Active Learning based on Random Forest and Its Application to Terrain Classification

Yingjie Gu, Dawid Zydek, and Zhong Jin

1 Introduction

In the machine learning literature many supervised algorithms have been proposed to perform pattern classification tasks. But in many pattern recognition tasks, labels are often expensive to obtain while a vast amount of unlabeled data are easily available. And redundant samples are often included in the training set, thus slowing down the training process of the classifier without improving classification results. To solve this problem, active learning [1][2] techniques are proposed to select the most valuable samples for manually labeling to train a classifier.

Uncertainty, density, and diversity are three of the most important criteria in active learning. Uncertain samples are usually able to improve the current classifier most. The most popular uncertainty sampling is SVM_{active} [3] [4] that selects the sample nearest to the current decision boundary. In density sampling, samples in dense regions are thought to be representative and informative. The cluster structure of unlabeled data is usually exploited to find samples in dense regions. The main weakness of uncertainty and density sampling is that they are unable to exploit the abundance of unlabeled data. Thus the diversity criterion was proposed to select a set of unlabeled samples that are as more diverse as possible in the feature space, which reduces the redundancy among the samples selected at each iteration.

Recently, some active learning algorithms tried to combine two criteria to find the optimal samples. In [5], Huang et al. tried to query informative and representative examples based on the min-max view of active learning [6]. Some active learning techniques also query a batch of unlabeled samples at each iteration by considering both uncertainty and diversity criteria [7] [8]. Shi et al. [9] proposed a batch mode active learning method for Networked Data with three criteria (i.e., minimum redundancy, maximum uncertainty, and maximum impact).

The processing platform for active learning should be considered as well. Among many others, the distributed processing systems are gaining many attention and are suitable for active learning system that gathers samples from many distributed locations, and processes them as one virtual entity. Such solution was proposed in [10] where the system that optimizes the processing task allocation in Peer-to-Peer based computing architecture was proposed. In [11], the decentralized approach was shown, also supporting the multiple data sources (suitable to obtain samples).

Large numbers of active learning algorithms are based on SVM and regression classifier. But there is little work about active learning using random forest classifier. According to the information we have, DeBarr et al. have made an exploration in random forest active learning [12]. In this paper, we proposed a novel active learning algorithm based on random forest that selects samples with large uncertainty, density, and diversity for manual labeling. For each unlabeled samples, we use the difference between the most votes and second most votes from the random forest classifier to measure its uncertainty. The average distance between the sample and its k-nearest unlabeled neighbors is used to measure the density while the distance between the sample and its nearest labeled neighbor is used to measure the diversity.

The rest of this paper is organized as follows. Section 2 describes the proposed active learning based on random

Y. Gu (✉)
Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Department of Electrical Engineering, Idaho State University, Pocatello 83209-8060, USA
e-mail: csyjgu@gmail.com

D. Zydek
Department of Electrical Engineering, Idaho State University, Pocatello 83209-8060, USA
e-mail: zydedawi@isu.edu

Z. Jin
Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
e-mail: zhongjin@njust.edu.cn

H. Selvaraj et al. (eds.), *Progress in Systems Engineering: Proceedings of the Twenty-Third International Conference on Systems Engineering*, Advances in Intelligent Systems and Computing 1089,
DOI 10.1007/978-3-319-08422-0_41,

forest. The experimental settings and results on several data sets are presented in Section 3. Finally, Section 4 discusses the conclusion of this work.

2 Active Learning based on Random Forest

2.1 Random Forest and Active Learning

Random forest is an ensemble classifier that is composed of many decision trees. It was first proposed to solve the classification problem [13]. For a new testing sample, each tree gives a prediction. So the testing sample receives a vote on that prediction class. The prediction label of the sample is the class with the most votes. In recent years, there have been a lot of applications [14] [15] [16] in computer vision, which employ random forest as classifier. Although it has been widely applied, there is little work apply random forest in active learning. According to what we know, DeBarr et al. have made an exploration in random forest based active learning [12]. They queried the sample whose probability assigned by the random forest model is closet to 0. 5. The probability of an instance is computed as the proportion of decision trees assigning the instance label. Similar to Tong's SVM active learning, it just selects the most uncertain sample for manual labeling.

