Spectral Clustering based Active Learning with Applications to Text Classification

Wenbo Guo , Chun Zhong ,Yupu Yang

Department of Automation, Shanghai Jiao Tong University, Shanghai，China

Abstract. Active learning is a kind of machine learning algorithms that spontaneously choose data samples from which they will learn. It has been widely used in many data mining fields such as text classification, in which large amounts of unlabelled data samples are available, but labels are hard to get. In this paper, an improved active learning algorithm is proposed, which takes advantages of the distribution feature of the datasets to reduce the labelling cost and increase the accuracy. Before the active learning process, spectral clustering algorithm is applied to divide the datasets into two categories, and instances located at the boundary of two categories are labelled to train the initial classifier. In order to reduce the calculation cost, an incremental method is added in the present algorithm. The algorithm is applied to several text classification problems. The results show it is more effective and more accurate than the traditional active learning algorithm.

1 Introduction

In the last few years, active learning [1] has become more and more popular because of its effectiveness, especially when dealing with the kind of learning tasks where class labels of each data sample are difficult to get and unlabeled data are sufficient or easy to collect. By applying active learning algorithms, the most informative samples are selected in order to learning the correct classifier with less labeled data samples. Another advantage of active learning is that by selecting the most informative data sample, the iteration times of the learning process can be reduced, thus the training time can be reduced. Also active learning introduces a select engine to select data samples, which makes active learning has a better generation performance, due to the violation of the assumption that training set and the test set are identically distributed.

In most situations of pool-based active learning scenarios, active learning deals with a small set of labeled data L and a large pool of unlabeled data U. With the use of few labeled instances, an initial model can be trained. Then in order to keep improving the model, the active learner selects a few unlabeled instances during each consecutive iteration which contain most information in term of the current model. The selected instances are used to update the classification model. This active learning process stops when some certain criterion is met. In this way, unnecessary and redundant samples are less likely to be included in the training set, thus greatly reducing the labeling cost and potentially the computational cost.

Active learning is an iteration process [2]. Each time a new instance is selected, the classification model will be updated until the stopping criterion is met. This process needs large amount of calculation. It is possible that after the first several iterations, the result changes little with the conduction of following iterations. What's more, because the instances are selected according to the current model, it happens that with the conduct of active learning, the selected instances may become useless, which increases the labeling cost. For example, if the current hyper-plane lies far away from the optimal one, the instances selected according to the current model will be useless for updating of the model and getting the correct hyper-plane. This cost dues ignoring the distribution feature of the training data.

Incremental learning method is applied, in this paper. When an instance (x, y) is labeled, the new classification model is updated based on only the new labeled instance (x, y), instead of training the former classification model. In order to cut down the labeling cost and make use of the information about the data distribution feature, a spectral clustering based active learning algorithm is proposed. Before the start of active learning, the whole datasets is clustered into two categories, and the instances located on the border of the two categories are picked to be the initial support vectors, and during the learning process, the points closest to the hyper plane will be chosen to be the new instance of the training set. The effect of this algorithm is show in the results of applying it to several text classification problems.

2 Active learning based on spectral clustering

2.1 Active learning with support vector machine

Active learning refers to machine learning algorithms, which autonomously select data samples. Thus any passive supervised learning method can be applied to train classification model. This paper focus on support vector machine (SVM), which has the only optimal solution with high accuracy. SVM learns a linear classifier, typical in a kernel-based feature space, which can be used to estimate the informativeness of data samples. This property make it perfectly suitable for active learning, because the data sample that is closest to the boundary is the most informative sample which will be labelled for the next iteration of learning.

2.1.1 Support vector machine

Support vector machine aims to learning a separating hyper plane classifier based on the idea of maximum margin. A SVM algorithm for binary classification has the input of N training data samples $\{(x_1, y_1),(x_2, y_2),\ldots,(x_N, y_N)\}$, where $x_i \in R^n$ and $y_i \in \{\pm 1\}$. The classification decision function of SVM is as followed.

$$f(x)=sign(\boldsymbol{w}^T \cdot \boldsymbol{x}+b) \tag{1}$$

where w and b determine the classification hyper plane.

