Active Learning with Sparse Reconstruction

Hanyi Zhang, Hanqing Lu, Ming Kong, Hanyin Fang, Zhou Zhao
College of Computer Science, Zhejiang University
and the Key Laboratory of Advanced Information Science and Network Technology of Beijing,
Beijing Jiaotong University, China
{3130102345, lhq110, zjukongming, fhy881229, zhaozhou}@zju.edu.cn

ABSTRACT

Active learning is a machine learning technique designed to select a few data points from the unlabeled data set to label, which helps to keep from expensive human labor. One reasonable direction among active learning approaches is to select the most representative data points. That means, the original data set can be approximated by linear combination of the selected points. However, traditional approaches unify the selection step and the reconstruction step into one framework, which leads to less precise results. In this paper, we propose a novel unsupervised framework named Active Learning with Sparse Reconstruction, taking the sparsity constraint into account. Specifically, we introduce l_1-norm regularizer into data reconstruction and separately do the selecting and reconstructing step. We further develop an efficient iterative gradient method to solve the optimization problem. Our empirical study shows encouraging results of the proposed algorithm in comparison to other state-of-the-art active learning algorithms.

CCS Concepts

•Computing methodologies → Active learning settings;

Keywords

active learning, data reconstruction, sparse coding

1. INTRODUCTION

In many real world scenarios, there is no shortage of unlabeled data but labeled data can be very time-consuming to get. The challenge is thus to reduce the number of labeled training examples while the quality of the trained model must be maintained, which asks researchers to determine which unlabeled samples are most informative. This paradigm is typically called active learning, which has been shown to benefit many real world domains such as image retrieval, document summarization and so on.

There has been a long history of research on active learning in machine learning community [5] [4] [1]. Recently, Yu et al. [6] have proposed Transductive Experimental Design which yielded impressive results on text categorization. However, most relevant approaches such as [2] unify the selection step and the reconstruction step into a reconstruction matrix A and select the most informative data points via A, which is inaccurate because the reconstruction contribution of a data point is ambiguous since it varies according to different data points it reconstructs. It is more reasonable to do the selection step and the reconstruction step separately so that we only allow selected data points to participate in reconstruction of the original dataset.

In this paper, we propose a novel unsupervised active learning algorithm called Active Learning with Sparse Reconstruction(ALSR). It selects the most informative samples in relation to the intrinsic geometrical structure of the data set. The underlying idea is that each data point can be approximately reconstructed by the linear combination of the selected data in its neighborhood. It is more reasonable to reconstruct it by using only its nearest neighbors since the points far away from the target point have little or even negative effect for the reconstruction [8]. Moreover, we introduce a data selection matrix to control each data point to be selected or not and punish the number of chosen data points to make sure that only a small number of data are selected. Our algorithm is unsupervised so that there is no need for labeled data to train a series of models in advance.

It is worthwhile to highlight several aspects of the proposed approach here:

1. We reformulate the problem of active learning from the viewpoint of separating the selection step and the reconstruction step in data reconstruction, which only allows selected data points to participate in reconstruction of the original dataset. Similarly, we update data selection matrix and data reconstruction coefficient matrix separately in our proposed algorithm.

2. We consider the process of active learning via sparse reconstruction over the composite objective function. We introduce l_1-norm onto the data selection matrix to enforce the sparsity of the selected data. The sparsity of the data selection matrix reduces the redundant or noisy data points.

3. We modify an iterative gradient algorithm of sparse coding, which can be conducted in parallel, to settle the proposed optimization problem. We evaluate the effectiveness of our approach in extensive experiments.

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

2. THE OBJECTIVE FUNCTION

Given a data matrix $X = (x_1, \cdots, x_m) \in \mathbb{R}^{n\times m}$, where m is the number of data and each data point x_i corresponds to a n-dimentional column vector, consider a selected data matrix $Z = (z_1, \cdots, z_m) \in \mathbb{R}^{n\times m}$, where z_k equals to x_k if data point k is selected or 0 if not selected, and the reconstruction coefficient matrix $A = (a_1, \cdots, a_m) \in \mathbb{R}^{m\times m}$. The goal of active learning is selecting the most informative data for labelling to reduce the number of labeled training examples while maintaining the quality of the trained model. That is, we want to determine the selected data matrix Z and the information loss is minimized. Inspired by [6], we propose a data selection criteria based on data reconstruction. Then the total information loss of the data reconstruction for all the data points is given by

$$\mathcal{L} = \sum_{i=1}^{m} \|x_i - Za_i\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm.

