Active Learning with Maximum Density and Minimum Redundancy

Yingjie Gu[1,2], Zhong Jin[1], and Steve C. Chiu[2]

[1] Computer Science and Engineering,
Nanjing University of Science and Technology, Nanjing 210094, China
csyjgu@gmail.com, zhongjin@njust.edu.cn
[2] Department of Electrical Engineering,
Idaho State University, Pocatello 83209-8060, USA
chiustev@isu.edu

Abstract. Active Learning is a machine learning technique that selects the most informative examples for labeling so that the classification performance would be improved to its maximum possibility. In this paper, a novel active learning approach based on Maximum Density and Minimum Redundancy (MDMR) is proposed. The objective of MDMR is to select a set of examples that have large density and small redundancy with others. Firstly, we propose new methods to measure the density and redundancy of examples. Then a model is built to select examples by combining density and redundancy and dynamic programming algorithm is applied to solve the problem. The results of the experiment on terrain classification have demonstrated the effectiveness of the proposed approach.

Keywords: active learning, classification, density, redundancy.

1 Introduction

In many real-world applications, there are large numbers of unlabeled data while the labels are expensive and difficult to get. And much redundant data, which slows down the training process without improving the classification result, also exist in the training set. Active learning [1] was proposed to select the most informative examples for labeling and training a classifier, thus the labels of testing examples can be predicted most precisely. The kernel problem of active learning is how to measure the value of each example and how to select the most informative examples from the unlabeled data set.

There are many criteria in active learning for examples selection. Uncertainty sampling is one of the most widely used criterion that queries the examples whose labels are most uncertain under the current trained classifier. The most popular uncertainty sampling is SVM_{active} [2], which selects the examples nearest to the current decision boundary. Other criteria like variance reduction [3], density [4], and diversity [5] also have been widely applied to active learning.

Optimum Experimental Design (OED) [6], which refers to the problem of selecting examples for labeling in statistics, has attracted an increasing amount

C.K. Loo et al. (Eds.): ICONIP 2014, Part I, LNCS 8834, pp. 103–110, 2014.

of attention [7] [8]. The example $\mathbf{x}$ is referred to as experiment and its label y is referred to as measurement. OED tries to select examples so that the variances of a parameterized model are minimized. OED has two types of criteria. One is D, A, and E-Optimal Design that choose data points to minimize the variance of the model's parameters. The other is I and G-optimal Design that minimize the variance of the prediction value.

Active learning based on OED selects the most informative points while it is unable to exploit the redundancy between selected points. In this paper, we proposed an active learning algorithm called MDMR to select a set of points with maximum density and minimum redundancy. By combining examples' density and redundancy, every selected example is informative and the redundancy between selected examples are small.

The rest of this paper is organized as follows: In Section 2, we elaborate the proposed active learning approach MDMR. The experimental settings and results are presented in Section 3. Finally, we discuss the conclusion and future work in Section 4.

2 Active Learning with Maximum Density and Minimum Redundancy

The general problem of active learning can be described as follows. Given a set of points $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where each $\mathbf{x}_i$ is an instance of d-dimensional vector, find a subset $\mathcal{Z} = \{\mathbf{z}_{s_1}, \mathbf{z}_{s_2}, ..., \mathbf{z}_{s_k}\} \subseteq \mathcal{X}$, which contains the most informative points. In other words, if the points $\mathbf{z}_{s_i}(i = 1, 2, ..., k)$ are labeled and used as training data, the labels of testing data can be predicted most precisely.

In this section, a novel active learning algorithm is proposed to select examples by considering examples' density and redundancy.

2.1 Density and Redundancy

Information density is an important criterion for active learning since examples in dense regions are expected to be representative and informative. Thus we aim to select a set of examples that have large density. Firstly, we use Gaussian kernel to construct a complete graph with all unlabeled examples. The weight W_{ij} between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as

$$W_{ij} = exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}) \tag{1}$$

where σ is the parameter of gaussian kernel. As shown in (1), W_{ij} is large if $\mathbf{x}_i$ and $\mathbf{x}_j$ are very close to each other. The large weight W_{ij} means $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly connected, or they have large similarity.