As to linear support vector machine, the decision function can be obtain by solving the following optimization problem.

$$\min_{w,b,\xi_i} \frac{1}{2}\left\|\boldsymbol{w}\right\| + C\sum_{i=1}^{\ell}\xi_i \tag{2}$$

$$s.t. \quad y_i(\boldsymbol{w}\cdot \Phi(\boldsymbol{x})+b) \geq 1-\xi_i, \ \forall i \tag{3}$$

$$\xi_i \geq 0, \ \forall i \tag{4}$$

where $\Phi(\cdot)$ is a mapping from input space to feature space, $C > 0$ is penalty coefficient, which controls the penalty degree of misclassification and ξ_i is relaxation coefficient. The aim of the introduction of these coefficients is to deal with the problem, whose training set is linear inseparable.

In practice, this problem is a convex quadratic problem, which can be solved by optimizing its dual problem according to the Lagrange duality.

$$\max_{\alpha} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j) \tag{5}$$

$$s.t. \quad 0 \leq \alpha_i \leq C, \ \forall i \tag{6}$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0 \tag{7}$$

where α_i is Lagrange multiplier, and C is a constant chosen by user.

In order to solve the non-linear classification problem, the kernel trick is introduced to linear support vector machine. The widely used inner product as the mapping from input space to feature space is replaced by kernel function in non-linear support vector machine. The kernel function chosen by this paper is Gaussian kernel [3].

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2} \tag{8}$$

where x_i and x_j are two data samples and γ is the parameter of kernel function.

The classification decision function can be written by:

$$f(x) = sign(\sum_{i=1}^{N}\alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b) \tag{9}$$

This decision function can be obtained by changing the $\Phi(x_i)\ \Phi(x_j)$ in equation (5) with equation (8) and solving the dual problem (5) ~ (7). Thus active learning can use this non-linear SVM model as classification model and the follow function can be used to select the most informative data sample.

$$x = \arg\min_{x\in U}\left|\sum_{i=1}^{N}\alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right| \tag{10}$$

where U is unlabeled datasets.

2.2 Spectral clustering

Spectral clustering is a cluster algorithm based on graph theory. Datasets are classified into two categories based on the similarities among samples. The graph is $G = (V, E)$, where the V denotes the data samples, the E represents the similarities among two samples. Then clustering problem is transformed into a graph cut problem. To solve this problem means finding a method who can cut graph G into two categories, by maximizing the value of E inside the categories and minimizing the value of E between two categories.

The spectral clustering algorithm used in this paper is proposed in Table 1:

Table 1: Spectral clustering algorithm

Algorithm 1: Spectral clustering algorithm [3]
Input: Unlabeled datasets $X=\{x_i\}$.
Output: Label vector Y, where $y_i \in \{\pm 1\}$.
Step 1：Build the graph $G=(V, E)$，and calculate adjacent matrix W, where $W_{ij} = K_{ij}(x_i, x_j)$.
Step 2: Calculate the degree of each point, $d_i = \sum_j W_{ij}$ (11)
Step 3：Calculate diagonal matrix D, $D_{ij} = \begin{cases} 0, i \neq j \\ d_i, i = j \end{cases}$ (12)
Step 4：Calculate the Laplace matrix of G: $L_{sym} = I - D^{-1/2}WD^{1/2}$ (13)
Step 5: Apply eigenvalue decomposition to L_{sym} and the second less feature of L_{sym} can be obtained, written as v_2.
Step 6： $\hat{y} = sgn(v_2)$, and return $\hat{y}$.

2.3 Spectral clustering based active learning

Suppose a classification task on a datasets containing samples pertaining to two categories, In order to explore its distribution feature, reduce redundancy and increase the accuracy, the spectral clustering is applied on the datasets.

Prior to active learning, the spectral clustering method is applied to the datasets U, then the data points located at the border of the two categories are chosen to be the initial support vectors. During the learning process, the data points near the hyper plane are selected to be labeled and applied to the incremental learning. The final classification model can be obtained after consecutive iterations.