Different from other TED based approaches [2], here we introduce a data selection matrix $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$, where the i-th data is selected if and only if $\lambda_i = 1$. Thus we can represent the selected data matrix Z by $X\Lambda$. Now the total data reconstruction error can be rewritten as

$$\mathcal{L} = \sum_{i=1}^{m} \|x_i - X\Lambda a_i\|_F^2 = \|X - X\Lambda A\|_F^2, \tag{2}$$

By adjusting the data selection matrix Λ, we can minimize the data reconstruction error $\mathcal{L}$ and thus get the optimal selected data set $X\Lambda$.

Now we obtain the definite optimization problem, which is given by

$$\min_{\Lambda, A} \|X - X\Lambda A\|_F^2, \tag{3}$$

Then we will consider the distance punishment. For every two data points in data space, we denote D_{ij} to be the distance between data point x_i and x_j. In this paper, we set the distance to be Euclidean distance. For each data point x_i, we claim that it can only be reconstructed by its neighbors. Intuitively, the smaller D_{ij} is, the greater effect data j should have on the reconstruction of data i and vice versa. Thus, if the data point is far away from the target point participates in the reconstruction, it should get large punishment. We reformulate our objective function as follows:

$$\min_{\Lambda, A} \|X - X\Lambda A\|_F^2 + \beta \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}^2 D_{ij} \tag{4}$$

where β is a regularization parameter.

However, we observe that the problem is computationally intractable since the data selection vector λ is constrained to be an integer vector. Therefore we adopt a commonly used relaxation to allow the components of the vector λ to take real numbers. Then each component λ_j corresponds to a scaling factor which indicates how much information the j-th data point contains.

We also notice that compelling the λ to have more zero components implies that fewer data points are selected, which agrees with the intention of active learning. So we enforce the sparsity of the diagonal matrix Λ by employing the l_1-norm regularization. Then the objective function can be rewritten as following:

$$\min_{\Lambda, A} \|X - X\Lambda A\|_F^2 + \beta \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}^2 D_{ij} + \alpha\|\Lambda\|_1, \tag{5}$$

where α and β are two positive trade-off parameters to control the degree of penalty on sparsity and reconstruction distance. The l_1-norm regularization on the data selection vector $\|\Lambda\|_1$ controls the number of selected data points. The regularization coefficient α is used to ensure the selection matrix Λ is suitable for data selection. The larger the value of α is, more sparse the vector λ will become.

3. THE OPTIMIZATION

In this section, we design a gradient method to solve Problem (5). Inspired by the iterative optimization method in [3], we divide the gradient method into two steps: learning the data selection matrix Λ while fixing the reconstruction coefficient matrix A and learning reconstruction coefficient matrix A while fixing the data selection matrix Λ.

3.1 Learning the Sample Selection Matrix Λ

We first discuss how to optimize the objective function by adjusting the data selection matrix Λ while fixing the reconstruction coefficient matrix A. Now Problem (5) can be reduced as follows:

$$\min_{\Lambda} \|X - X\Lambda A\|_F^2 + \alpha\|\Lambda\|_1, \tag{6}$$

which is a l_1-norm regularized optimization problem. Due to the fact that Problem (6) is non-differentiable when some components of data selection vector λ equal to zero, we adopt an optimization method based on coordinate descent to solve the problem.

We rewrite the optimization problem as follows:

$$\min_{\lambda_p} f(\lambda_p), f(\lambda_p) = \sum_{i=1}^{n}\sum_{j=1}^{m} \left(r_{ij}^{-p} - \lambda_p x_{ip} a_{pj}\right)^2 + \alpha|\lambda_p| \tag{7}$$

where $r_{ij}^{-p} = x_{ij} - \sum_{k\neq p} \lambda_k x_{ik} a_{kj}$. The residue r_{ij}^{-p} is a constant given the variable p.