For an example in dense region, it should be very close to its neighbors, which means the weight between the example and its neighbors should be large. Therefore, the average weight between an example and its p-nearest neighbors

is able to measure the density of the example. The density of $\mathbf{x}_i$ is defined as follows:

$$den(\mathbf{x}_i) = \frac{1}{p} \sum_{\mathbf{x}_j \in N_p(\mathbf{x}_i)} W_{ij} \tag{2}$$

Where $N_p(\mathbf{x}_i)$ is the p-nearest neighbors of $\mathbf{x}_i$.

Density-based active learning is able to select the most representative examples, but it is unable to exploit the redundancy between the selected examples. In other words, some selected examples may have similar information. Hence each example has maximum information can't guarantee the global information is maximum. Here we exploit the redundancy among the selected examples.

The examples have large weight are usually highly connected to each other. They probably have more redundant information than the examples whose weight is small. So the selected examples are required to have small weight with each other. Here, the maximum weight between an example and other selected examples are used to measure the redundancy of the example. If the maximum weight is very small, the example has little redundancy with other selected examples.

Suppose we have selected a set of k examples $Z_k = \{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, ..., \mathbf{x}_{s_k}\}$ from $\mathcal{X}$. The redundancy between example $x_{s_i} (i > k)$ and Z_k can be described as follows:

$$red(\mathbf{x}_{s_i}, Z_k) = \max_{\substack{1 \le j \le k \\ j \ne i}} W_{s_i s_j} \tag{3}$$

2.2 The Proposed Approach

In this work, we aim to select k examples ($\mathcal{Z}$) with maximum density and minimum redundancy from $\mathcal{X}$. Suppose Z_k is an arbitrary subset of $\mathcal{X}$ that contains k examples and $Z_k = \{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, ..., \mathbf{x}_{s_k}\}$. The final selected k examples $\mathcal{Z}$ can be obtained by solving the following problem:

$$\mathcal{Z} = \arg\max_{Z_k \subseteq \mathcal{X}} \sum_{i=1}^{k} (den(\mathbf{x}_{s_i}) - \lambda \, red(\mathbf{x}_{s_i}, Z_k)) \tag{4}$$

where λ is the tradeoff parameter that can determine the importance of density and redundancy.

Unfortunately, the optimization problem (4) is a highly complicated problem. To get the optimal subset $\mathcal{Z}$, we would have to search over all possible sets to determine the unique optimal $\mathcal{Z}$. It is impossible to finish in short time with the number of examples increased.

However, it should be noted that $den(\mathbf{x}_{s_i})$ is only dependent on s_i while $red(\mathbf{x}_{s_i}, Z_k)$ is related with $\{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, ..., \mathbf{x}_{s_k}\}$. Suppose $X_u = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_u\} (u = 1, ..., n)$ and $Z(u, v) (v \le u)$ denotes the optimal solution of selecting v examples from X_u. We transform the problem (4) into a relatively simple form:

$$\mathcal{Z} = \arg\max_{Z_k \subseteq \mathcal{X}} \sum_{i=1}^{k} (den(\mathbf{x}_{s_i}) - \lambda \, red(\mathbf{x}_{s_i}, Z(s_i - 1, i - 1))) \tag{5}$$

where $red(\mathbf{x}_{s_i}, Z(s_i - 1, i - 1))$ is the redundancy between $\mathbf{x}_{s_i}$ and the selected $i - 1$ examples from $X_{s_i-1} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{s_i-1}\}$.

Obviously, $red(\mathbf{x}_{s_i}, Z(s_i - 1, i - 1))$ is relevant with $\{s_1, ..., s_{i-1}\}$ but irrelevant with $\{s_{i+1}, .., s_k\}$. This means that when we select the $i-th$ example, it is required to have small redundancy with the selected examples $\{\mathbf{x}_{s_1}, ..., \mathbf{x}_{s_{i-1}}\}$. This guarantees that the next selected example must be different from the previous selected examples. This idea is in accord with the process of sequential examples selection.

