Improved Margin Sampling for Active Learning

Jin Zhou and Shiliang Sun

Department of Computer Science and Technology
East China Normal University
500 Dongchuan Road, Shanghai 200241, China
Email: jinjin.zhou12@gmail.com, slsun@cs.ecnu.edu.cn

Abstract. Active learning is a learning mechanism which can actively query the user for labels. The goal of an active learning algorithm is to build an effective training set by selecting those most informative samples and improve the efficiency of the model within the limited time and resource. In this paper, we mainly focus on a state-of-the-art active learning method, the SVM-based margin sampling. However, margin sampling does not consider the distribution and the structural space connectivity among the unlabeled data when several examples are chosen simultaneously, which may lead to oversampling on dense regions. To overcome this shortcoming, we propose an improved margin sampling method by applying the manifold-preserving graph reduction algorithm to the original margin sampling method. Experimental results on multiple data sets demonstrate that our method obtains better classification performance compared with the original margin sampling.

Keywords: Active learning, Margin sampling, Support vector machine, Manifold-preserving graph reduction

1 Introduction

In machine learning, supervised models, such as support vector machines (SVMs) are commonly used in classification problems [1, 12], owing to their valuable generalization properties and the uniqueness of the solution. However, as any other supervised classifier, SVMs rely on the quality of labeled examples used for training. Therefore, the training examples should completely represent the surface-type statistical properties in order to allow the classifier to find the correct solution. Although people can easily get a large number of unlabeled examples, it usually needs much manual labor to label them, which can be expensive, difficult or time-consuming. Therefore, there is a need for procedures to find a suitable training set automatically, or semi-automatically.

In the machine learning literature, this approach is known as active learning. Active learning is a sampling process by actively selecting and labeling the most informative candidates from a large pool of unlabeled examples. Instead of randomly picking unlabeled examples, active learning selects the examples that are considered the most valuable and informative for human labeling [16]. Through this, a predictor trained on a small set of well-chosen examples can perform as

well as a predictor trained on a large number of randomly chosen examples [9, 3, 18].

There are mainly three classes of methods used in active learning [17]. The first class of active learning methods is large margin-based heuristics, for instance, the margin sampling (MS) strategy which relies on SVM specificities [2, 13, 10]. MS selects the unlabeled data which lies within the margin of the current SVM since these examples are the most likely to become new support vectors. The second class is committee-based heuristics, for example, entropy query-by-bagging (EQB) [5]. The committee members with different hypotheses about parameters are trained to label the unlabeled data. It tends to select for labeling the unlabeled examples where the disagreement among the classifiers is maximal. The third one relies on the estimation of the posterior probability distribution function of the classes [8]. It selects the examples for manual labeling based on the values of their posterior probabilities. For a binary problem, the selected examples are the ones which give the class membership probability closest to 0.5.

In this paper, we mainly focus on MS which is a state-of-the-art active learning method and has widely applied in many practical issues, such as text mining [14] and remote sensing image retrieval [15]. However, as stated in [17], one of the drawbacks of MS is that the method is optimal only when a single example is chosen each iteration. When several samples are chosen simultaneously, the structural information and distribution in the feature space are not considered. This will lead to a consequence that several samples lying in the same area close to the hyperplane are selected into the training set. However, these points possibly provide the same information, and thus there is no need to select all of them. More importantly, it will cause data redundancy which decreases the classification accuracy. Considering both the distribution structure and uncertainty of the selected examples is an effective way to overcome this shortcoming. Several active learning algorithms have been proposed. For example, Huang et al. [7] presented a principled approach, termed QUIRE, to combine the informativeness and representativeness of an instance and Nguyen [10] proposed a formal model for incorporation of clustering into active learning. In this paper, we propose an improvement of MS by applying an algorithm called manifold-preserving graph reduction (MPGR) [15] beyond the original MS method. MPGR is a simple example sparsification algorithm which takes the space connectivity among samples into account and simultaneously effectively removes outliers and noisy points. By using MPGR, we can construct a subset which represents the global structure of the original distribution of samples. Such a modification of MS can avoid oversampling on dense regions to a large extent. Previously, we have applied MPGR to a different context, that is, active learning with probabilistic models, and got good performance improvements [19].

The remainder of this paper proceeds as follow. In Section 2, we briefly introduce some background about MS. In Section 3, we describe our method which applies MPGR to the original MS method. In Section 4, we show the

experimental results on three real data sets to demonstrate the effectiveness of our method. Finally, we provide concluding remarks in Section 5.