Details of the algorithm is proposed in Table 2.

Table 2: Active learning with spectral clustering

Algorithm 2: Active learning with spectral clustering
Input: Unlabeled datasets U, the number of initial labeled data instance k,
Output: An SVM classifier
Step 1: Divide U into two categories with the use of spectral clustering algorithm, where V_i is the index of which category X_i belongs to, if V_i <0, X_i belongs to the first category, if V_i >0, X_i be- longs to the second category, if $\|V_i\|\to 0$, X_i is located at the border of two categories.
Step 2: Choose k data points which have the least value of V for labeling, save them as datasets L.
Step 3: Learn a classification model $f^{(0)}$ from the labeled datasets L.
Step 4: Set t=0.
Step 5: Actively select the unlabeled instance X^* which is located nearest to the hyper plane, delete X^* from U and let: $$L = L \cup X^* \quad (14)$$
Step 6: Learn the new classification model $f^{(t+1)}$ from the former classifier $f^{(t)}$ and new labeled data X^*, that is: $$f^{(t)} + X^* \xrightarrow{S.V.M} f^{(t+1)} \quad (15)$$
Step 7: $t = t + 1$.
Step 8: Repeat step 5 ~ step 8 until the stopping criterion is met.
Step 9: Return final classification model $f^{(t)}$.

3 Application to text classification

In this paper, we apply this method to text classification problem, especially the widely used datasets: news20 and w2a. These two datasets come from the LIBSVM [4]. By comparing the result of Algorithm 2 with traditional active learning method, the effectiveness of it can be shown. The news20 contains 20,000 instances, each representing an item of news written in English. There are 20 different kinds of news. The w2a datasets has 3470 instances in training set and 46279 instances in testing set.

Linear support vector machine was applied to news20, for the instances in this datasets are linear separable. A single SVM can deal with binary classification task. SVM with Gaussian kernel was employed to be the classifier for w2a.In this experiment, γ is set to be 0.01 and the other parameter of SVM is set as C=100.

The traditional active learning support vector machines (ALSVM) are used for comparison. They are applied on the two datasets respectively. The dimension of the datasets is fixed during the active learning process.

Each experiment is repeated 100 times, and each time initialized with 5 selected instances. Both the traditional ALSVMs and proposed method use the same set of instances for initialization. In each active learning iteration, the data samples nearest to the current separating hyper plane are selected according to equation (10).

Average classification error rates of the datasets news20 with respect to the number of iterations are shown in the following two figures. In each figure, the blue line represents the traditional linear ALSVMs, the red line means the proposed algorithm. The results of iteration range from 20 to 51 times are shown, in which the difference of two method can be clearly distinguished.

Figure 1 shows that correct rates of proposed algorithm always beat the traditional ALSVMs. When it comes to the iteration 51, error rates of proposed algorithm are relatively 33% lower than the traditional ALSVMs

Figure 2 exhibits that error rates of proposed algorithm is lower than the traditional ALSVMs at the beginning, and as the process goes on, the traditional active learning method reaches almost the same correct rates with the proposed method.

