the samples which is located at the geometric position around the hyperplane for training samples to learn. In S3VM learning, the final result is to obtain a better classifier and will label the unlabeled samples annotation to the correct category. Thus, when we need to know the S3VM training and classification result which is well or bad, we can judge it by means of final label of the samples and the correctness of category. However, in the second-class classification, the vast majority of samples can only belong to one of two categories, this is because that the boundaries of categories of some samples are little different such as news and sports categories, and sometimes a sample may also belong to two categories, so we can define a threshold T to the end of training by the last two results labeled and all samples needed labeled,

Where, A is the last two results number of samples of different labeled, B is the total number of samples should be labeled.

This is because that the value of T reflects the stability of the results of the samples categories labeled recently. When the smaller of the value of T, the greater consistency of, and the more credible of the classification results of the classifier, we can end the training. Otherwise, when the greater of the value of T, the more greatly of the results of the last two classification results vary, and the more not credible of the classification results of the classifier, we should continue to train. Therefore, the value of T can be used for the conditions to determine when we should end the training. Of course, for the above formula, we can see that when T is 0, the classifier is the best, but the boundaries of categories of some samples are little different and a sample may also belong to two categories sometimes, also we should reduce the training time, we can set a condition of ending training, which is the value of T is recognized.

In summary, the training algorithm of the main steps in this paper is:

1. Select the appropriate type of kernel functions and related parameters, and then train the initial samples which are labeled in initial S3VMlearning to obtain the initial classifier.
2. Label the all unlabeled samples used the obtained classifier, use the active learning strategies mentioned earlier to select the samples which may become support vector from them, and combine the initial labeled training set for a new training set.
3. Train the new set of training samples to obtain a new classifier again, and delete the non-support vector from the new training set and then it is a new training set again.
4. Label all the initial unlabeled samples used the obtained classifier, calculate the value of T, if the value of T is scheduled to meet the requirements, then go to step 5. Otherwise, use the active learning strategies mentioned earlier to select the samples which may become support vector from all samples, and combine the last training set for a new training set, then repeat step 3 to step 4.
5. End the training, output the results.

SECTION IV.

## Experiment and argumentation

In order to test the performance of AS3VM algorithm, we apply this algorithm into Chinese webpage classification. We compare the performance using algorithm of TSVM, PTSVM and CSSA.

In the experiment, the webpage of Chinese training text comes from sina web net. In the first experiment, the number of subject webpage is 500. And they belong to health and science and technology and they both have 250 web pages. Now, we select 50(both 25) pages to label for the initial labeled samples; we select 350(both 175) pages from the rest of pages for the unlabeled samples; and we label the rest of pages as the test set. In the second experiment, the number of subject webpage is 1000. And they also belong to health and science and technology and they both have 500 web pages. Now, we select 50(both 25) pages to label for the initial labeled samples; we select 850(both 425) pages from the rest of pages for the unlabeled samples; and then we label the rest of pages as the test set. In the third experiment, the number of subject webpage is also 1000. And they belong to health and science and technology and others, and health and science and technology both have 490 web pages and others have 20 web pages. Now, we select 50(both 25 in health and science and technology) pages to label for the initial labeled samples; we select 850(both 425 in health and science and technology and 20 in others) pages from the rest of pages for the unlabeled samples; and then we label the rest of pages as the test set. Of course, we should pretreat the web pages and transfer them to the form of vectors, then train the samples using the above-mentioned four kinds of algorithms for learning respectively. All of the algorithms are carried out in the Libsvm-mat-2.89–3 saddlebag [7]. The platform of the experiment is Pentium 4, 1.0GRAM, operating system is the Windows XP. In the experiment, kernel is REF function, C=1 T=5%. The results of the experiment are shown in table 1.

Where, A is the number of documents classified correctly, B is the number of the documents of the category.

From the figure 2, we can find that the four kinds of training algorithms are almost similar when the training set is small, otherwise, we can find that PTSVM algorithm is the most time-consuming and AS3VM algorithm is a little better than the rest of the training algorithms. From the figure 3, the performance of all four kinds of training algorithms in the second experiment is better in the first one and also better in the third one. Also, we can find that the performance of AS3VM algorithm is the best. So, in the overall, the performance of AS3VM algorithm is better than the other three kinds of algorithms from the experiment results of the above table and figures, specifically, the more of the training samples and the littler of the noise training samples, the better of the classification results. This is that because AS3VM algorithm reduces the training samples by the method of preselecting support vectors which are small-scale in all training set, and reduces the training time effectively; in addition, choosing the best samples, re-training every time and ending training with threshold T reduces the accumulated error and improves the accuracy rate of classification.

SECTION V.

## Conclusion

This paper presents a new semi-supervised support vector machine learning algorithm based on active learning, which would effectively overcome slow training speed and low training accuracy rate with the traditional TSVM algorithm, could adapt to the training sample in different distribution better and have the practical significance for promoting. Next, the research will be focused on how to choose better active learning strategies, how to make the algorithm to better adapt to the different distribution of training samples, as well as how to further improve the TSVM classification performance, especially, how to train the noise samples to meet the higher requirements of the classification problem.