2 Background

In this section, we brief reviews some background of margin sampling.

Margin sampling is specific to margin-based active learning algorithm which takes advantage of SVM geometrical properties [13]. An SVM uses a linear optimal hyperplane to discriminate classes, which is induced from the maximum margin principle between two classes [11]. For detailed information about SVM, see [1, 12].

As we all know, the distance to the separating hyperplane can straightforwardly estimate the classifier confidence on an unlabeled example. The nearer the distance of an example to the hyperplane is, the lower the classifier confidence on it is. That is to say, the more information the example possesses. Therefore, the points are the most likely to become new support vectors which fall within the margin of the current classifier. Given a labeled training set $L = \{(x_1, y_1), \cdots (x_m, y_m)\}$ $(x_i \in R^d)$ with the corresponding labels $y_i \in \{\pm 1\}$. The goal of MS is to choose the examples with the minimum distance to the decision boundary from a set of n unlabeled examples U $(n \gg m)$.

Consider a binary problem. The distance of a sample to the decision boundary is given by

$$f(q_j) = \sum_{i=1}^{m} \alpha_i y_i K(x_i, q_j) + b, \tag{1}$$

where K is a kernel matrix, which defines the similarity between the candidate q_j and the support vector x_i, α represents the support vector coefficient $(\alpha \neq 0)$, and y_i are the labels of the support vectors with the value $\{1, -1\}$. As to multi-class classification, we can just use one-vs-rest to convert the multi-class problem to multiple two-class problems.

Therefore, the candidate selected into the training set is the one respecting the condition

$$x' = \arg\min_{q_j \in U} |f(q_j)|. \tag{2}$$

Then x' and its ture label are is added into L and x' is removed from U simultaneously.

3 Our proposed approach

One of the drawbacks of the MS is that the method is optimal only when a single candidate is chosen per iteration. When selecting several examples simultaneously, the problem of oversampling on a small area is unavoidable. In order to remedy the problem, we propose an improvement of MS by considering the space connectivity and the distribution in the feature space of the unlabeled candidates.

3.1 MPGR

In machine learning, manifold assumption is an important assumption which indicates that samples in a small area have similar properties and thus their labels should also be similar. This assumption reflects local smoothness of the decision function which can alleviate the overfitting problems. In addition, sparse manifolds have significant advantages as follows: it can effectively eliminate the influence of noisy points and simultaneously accelerate the evaluation of predictors learned from the manifolds [15].

Manifold-preserving graph reduction (MPGR) is a simple but efficient graph reduction algorithm based on the manifold assumption [15]. For a graph, normally speaking, weights can measure the similarity of linked points. It means that the higher the weight is, the more similar the linked examples are. Here we introduce the definition of degree $d(p)$.

$$d(p) = \sum_{p \sim q} w_{pq} \tag{3}$$

where $p \sim q$ means that example p is connected with example q (the k-nearest-neighbor rule is used to construct the adjacency graph where k is set to 10 in this paper) and w_{pq} is their corresponding weight. The weight is defined as:

$$W_{pq} = \begin{cases} \exp(-\frac{||x_p - x_q||^2}{t\eta}), & \text{if } x_p, x_q \text{ are neighbors ,} \\ 0, & \text{otherwise .} \end{cases} \tag{4}$$

Here t is a parameter varying in {1,5,10}, and η is the mean of all the smallest distances between one point and its neighbors. If two examples are not linked, their weight is regarded as 0. $d(p)$ is generally used as a criterion to construct sparse graphs because of its simplicity. A bigger $d(p)$ means the example p has more information. That is to say, the example p is more likely to be selected into the sparse graphs.

Through the MPGR algorithm, we can construct a manifold-preserving sparse graph, which means that an example outside of the sparse graph has a high space connectivity with an example retained in it. Given a graph composed of all unlabeled examples, the manifold-preserving sparse graph is composed of the candidates which have a high space connectivity with the original unlabeled examples [15]. In other words, the subset constructed by MPGR is high representative and maintains a good global manifold structure of the original data distribution. In addition, when there are noisy examples and outliers in the training data, the MPGR algorithm can remove them effectively.

3.2 Improved margin sampling method

In this section, we introduce our method which applies MPGR to the original MS.