2.3 The Dynamic Programming Approach

The problem (5) can be solved by dynamic programming that breaks it down into simpler subproblems. Suppose $F(u, v)(u \geq v)$ denotes the maximum volume of information of selecting v examples from X_u. As defined above, $Z(u, v)$ denotes the optimal solution of selecting v examples from X_u, hence $\mathcal{Z} = Z(n, k)$. $F(u, v)$ and $Z(n, k)$ can be described as follows:

$$F(u, v) = \max_{Z_v \subseteq X_u} \sum_{i=1}^{v} (den(\mathbf{x}_{s_i}) - \lambda \, red(\mathbf{x}_{s_i}, Z(s_i - 1, i - 1))) \tag{6}$$

$$Z(u, v) = \arg\max_{Z_v \subseteq X_u} \sum_{i=1}^{v} (den(\mathbf{x}_{s_i}) - \lambda \, red(\mathbf{x}_{s_i}, Z(s_i - 1, i - 1))) \tag{7}$$

where $u \in \{1, 2, ..., n\}, v \in \{1, 2, ..., k\}$, and $u \geq v$.

Our final goal is to find $Z(n, k)$ that decides which k examples should be selected from the n unlabeled examples.

It should be noted that there are two special situations: $v = 1$ and $u = v$. If $v = 1$, there are no redundancy since only one example is selected. Therefore, the example with maximum density should be selected. If $u = v$, obviously, all of the examples in X_u should be selected. So

$$F(u, v) = \begin{cases} \max\limits_{\mathbf{x}_i \in X_u} den(\mathbf{x}_i) & if \;\; v = 1 \\ \sum_{i=1}^{u} den(\mathbf{x}_i) - \lambda \, red(\mathbf{x}_i, X_{i-1}) & if \;\; u = v \end{cases} \tag{8}$$

Suppose we have already obtained the optimal solution of selecting $v - 1$ and v examples from X_{u-1}, now we consider the optimal solution of selecting v examples from u unlabeled examples. If the example $\mathbf{x}_u$ has small density and large redundancy with $Z(u - 1, v - 1)$, obviously we will not select the example $\mathbf{x}_u$. Hence the optimal solution of selecting v examples from X_u should be the same as selecting v examples from X_{u-1}. On the contrary, if the example $\mathbf{x}_u$ has large density and small redundancy with $Z(u - 1, v - 1)$, we prefer to select it for labeling. In this situation, since $Z(u - 1, v - 1)$ is the optimal solution of selecting $v - 1$ examples from X_{u-1}, the optimal solution of selecting v examples from X_u is $Z(u, v) = Z(u - 1, v - 1) \cup \mathbf{x}_u$.

Table 1. The process of the proposed active learning algorithm

Input:
Initial unlabeled data set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_n\}$, the gaussian parameter (σ), the number of nearest neighbor (p), the tradeoff parameter (λ), the number of examples need to select (k)
Output:
$Z(n,k)$: the k selected examples
Procedure:
compute weight W, $den(\mathbf{x}_i), (i = 1, 2, ..., n)$
Initilize X_u, $F = \mathbf{0}_{nk}$, $Z(u,v) = \emptyset$
For $u = 1 : n$
$F(u,1) = \max_{\mathbf{x}_i \in X_u} den(\mathbf{x}_i)$
$Z(u,1) = \arg\max_{\mathbf{x}_i \in X_u} den(\mathbf{x}_i)$
End
For $u = 2 : n$
For $v = 2 : k$
$C(u) = den(\mathbf{x}_u) - \lambda\, red(\mathbf{x}_u, Z(u-1, v-1))$
$F(u,v) = \max(F(u-1,v), F(u-1,v-1) + C(u))$
$Z(u,v) = \begin{cases} Z(u-1,v) & if \;\; F(u,v) = F(u-1,v) \\ Z(u-1,v-1) \cup \mathbf{x}_u & else \end{cases}$
End
End
Return $Z(n,k)$