A general active learning procedure is as follows:

step 1. Randomly select several samples to construct an initial training set $\mathcal{L}$ to train a classifier.

step 2. According some criteria, select a set of samples from unlabeled pool $\mathcal{U}$ for manual labeling.

step 3. Selected samples are added to $\mathcal{L}$ and the classifier is retrained by updated training set.

step 4. Repeat 2 and 3 until a stop criterion is satisfied.

The key problem in active learning is how to select a set of samples or a sample from unlabeled pool $\mathcal{U}$ in Step 2. In this paper, a novel active learning algorithm is proposed to select samples with maximum uncertainty, density, and diversity to improve the classifier.

2.2 The Proposed Approach

Given a dataset by $\mathcal{D} = \{x_1, x_2, ..., x_{n+m}\}$, where $\mathbf{x}_i$ is a sample of d dimension vector. The labeled data is $\mathcal{L} = \{x_1, x_2, ..., x_n\}$ while the unlabeled data is $\mathcal{U} = \{x_{n+1}, x_{n+2}, ..., x_{n+m}\}$, so $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$. The label of sample $\mathbf{x}_i$ is $y_i \in \{1, 2, \ldots, c\}$, $i = 1, \ldots, n$.

In the following, we will introduce how to select samples with maximum uncertainty, density, and diversity.

Uncertainty step A model can be trained with random forest based on labeled data $\mathcal{L}$. Perform classification on unlabeled data $\mathcal{U}$, the vote of each sample assigned to each class can be obtained. We use $\mathcal{V} = \{v_{ij} \mid i = n+1, ..., n+m, j = 1, ..., c\}$ to denote the votes of all unlabeled samples assigned to each class. v_{ij} denotes the vote of the unlabeled data $x_i \in \mathcal{U}$ assigned to class $j \in \{1, 2, \ldots, c\}$.

In prediction, random forest assigns each sample to the class that gets the maximum vote. The maximum vote of sample $\mathbf{x}_i$ is defined as $\overline{v}_i$, so

$$\overline{v}_i = \max_{j=1,...,c} \{v_{ij}\} \tag{1}$$

In traditional active learning based on random forest, among all unlabeled samples, the one with the minimum $\overline{v}_i$ [12] is selected for manual labeling. In their opinion, the smaller $\overline{v}_i$ is, the more uncertain the classification result is.

Here we propose a new method to measure the uncertainty of samples. As Figure 1 shows, the sample in left figure is denoted as $\mathbf{s}_1$ while the sample in right figure is denoted as $\mathbf{s}_2$. It can be seen that $\mathbf{s}_1$ got the maximum vote less than 300 while $\mathbf{s}_2$ got the maximum vote more than 300. Traditional random forest active learning will select $\mathbf{s}_1$ since its maximum vote is smaller than that of $\mathbf{s}_2$. However, the maximum vote of $\mathbf{s}_1$ is much larger than the votes of any other class. On the contrary, the maximum vote of $\mathbf{s}_2$ is just slightly larger than the vote of class 4. Thus we suppose $\mathbf{s}_2$ is more uncertain than $\mathbf{s}_1$ since the label of $\mathbf{s}_2$ is more ambiguous.

In view of the above reason, the difference between the maximum vote and the second maximum vote can be used to measure samples' uncertainty. Smaller difference means more uncertainty of a sample.

If sample $\mathbf{x}_i$ get the maximum vote on class p and second maximum vote on class q, namely

$$p = \arg\max_{j=1,...,c} \{v_{ij}\}$$

$$q = \arg\max_{j=1,...,c, j \neq p} \{v_{ij}\}$$

The difference between the maximum vote and the second maximum vote is

$$unc_i = v_{ip} - v_{iq} \tag{2}$$

unc_i is able to measure uncertainty of sample $\mathbf{x}_i$.