As mentioned above, there are some drawbacks in the traditional MS method, such as not exploiting the space connectivity and not considering the distribution

of the examples in the feature space. In order to overcome these shortcomings, we apply the MPGR algorithm to the original MS. We denote the new method as iMS. By exploiting the aforementioned MPGR, iMS tends to select the examples with high space connectivity, that is globally representative examples. As these examples are high representative, we can just use them to represent the whole data set to a large extent. Compared with the original MS, iMS considers the distribution and the manifold structure among the unlabeled data. Moreover, iMS can effectively eliminate the influence of noisy points which will be excluded due to the low space connectivity.

The difference between MS and iMS is on the scale of unlabeled examples needed to be queried. Assuming there are $s1$ unlabeled examples in total and the number of subset is $s2$ ($s1 \geq s2$). MS queries all the $s1$ unlabeled examples, while iMS is just querying $s2$ examples constructed by the MPGR algorithm. Since the subset takes into account the distribution and global structural information of unlabeled examples, it can effectively avoid the aforementioned oversampling in the same region in MS. It is important to notice that if $s1 = s2$, the iMS algorithm is identical to MS.

It can be seen that our method is a refinement of the original MS. Essentially, it consists of two steps. First, we construct a sparse subset by MPGR. Then we use MS to reselect unlabeled points using Eq. 2 from the subset. Thus it can not only reduce the number of unlabeled points to be queried, but also avoid oversampling on dense regions. Moreover it can effectively remove the noisy points and outliers from the candidate points .

4 Experiments

We evaluate our method on three real data sets which are the Ionosphere data set, the Vertebral Column (VC) data set and the Balance Scale (BS) data set. All the three data sets are publicly downloaded from the UCI Machine Learning Repository [1]. To demonstrate the generality of our method, these data sets include both binary classification and multi-class classification tasks.

4.1 Experimental Settings

The optimal parameters $\{C, \sigma\}$ (Gaussian RBF kernel is used) are found by grid search after five-fold cross-validation. C is a parameter which controlls the trade-off between the minimization of the number of misclassified training points and the maximization of the margin [6]. σ is the band-width parameter of Gaussian kernel $k(x, y) = \exp(-\frac{||x-y||^2}{2\sigma^2})$, which is optimized in the range of $\{2^{-3}, 2^{-2}, ...2^0, ...2^3\}$.

For each data set, the algorithms start with a small labeled data set L and select examples iteratively from the candidates set U. We have chosen here to consider the unlabeled set U from the training set ($U = [training\ set] - L$). To

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

show the effectiveness of our method, three methods are compared in this paper: iMS, MS and random selection. During each iteration, we select p points to add into L and accordingly reduce them from U, and then calculate the error rates on the test set T. The difference is that iMS selects the p points from a subset of U (constructed by MPGR), while MS selects them from the whole unlabeled data set U. As a baseline, the method of random selection is randomly picking those points.

In our experiments, we focus on the average error rates and the entire procedure has been repeated 15 times on each data set. In the following experiments, the size of the initial data set L is set to 10. Each algorithm adds five most relevant examples into L per iteration. Note that the size of the subset constructed by MPGR should not be too large. From [15], we can see that the classification accuracy often first increases and then decreases as the proportion of unlabeled examples retained increases.

4.2 Binary classification

The Ionosphere data set was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The data set includes 351 examples in total where each data point includes 34 features and an output attribute. It is a binary classification (good/bad) problem. The whole data set of 351 examples are randomly split into a training set (used for both L and U) of 271 examples and a test set T of 80 examples. The size of the subset constructed by MPGR is 100. Fig. 1 shows the performance comparison of iMS with MS and random selection on this data set.

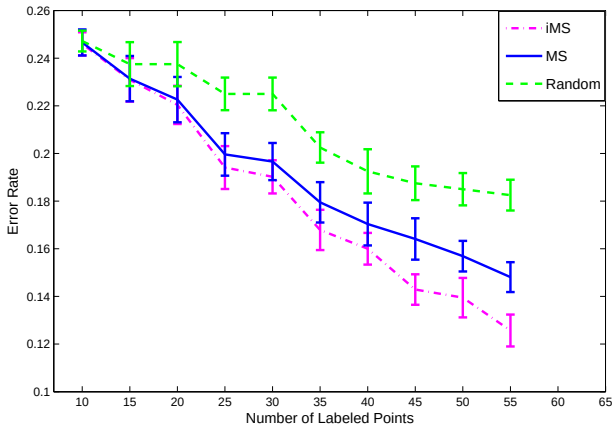


Fig. 1. Performance comparison of iMS, MS and random selection on the Ionosphere data set.