When we optimize the variable λ_p for the p-th data point, we keep other variables $\{\lambda_i\}_{i\neq p}$ fixed. We define $h(\lambda_p) = \sum_{i=1}^{n}\sum_{j=1}^{m} \left(r_{ij}^{-p} - \lambda_p x_{ip} a_{pj}\right)^2$, and $f(\lambda_p) = h(\lambda_p) + \alpha|\lambda_p|$. Based on the definition of sub-gradients [3], the sub-differential value $\frac{\partial|\lambda_p|}{\partial\lambda_p}$ is a sub-gradient of $|\lambda_p|$ if and only if

$$|\, \lambda_p' \,| \geq |\lambda_p| + \frac{\partial|\lambda_p|}{\partial\lambda_p}(|\, \lambda_p' \,| - |\lambda_p|) \tag{8}$$

Then the condition of getting the optimal value of $f(\lambda_p)$ can be written as following:

$$\begin{cases} \frac{\partial}{\partial\lambda_p} h(\lambda_p) + \alpha sign(\lambda_p) = 0 & \text{if } |\lambda_p| > 0 \\ |\frac{\partial}{\partial\lambda_p} h(\lambda_p)| \leq \alpha & \text{if } \lambda_p = 0 \end{cases} \tag{9}$$

Thus, we can remove the l_1-norm on λ_p by replacing $|\lambda_p|$ with either λ_p if $sign(\lambda_p) = 1$, $-\lambda_p$ if $sign(\lambda_p) = -1$ or 0 if $\lambda_p = 0$. In this way, Problem (7) can be reduced to a standard unconstrained quadratic optimization problem, which can be solved by the existing optimization methods.

In the algorithm, we maintain an active set S for potentially nonzero coefficients of λ (i.e. $S = \{p|\lambda_p = 0, |\frac{\partial}{\partial\lambda_p} h(\lambda_p)| > \alpha\}$) and their corresponding signs $\theta = [\theta_1, \theta_2, \cdots, \theta_k]$. In each active step, the variable λ' whose violation of the optimality condition $|\frac{\partial}{\partial\lambda_p}h(\lambda_p)| > \alpha$ is the largest is selected and added to S. Then a series of feature-sign steps are proceeded: at each iteration, given the active set S and signs θ, it computes the new analytic solution λ^{new} to the unconstrained quadratic problem for all p in S:

$$\lambda_p^{new} = \frac{\sum_{i=1}^n \sum_{j=1}^m r_{ij}^{-p} x_{ip} a_{pj} - \frac{\alpha\theta_p}{2}}{\sum_{i=1}^n \sum_{j=1}^m x_{ip}^2 a_{pj}^2}$$

Then an efficient discrete line search between the current value λ to the new value λ^{new} is invoked. Afterwards, two optimality conditions are checked and the data selection matrix Λ is returned if all the variables meet the conditions. The procedure of the algorithm is as follows:

1. For each λ_p, search for its sign θ_p
2. Solve the reduced unconstrained quadratic optimization problem to get the optimal $\lambda_p^\star$ which minimizes the objective function.
3. return the optimal data selection matrix $\Lambda^\star = diag(\lambda^\star)$

3.2 Learning the Reconstruction Coefficient Matrix A

We now describe the method of learning the coefficient matrix A while fixing the data selection matrix Λ. The problem becomes an unconstrained least squares problem as following:

$$\min_A \|X - X\Lambda A\|_F^2 + \beta \sum_{i=1}^m \sum_{j=1}^m a_{ij}^2 D_{ij} \tag{10}$$

We use coordinate descent method to optimize A and the solution to Problem (10) is given by

$$a_{pq} = \frac{\sum_{i=1}^n \lambda_p x_{ip} x_{iq} - \sum_{i=1}^n \lambda_p x_{ip} (\sum_{k\neq p}^m x_{ik} \lambda_k a_{kq})}{\sum_{i=1}^n \lambda_p^2 x_{ip}^2 + \beta D_{pq}} \tag{11}$$

We update every a_{pq} in turn until it is converged.

Our ALSR algorithm is summarized in Algorithm (1).

Algorithm 1 Active Learning with Sparse Reconstruction

```
Input: the data set of m data points $X = [x_1, x_2, \cdots, x_m]$,
    the parameters $\alpha$
1: Initialize reconstruction coefficient matrix $A_0$ randomly
   and data selection matrix $\Lambda_0$ by zero matrix, k=1
2: repeat
3: Update $\Lambda_k$ with algorithm in 3.1
4: Update $A_k$ with Equation (11)
5: k:=k+1
6: until it is converged
7: return $\Lambda$
```

4. EXPERIMENT RESULTS

In this section, we evaluate the effectiveness of our proposed unsupervised active learning method. The experiments is conducted on three datasets from IDA Benchmark repository: banana, twonorm and waveform [1]. We compare our method with random selection and other state-of-the-art active learning methods. All the methods used to compare in the experiment are listed as follows:

1. Random Sampling method which randomly selects training samples from the data set.
2. Simple Margin [5] method which selects the examples which are closest to the decision boundary of the classifier, using the hinge loss. [2]
3. Sequential Transductive Experimental Design (Sequential TED) [7] method which greedily selects the examples which minimize the loss.
4. Active Learning based on Neighborhood Reconstruction (ALNR) [2] method which is based on Transductive Experimental Design and takes the local geometric structure of the data set into account.
5. Active Learning with Sparse Reconstruction (ALSR) method which is proposed in this paper. The algorithm selects the data points according to the ordered values of the optimal data selection vector λ.