Density step Many active learning algorithms select samples that are most representative to unlabeled data. These approaches aim to exploit the cluster structure of unlabeled data, usually by a clustering method. Instead, we propose a novel idea to select representative samples. As we know, samples in dense regions are usually thought to be representative. In other words, a representative point is usually near to

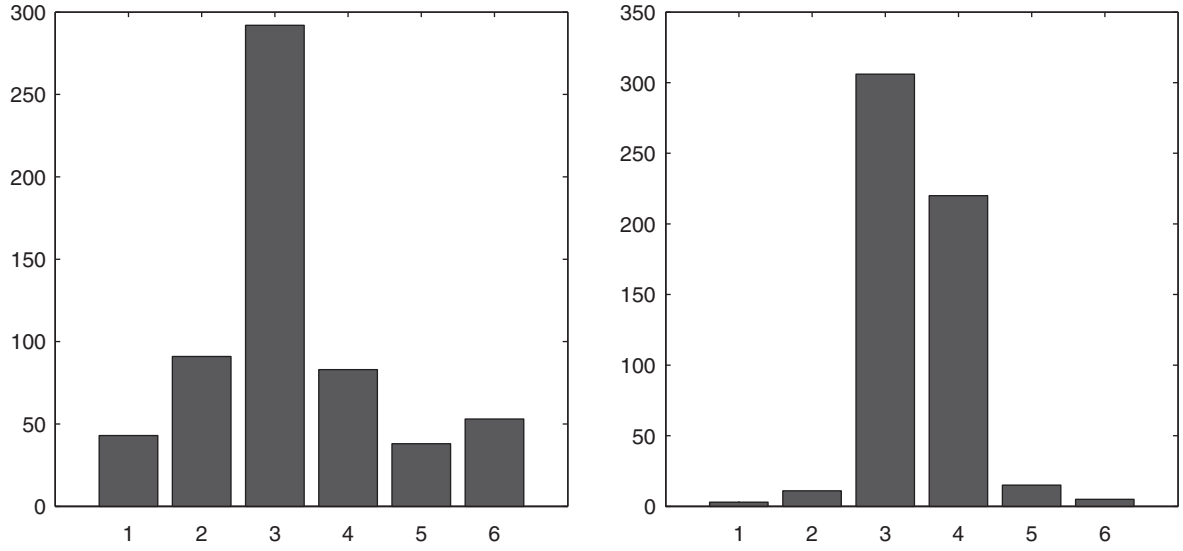

Fig. 1. Votes of two samples: X-axis indicate the class label while Y-axis indicate the vote of each class.

its neighbors. On the contrary, a point far from its neighbors is usually not representative or an outlier. Thus we use the average distance from a sample to its k-nearest neighbors to measure the representativeness of the sample.

For any $x_i \in \mathcal{U}$, define its k nearest neighbors from $\mathcal{U}$ as $x_{i_j}, j = 1, ..., k, x_{i_j} \in \mathcal{U}$. The average distance from $\mathbf{x}_i$ and its k nearest neighbors can be computed:

$$den_i = \frac{1}{k}\sum_{j=1}^{k}\left\|x_i - x_{i_j}\right\|^2 \tag{3}$$

We use den_i to measure the density of sample $\mathbf{x}_i$.

Diversity step Some active learning algorithms select samples that are similar with labeled samples. So it will not improve the classifier obviously. In our proposed approach, the distance between the sample and its nearest labeled neighbor is used to measure the similarity between the sample and labeled samples. If the sample is far from its nearest labeled neighbor, it is dissimilar with other labeled samples. On the contrary, if it is close to the nearest labeled sample, there is at least one sample that is similar with it in the labeled set. Therefore, we select the sample that is far from its nearest labeled neighbor for manual labeling.

For any $x_i \in \mathcal{U}$, compute the distance between $\mathbf{x}_i$ and its nearest labeled neighbor:

$$div_i = \min_{j=1,2,...,n}\left\|x_i - x_j\right\|^2 \tag{4}$$

$$x_i \in \mathcal{U}, x_j \in \mathcal{L}$$

Large div_i indicates that sample $\mathbf{x}_i$ has little similarity with labeled samples.