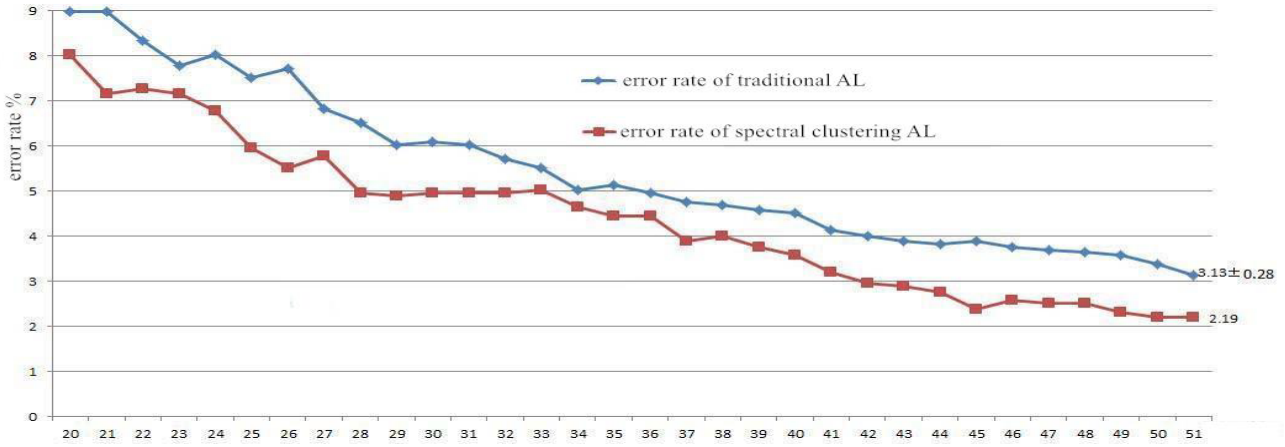

Figure 1. Classification error rates of No.11 and No.12 news in news20.

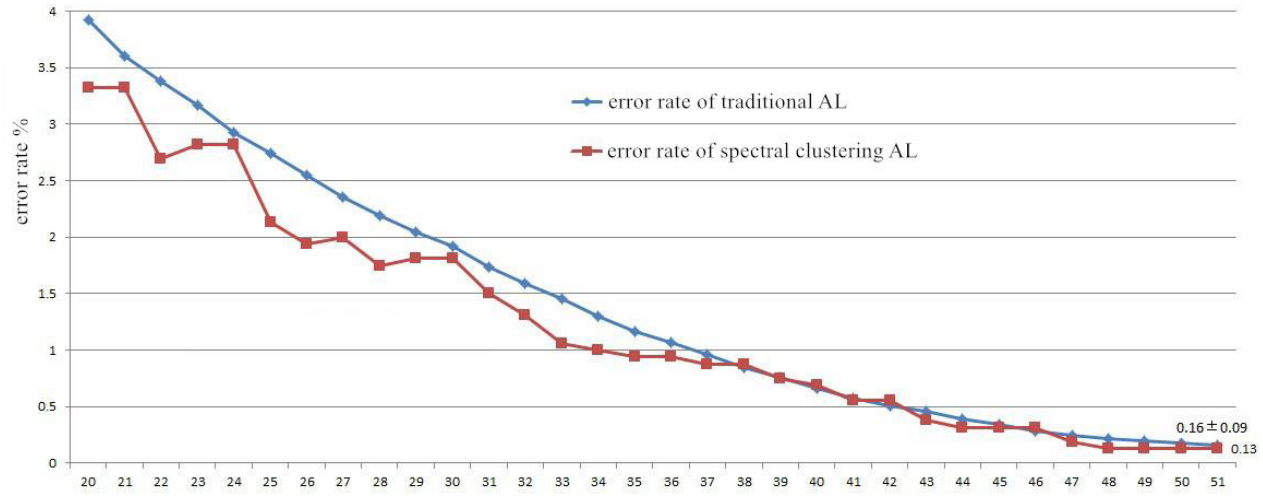

Figure 2. Classification error rates of No.7 and No.15 news in news20.

Figure 3. Classification error rates of No 19 and 20 news in news20.