4.1 Performance Evaluation

4.1.1 Banana Dataset

In Banana dataset, we apply each of 5 active learning algorithms to select k (=5, 10, 15, ..., 75) samples from the training set and conduct a SVM training to get a classifier for each algorithm. Then we use the data from the test set to get a classification accuracy of 5 classifiers. Table 1 demonstrate the performance of 5 active learning methods on banana dataset. Fig.1(a) shows the average classification accuracy versus the number of selected samples by using SVM as the classification algorithm. The result demonstrates the efficiency of our algorithm. When there are only 5 selected training examples, the test accuracy of all the 5 algorithm is relatively low. However, our algorithm converges super fast: when the number of selected training examples reaches 15, our algorithm starts to converge and the accuracy far outweigh the rest of other algorithms.

4.1.2 Twonorm Dataset

On this dataset, we apply each active learning algorithm to select k (=2, 4, 6, ..., 30) training samples. The average classification accuracy is shown in Fig.1(b). As is shown in the picture, our algorithm outperform all the other algorithms when the number of selected training samples is pretty small. As the number of selected samples increases, the accuracy of all of the algorithms increase. Moreover, the convergence point of our algorithm is slightly ahead of the ALNR and Sequential TED algorithm while random sampling starts to converge much later. Table 2 also demonstrate the performance of 5 active learning methods on twonorm dataset.

4.1.3 Waveform Dataset

As before, we apply each active learning algorithm to select k (=3, 6, 9, ..., 36) training samples. The results are shown in Table 3 and Fig.1(c). As we can see, Sequential TED, ALNR and our algorithm our algorithm outperform the random sampling in most cases. These three algorithms perform comparably but our algorithm converges ahead of the other two algorithms.

[1] http://mldata.org/repository/data/viewslug/

[2] In the experiment, we first randomly select 10 samples to perform a SVM training. Then we use the classifier we obtain to select data that are closest to the decision boundary.

Table 1: Accuracy on Banana dataset

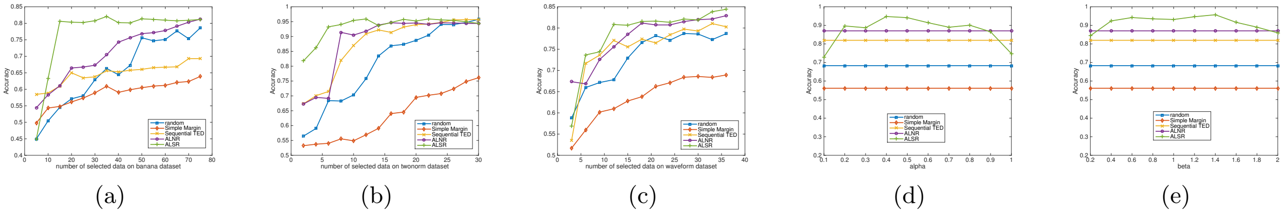
	10 Samples	20 Samples	30 Samples	40 Samples	50 Samples
Random	0.504 ± 0.032	0.571 ± 0.066	0.629 ± 0.080	0.644 ± 0.075	0.756 ± 0.057
Simple Margin	0.543 ± 0.070	0.562 ± 0.104	0.589 ± 0.107	0.591 ± 0.094	0.605 ± 0.085
Sequential TED	0.589 ± 0.031	0.650 ± 0.035	0.638 ± 0.028	0.653 ± 0.021	0.660 ± 0.034
ALNR	0.583 ± 0.058	0.665 ± 0.069	0.673 ± 0.057	0.743 ± 0.051	0.769 ± 0.037
ALSR	**0.633 ± 0.057**	**0.803 ± 0.083**	**0.808 ± 0.059**	**0.802 ± 0.040**	**0.814 ± 0.034**