Selection Function In this section, we want to select an uncertain, representative, and diverse sample that is measured by unc_i, den_i, and div_i. To combine these three criteria together, unc_i, den_i, and div_i $(i = n+1, ..., n+m)$ are normalized to [0, 1], respectively. For any vector $\mathbf{p} = [p_1, \ldots, p_n]$, the normalization operation is as follow:

$$p_i^{'} = \frac{p_i - min}{max - min} \tag{5}$$

where $max = \max_{j=1,...,n}\{p_j\}$ and $min = \max_{j=1,...,n}\{p_j\}$.

The query criteria of proposed active learning can be described as:

$$s = \underset{i=n+1,...,n+m}{argmin}\ \{unc_i + den_i - div_i\} \tag{6}$$

As above analysis, proposed algorithm would select the sample $\mathbf{x}_s$ that has large uncertainty, density, and diversity.

The proposed active learning approach is summarized in Table 1.

3 Experiments

To demonstrate the effectiveness of our proposed algorithm, we compare it with other three active learning methods:

- **Random Sampling** method, which randomly select samples from the unlabeled data.
- **Random Forest active learning** (RFAL), proposed by DeBarr[12] which select samples with min-max votes.
- **Support Vector Machine active learning** (SVMAL) [4], which selects the point closest to the current decision boundary.

Table 1 The proposed active learning approach

Input:
$\mathcal{D}_{train}$: A data set for training
$\mathcal{D}_{test}$: A data set for testing
Initialize:
$\mathcal{L}$: random select 10 samples from $\mathcal{D}_{train}$
$\mathcal{U}$: $\mathcal{U} = \mathcal{D}_{train} - \mathcal{L}$
Repeat:
for $i = n + 1$ to $n + m$ do
Calculate unc_i, den_i, div_i
Calculate $score_i = unc_i + den_i - div_i$
end for
$s = \underset{i=n+1,\ldots,n+m}{argmin} \; score_i$
$\mathcal{L} \leftarrow \mathcal{L} \cup \mathbf{x}_s$
$\mathcal{U} \leftarrow \mathcal{U} - \mathbf{x}_s$
Until the number of selected or the required accuracy is reached

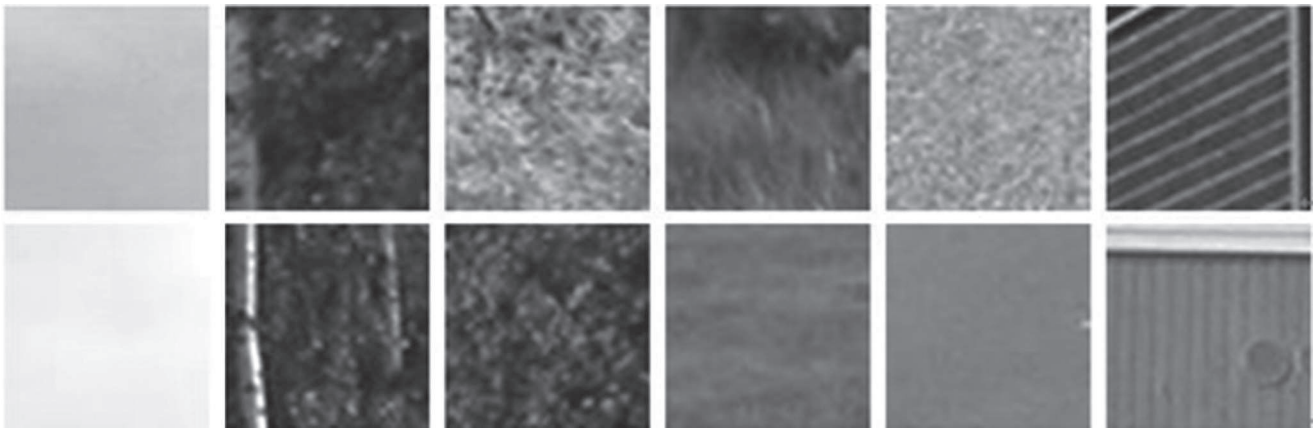

Fig. 2. Patch examples of Outex: bush, grass, tree, sky, road, and building

– **Proposed algorithm**, which query samples with maximum uncertainty, density, and diversity based on random forest. Random forest tool is available here[17].