In general, the relationships between $F(u-1, v-1)$, $F(u-1, v)$, and $F(u, v)$ can be described as follows:

$$C(u) = den(\mathbf{x}_u) - \lambda\, red(\mathbf{x}_u, Z(u-1, v-1)) \tag{9}$$

$$F(u,v) = \max(F(u-1,v), F(u-1,v-1) + C(u)) \tag{10}$$

where $2 \le u \le n$, $2 \le v \le k$ and $v \le u$. $Z(u,v)$ can be computed as follows:

$$Z(u,v) = \begin{cases} Z(u-1,v) & if \;\; F(u,v) = F(u-1,v) \\ Z(u-1,v-1) \cup \mathbf{x}_u & else \end{cases} \tag{11}$$

Since $Z(1,1), Z(2,1)$ is easy to obtain, the global optimal solution $Z(n,k)$ can be obtained by iteration. The dynamic programming approach to solve the examples selection problem is summarized in Table 1. As can be seen from Table 1, the proposed active learning algorithm is easy to perform and the computational cost is low.

3 Experiments

In this section, experiments of terrain classification are performed with different active learning algorithms. In order to demonstrate the effectiveness of our proposed algorithm, we evaluate and compare four active learning methods:

- **Random Sampling** (Rand) method, which selects examples randomly from unlabeled data set.
- **D-Optimal Design** (DOD) as described in Section 2.1.
- **Manifold Adaptive Experimental Design** (MAED) Algorithm [9], which performed convex TED in manifold adaptive kernel space.
- **Active Learning with Maximum Density and Minimum Redundancy** (MDMR), which is proposed in this paper.

3.1 Data and Experimental Settings

Terrain image dataset used in the experiment was constructed by us from the Outex Database [10], which is consisted of two data sets: Outex-0 and Outex-1. Each of them includes 20 outdoor scene images and the size of each image is 2272×1704. The images are marked as one type of bush, grass, tree, sky, road, and building. The marked area of each image is cut into patches with size 64×64 and each patch is regarded as an example. In this work, we extract 50 patches of each class (totally 300 patches) to construct a pool of unlabeled data set for examples selection. The testing examples, which are predicted to evaluate active learning algorithms' performance, are also consisted of 300 patches (50 patches from each class).

Fig. 1. Patch examples of Outex: sky, tree, bush, grass, road, and building

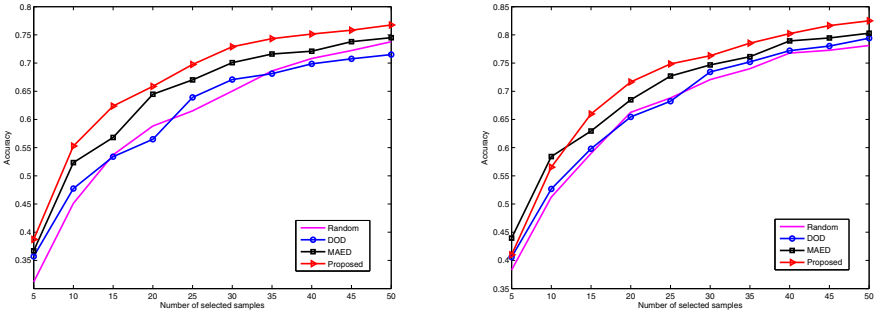
Two examples of each class are shown in Figure 1. It is difficult to classify these terrains directly in image color space. Thus color histogram feature [11] and texture feature with the rotation-invariant operators $LBP^{riu2}_{8,1+16,3}$ [12] are extracted and combined together. As a result, each example is represented by a 43-dimensional feature vector.

Logistic regression with l_2 regularization is used as classifier and the regularization parameter is set to be 0.5. The parameters in our proposed active learning algorithms are set as follows: the gaussian kernel parameter ($\sigma = 0.1$), the number of nearest neighbor ($p = 15$), and the tradeoff parameter ($\lambda = 10$).