Table 2: Accuracy on Twonorm dataset

	4 Samples	8 Samples	12 Samples	16 Samples	20 Samples
Random	0.590 ± 0.174	0.683 ± 0.164	0.758 ± 0.154	0.869 ± 0.094	0.887 ± 0.120
Simple Margin	0.537 ± 0.103	0.555 ± 0.060	0.569 ± 0.072	0.640 ± 0.088	0.695 ± 0.087
Sequential TED	0.701 ± 0.216	0.819 ± 0.166	0.909 ± 0.067	0.914 ± 0.079	0.940 ± 0.028
ALNR	0.695 ± 0.215	0.871 ± 0.021	0.918 ± 0.049	0.946 ± 0.006	0.945 ± 0.007
ALSR	**0.863 ± 0.028**	**0.940 ± 0.044**	**0.959 ± 0.013**	**0.947 ± 0.007**	**0.952 ± 0.008**

Table 3: Accuracy on Waveform dataset

	6 Samples	12 Samples	18 Samples	24 Samples	30 Samples
Random	0.660 ± 0.163	0.678 ± 0.048	0.766 ± 0.040	0.771 ± 0.067	0.786 ± 0.043
Simple Margin	0.559 ± 0.172	0.610 ± 0.149	0.638 ± 0.121	0.671 ± 0.073	0.686 ± 0.073
Sequential TED	0.716 ± 0.037	0.771 ± 0.049	0.774 ± 0.061	0.784 ± 0.078	0.793 ± 0.050
ALNR	0.668 ± 0.160	0.752 ± 0.042	0.812 ± 0.033	0.807 ± 0.019	**0.819 ± 0.017**
ALSR	**0.737 ± 0.039**	**0.808 ± 0.071**	**0.815 ± 0.073**	**0.814 ± 0.060**	0.819 ± 0.068



(a) (b) (c) (d) (e)

Figure 1: Subfigure (a)(b)(c) show the average classification accuracy versus the number of training samples on banana, twonorm and waveform dataset respectively. Subfigure (d)(e) show the performance of ALSR versus the parameters α and β on twonorm dataset.

4.2 Parameters Selection

There are two essential parameters, α and β, in our approach. The former one is the l_1-norm regularizer, which controls the sparsity of the data selection vector λ. The latter one is the distance penalty regularizer, which is used to control the locality. In this part, we examine how the performance of our proposed algorithm varies with the parameters α and β separately. We vary the value of α from 0.1 to 1 and β from 0.2 to 2. The experiment is conducted on twonorm dataset and we select top 8 informative samples to do a SVM training for a classifier of each algorithm, after which a validation test is executed.

When β is fix to be 0.6, the impact of α for the algorithm performance is shown in Fig.1(d), where we can see our ALSR achieves good performance when α varies from 0.2 to 0.8. Fig.1(e) also shows the experimental results with α fixed to be 0.5 and β varying from 0.2 to 2, where ALSR has a consistent good performance between 0.4 and 1.8.

5. CONCLUSIONS

We reformulate the problem of active learning from a new view point of sparse reconstruction. Comparing to previous active learning approaches such as Sequential TED and ALNR, our proposed approach explicitly separate the selecting step and the reconstruction step and take the local manifold structure into account. Therefore, the selected points by using our approach can improve the classifier the most if they are used as training samples. Experimental results on two datasets show the effectiveness of our approach.

6. ACKNOWLEDGEMENTS

The project is supported by the Key Laboratory of Advanced Information Science and Network Technology of Beijing (XDXX1603).

7. REFERENCES

[1] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[2] Y. Hu, D. Zhang, Z. Jin, D. Cai, and X. He. Active Learning via Neighborhood Reconstruction. *Proceedings of the 23th International Joint Conference on Artificial Intelligence*, pages 1–7, June 2013.

[3] H. Lee, A. Battle, R. Raina, and A. Y. Ng. Efficient sparse coding algorithms. *NIPS*, pages 801–808, 2006.

[4] H. S. Seung, M. Opper, and H. Sompolinsky. Query by Committee. *COLT*, pages 287–294, 1992.

[5] S. Tong and D. Koller. Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research ()*, 2001.

[6] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. *ICML*, pages 1081–1088, 2006.

[7] K. Yu, S. Zhu, W. Xu, and Y. Gong. Non-greedy active learning for text categorization using convex ansductive experimental design. *SIGIR*, pages 635–642, 2008.

[8] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, and T. S. Huang. Active Learning Based on Locally Linear Reconstruction. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2026–2038, Oct. 2011.