The analysis of outdoor terrain images for navigating a mobile robot is very challenging. In experiments, above four active learning algorithms are performed on two terrain image data sets.

3.1 Outex Data Sets

Outex data [18] contains two data set: Outex0 and Outex1. Both of them include 20 outdoor scene images and the images' size is 2272 × 1704. The labeled area of each image is cut into patches of size 64 × 64. The patches contain 6 terrain classes defined as bush, grass, tree, sky, road, and building with considerable changes of illumination. Two sample patches of each class are shown in Figure 2. Each terrain patch is represented by a 64 × 64 dimensional vector in image space. It is difficult to classify these terrains directly in image space. We extract color histogram feature [19] and texture feature using rotation-invariant operators $LBP^{riu2}_{8,1+16,3}$ [20] [21]. Both of these features were proved to be effective in performing outdoor scene classification tasks.

For each class, 50 patches are randomly selected to construct a training set while 50 patches are randomly selected for testing. Then 10 patches in training set are randomly selected to construct an initial labeled set $\mathcal{L}$ and the rest in training set construct unlabeled set $\mathcal{U}$. At each iteration, we

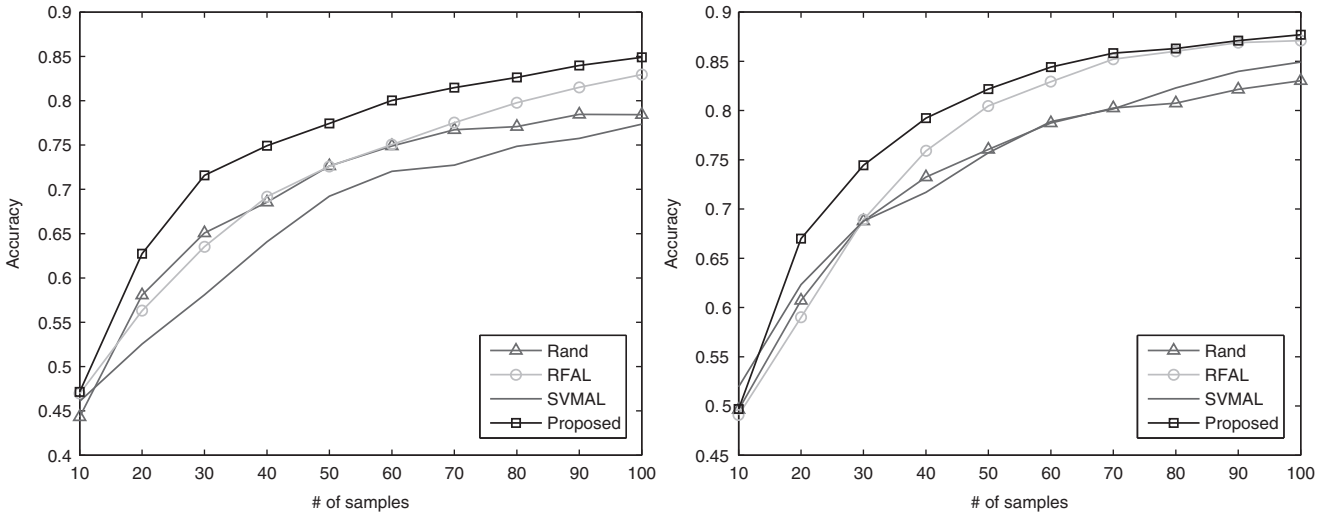

Fig. 3. Classification accuracy on Outex-0, Outex-1

query 10 samples from $\mathcal{U}$ for manual labeling and add them to $\mathcal{L}$. The max number of iteration is fixed at 10. The experiment is repeated for 20 times and the average classification accuracy is shown in Figure 3.

From Figure 3, we can see that our proposed algorithm outperforms other methods. SVM active learning performs the worst nearly all the case since it is not developed for multi-class active learning. Traditional random forest active learning is worse then proposed method because it just selects uncertain sample while ignore the sample's density and diversity.