The VC data set contains six biomechanical features and an predicted attribute, which is used to classify orthopaedic patients into two classes (normal or abnormal). There are 310 examples in total which are randomly split into a training set of 230 examples and a test set of 80 examples. The subset size is 100. Fig. 2 shows the comparison results of iMS, MS and random selection.

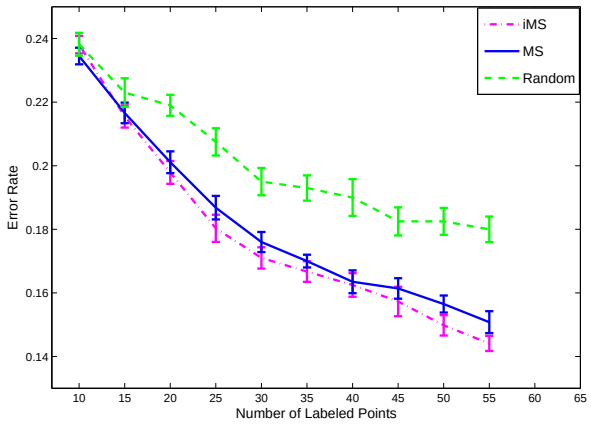


Fig. 2. Performance comparison of iMS, MS and random selection on the VC data set.

4.3 Multi-class classification

The BS data set is generated to model psychological experimental results. Each example is classified as: tip to the left (left), tip to the right (right), or to be balanced (balance). There are 625 examples in which each example contains four feature attributes and an output attribute. It is a multi-class classification problem. We set the subset size to be 150. The 625 examples are randomly split into a training set of 525 examples and a test set of 100 examples. Fig. 3 shows the performance comparison of iMS classification with MS and random selection on the BS data set.

The experimental results on the three data sets, which include binary classification and multi-class classification problems, show that our approach iMS obtains a better performance than MS and random selection. This might be due to the following reasons. Firstly, compared with MS, iMS constructs an important and informative subset which takes into account the global manifold structure and the distribution of the unlabeled examples. Secondly, by using MPGR, the influence of noisy points and outliers can be effectively eliminated.

5 Conclusions

In this paper, we applied the MPGR algorithm to an active learning method, the SVM-based MS and presented our improved new method iMS. Compared with the original MS, iMS is a refinement making use of the MPGR algorithm, which takes the distribution in the feature space and the structural space connectivity of the unlabeled candidates into account. Especially when there are noisy examples and outliers in the training data, the MPGR algorithm can effectively remove them. Consequently, oversampling on dense regions is avoided. Experimental results on multiple data sets show that our new method iMS outperforms MS and random selection. Extensions of the MPGR algorithm to other learning contexts will be interesting future research.

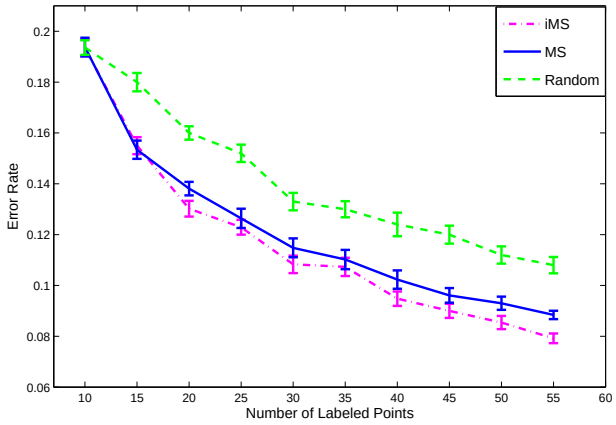
Acknowledgements

This work is supported by the National Natural Science Foundation of China under Projects 61370175 and 61075005, and Shanghai Knowledge Service Platform Project (No. ZF1213).