3.2 Results

The process of the experiments are as follows: Firstly, we select $k(k = 5, 10, 15, ..., 50)$ examples from unlabeled data set for labeling and training a classifier.

Then we perform classification on testing data set and the accuracy is defined as Correct Classification Rate (CCR). The experiments are repeated 20 times and the average accuracy is computed as the final result.



(a) Results on Outex0 (b) Results on Outex1

Fig. 2. Classification performance on Outex0 and Outex1 dataset using Rand, DOD, MAED, and the proposed active learning algorithm

Figure 2 shows the average classification accuracy versus the number of training (selected) examples. As can be seen, our proposed MDMR algorithm significantly outperforms the other active learning algorithms in most cases. The MAED algorithm outperforms Random Sampling and DOD method in most cases. DOD and Random Sampling perform comparably to each other. When only five examples are selected, there exists at least one class that does not have any labeled examples. Therefore, in this case, all of the algorithms yield low classification accuracy. As the number of selected examples increases, the classification accuracy of all of the algorithms increases. As shown in 2, with only 40 selected examples, MDMR algorithms performs comparably to or even better than the other algorithms with 50 selected examples. Our MDMR algorithm yields the highest classification accuracy.

4 Conclusion

In this paper, we have introduced a novel active learning algorithm called MDMR, which selects the examples with maximum density and minimum redundancy. The experimental results on terrain classification demonstrate that it is better than other popular active learning algorithms.

The disadvantage of this proposed algorithm is that it is not global optimal since the examples are sequentially selected. The redundancy of selected example $\mathbf{x}_{s_t}$ is measured by the redundancy between $\mathbf{x}_{s_t}$ and previous selected $t-1$ examples $\{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, ..., \mathbf{x}_{s_{t-1}}\}$. Thus the redundancy among all the selected examples may not be minimum. Moreover, combining different criteria such as density and redundancy is a significant problem in active learning. There is a lot of work that needs to be explored.

Acknowledgements. This work is partially supported by National Natural Science Foundation of China under Grant Nos. 61373063, 61233011, 61125305, 61375007, 61220301, and by National Basic Research Program of China under Grant No. 2014CB349303.

References

1. Settles, B.: Active learning literature survey. University of Wisconsin, Madison (2010)
2. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. The Journal of Machine Learning Research 2, 45–66 (2002)
3. Ji, M., Han, J.: A variance minimization criterion to active learning on graphs. In: International Conference on Artificial Intelligence and Statistics, pp. 556–564 (2012)
4. Hu, X., Wang, L., Yuan, B.: Querying representative points from a pool based on synthesized queries. In: The 2012 International Joint Conference on Neural Networks (IJCNN), pp. 1–6. IEEE (2012)
5. Chakraborty, S., Balasubramanian, V., Panchanathan, S.: Dynamic batch mode active learning. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2649–2656. IEEE (2011)
6. Atkinson, A.C., Donev, A.N., Tobias, R.D.: Optimum experimental designs, with SAS. Oxford University Press, Oxford (2007)
7. He, X.: Laplacian regularized d-optimal design for active learning and its application to image retrieval. IEEE Transactions on Image Processing 19(1), 254–263 (2010)
8. Gu, Y., Jin, Z.: Neighborhood preserving d-optimal design for active learning and its application to terrain classification. Neural Computing and Applications 23(7-8), 2085–2092 (2013)
9. Cai, D., He, X.: Manifold adaptive experimental design for text categorization. IEEE Transactions on Knowledge and Data Engineering 24(4), 707–719 (2012)
10. University of oulu texture database, `http://www.outex.oulu.fi/temp/`
11. Procopio, M.J., Mulligan, J., Grudic, G.: Learning terrain segmentation with classifier ensembles for autonomous robot navigation in unstructured environments. Journal of Field Robotics 26(2), 145–175 (2009)
12. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)