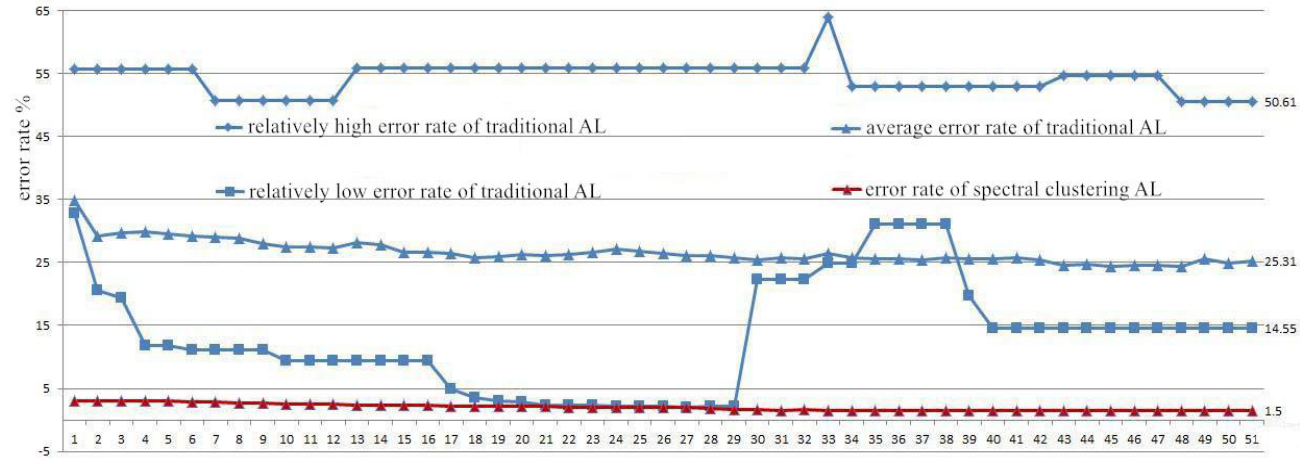

Figure 4. Classification error rates of datasets w2a.

Figure 3 shows that despite the lower error rates of traditional ALSVMs at the early iterations, the proposed algorithm eventually beats the traditional method. Also from the former three figures, the fluctuations of the results are caused by the shortage of active learning that the current selected instances may have less influence to the model as iteration goes on.

The traditional ALSVMs was applied to datasets w2a 100 times in this experiment, and a result with relatively high error rate, a result with relatively low error rate and average error rate are shown in the following figure, with the average error rates of proposed algorithm. The results of iteration range from 1 to 51 times are shown in Figure 4.

In the Figure 4, we can see that the error rates of spectral clustering active learning are lower than the traditional method and it has a faster convergence speed. The error rates of news20 are relatively lower than the w2a for

news20 was applied with linear SVM and has less noise. The Figure 4 also shows that the results of traditional active learning algorithm can reach an error rate up to 65%, and vary a lot, which indicate that traditional active learning is affected by initial states and has strong randomness. However, spectral clustering based active learning is less affected by initial states, and the results are more stable.

4 Conclusion

The experiment results indicate that the proposed spectral clustering active learning algorithm is superior to traditional ones，since the error rates are lower and the results are more stable. Taking advantages of the distribution feature of the datasets before selecting instances, the performance of active learning can be promoted.

Actually, the proposed method acts as a framework, spectral clustering is used for digging out the distribution feature of the datasets. As the datasets grows more complicated, more advanced clustering algorithm can be applied, such as Greedy Gradient Max-Cut (GGMC) [5], Sparse Subspace Clustering (SSC) [6], Spectral multi-manifold Clustering (SMMC) [7] and so on.

References

1. Schohn G, Cohn D, Less is more: Active learning with support vector machines, *International Conference on Machine Learning*, 839-846, (2000).
2. Tong S, *Active learning: theory and application*, Stanford University, (2001)
3. Zhou D, Bousquet O, Lal T N, et al, Learning with local and global consistency, *Advances in neural information processing systems*, **16**(16): 321-328, (2004)
4. Chang C C, Lin C J, LIBSVM: A library for support vector machines, ACM TIST , **2**(3): 27, (2011)
5. Wang J, Jebara T, Chang S F, Semi-supervised learning using greedy max-cut, J. Mach. Learn. Res. (JMLR), , **14**(1): 771-800, (2013)
6. Elhamifar E, Vidal R, Sparse subspace clustering: Algorithm, theory, and applications, ” IEEE Trans. Pattern Anal. Mach. Intell.(TPAMI), **35**(11): 2765-2781, (2013)
7. Wang Y, Jiang Y, Wu Y, et al, Spectral clustering on multiple manifolds, Neural Networks, IEEE Trans. Neural Netw., **22**(7): 1149-1161, (2011)