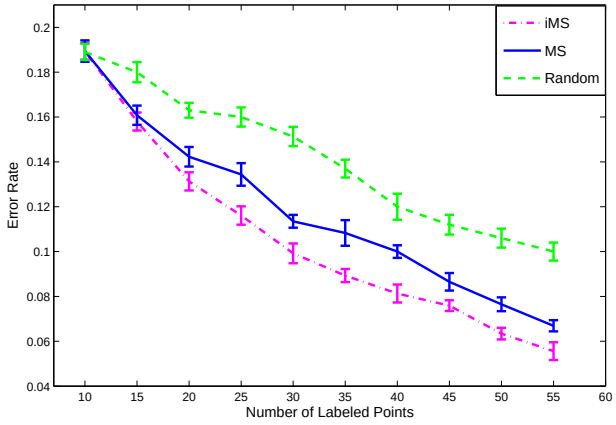
References

1. Boser, B.E., Guyou, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: 5th workshop on Computational learning theory, pp. 144-152. Pittsburgh (1992)
2. Campbell, C., Cristianini, N., Smola, A.: Query learning with large margin classifiers. In: 17th International Conference on Machine Learning, pp. 111-118. Stanford (2000)
3. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning. vol. 15, pp. 201-221 (1994)
4. Ferecatu, M., Boujemaa, N.: Interactive remote-sensing image retrieval image retrieval. IEEE Transactions on Geoscience Remote Sensing. vol. 45, pp. 818-826 (2007)
5. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning. vol. 28, pp. 133-168 (1997)
6. Hernández, E.P., Ambroladze, A., Taylor, J.S., Sun, S.: PAC-Bayes bounds with data dependent priors. The Journal of Machine Learning Research. vol. 13, pp. 3507-3531 (2012)
7. Huang, S., Jin, R., Zhou, Z.: Active learning by querying informative and representative examples. In: 24th Annual Conference on Neural Information Processing Systems, pp. 892-900. Vancouver (2010)
8. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Active learning with Gaussian processed for object categorization. In: 11th International Conference on Computer Vision, pp. 1-8. Rio de Janeiro (2007)
9. Mackay, D.J.C.: Information-based objective functions for active data selection. Neural Computation. vol. 4, pp. 590-604 (1992)
10. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: 21th International Conference on Machine Learning, pp. 623-630. Banff, Canada (2004)

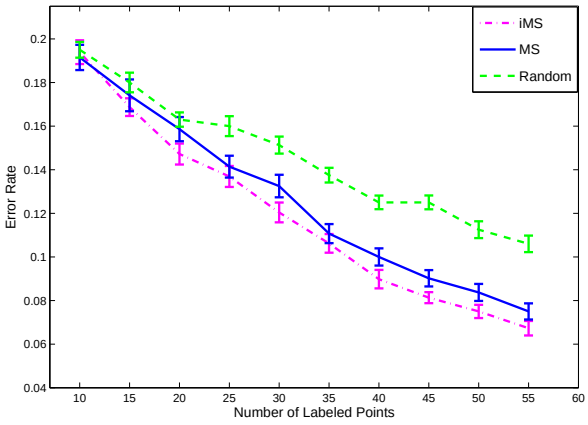
11. Oskoei, M.A., Hu, H.: Support vector machine-based classification scheme for myoelectric control applied to upper limb. IEEE Transactions on Biomedical Engineering. vol. 55, pp. 1956-1965 (2008)
12. Schölkopf, B., Smola, A.J.: Learning with Kernels. MIT press, Cambridge (2002)
13. Schohn, G., Cohn, D.: Less is more: Active learning with support vectors machines. In: 17th International Conference on Machine Learning, pp. 839-846. Stanford (2000)
14. Silva, C., Ribeiro, B.: Margin-based active learning and background knowledge in text mining. In: 4th International Conference on Hybird Intelligent Systems, pp. 8-13. Washington (2004)
15. Sun, S., Hussain, Z., Taylor, J.S.: Manifold-preserving graph reduction for sparse semi-supervised learning. Neurocomputing. vol. 124, pp. 13-21 (2013)
16. Sun, S., Hardoon, D.: Active learning with extremely sparse labeled examples. Neurocomputing. vol. 73, pp. 2980-2988 (2010)
17. Tuia, D., Ratle, F., Pacifici, F., Kanevski, M.F., Emery, W.J.: Active learning methods for remote sensing image classification. IEEE Transactions on Geoscience Remote Sensing. vol. 47, pp. 2218-2232 (2009)
18. Zhang, Q., Sun, S.: Multiple-view multiple-learner active learning. Pattern Recognition. vol. 43, pp. 3113-3119 (2010)
19. Zhou, J., Sun, S.: Active learning of Gaussian processes with manifold-preserving graph reduction. Neural Computing & Applications. DOI: 10.1007/s00521-014-1643-8 (2014)



(a) Left vs rest



(b) Right vs rest



(c) Balance vs rest

Fig. 3. Performance comparison of iMS, MS and random selection on the BS data set.