3.2 Hand-Labeled DARPA LAGR Datasets

Hand-Labeled DARPA LAGR Data sets [19] contain 3 scenes with different lighting condition. Each data set consists of 100-frame, hand-labeled image sequence. Each image was manually labeled, with each pixel being placed into one of three classes: OBSTACLE, GROUNDPLANE, or UNKNOWN. Feature extraction method is fixed as color histogram [22]. To create a color histogram, color intensities in each of the three color channels(R, G, and B) in the neighborhood of the reference pixel are binned. The number of bins is fixed at 5 and the window size is fixed at 16×16. Using three color channels and 5 bins per channel results in a feature image with feature depth of 15 values(3 channels $\times$ 5 bins per channel).

Active learning methods are performed on one of the data sets, DS2A. 5 points of each class in each frame are randomly selected to construct a training set. So the training set consists of 1000 samples. In the same way, a testing set can be constructed. Firstly, 10 points from training set were randomly selected for manual labeling to construct an initial labeled set $\mathcal{L}$ and the rest in the training set construct an unlabeled set $\mathcal{U}$. At each iteration, we query 5 samples from $\mathcal{U}$ for manual labeling and add them to labeled set $\mathcal{L}$. The max number of iteration is fixed at 10. The experiment is repeated for 20 times and the average classification accuracy is shown in Figure 4.

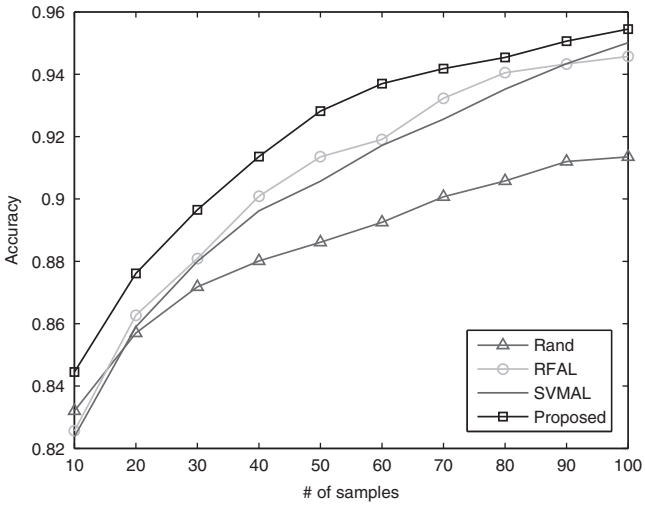

Fig. 4. Classification accuracy on DS2A

As can be seen from Figure 4, our proposed active learning algorithm performs the best. SVM active learning and random forest active learning perform worse since they just query the most uncertain sample but ignore the samples' density and diversity. Random sampling performs worst because it selects samples without any criteria.

4 Conclusion

In this paper, we propose a novel active learning technique for solving multiclass classification problem with random forest classifier. The proposed technique combines samples' uncertainty, density, and diversity information and selects the most valuable one. The results of experiment indicate the proposed method outperforms other methods.

There are several advantages of the proposed algorithm:

- The proposed active learning algorithm can also initialize the labeled set $\mathcal{L}$. The selection function is:

$$s = \underset{i=n+1,\ldots,n+m}{argmin} \{unc_i + den_i - div_i\} \tag{7}$$

  if we set $unc_i = 0$ and $div_i = 0$, we can decide which sample should be firstly labeled. Then the labeled set can be constructed according to selection function.
- It can be expanded to other classifiers through altering the first criterion that measures uncertainty of samples.
- There are no other parameters except two parameters in random forest classifier.
- It is independent with samples' label so it can be used for multi-class active learning.

Moreover, there are several interesting directions for extending present work. In this paper, we choose euclidean distance to measure the similarity of two samples, how about mahalanobis distance or consine distance? And how to measure the uncertainty, density, and diversity of samples more effectively. Last but not the least, how to combine different criteria together is an extremely difficult and significant problem.

Acknowledgment 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. D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.
2. B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, 2010.
3. S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*. ACM, 2001, pp. 107–118.
4. S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *The Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2002.
5. S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples." in *NIPS*, vol. 23, 2010, pp. 892–900.
6. S. C. Hoi, R. Jin, J. Zhu, and M. R. Lyu, "Semi-supervised svm batch mode active learning for image retrieval," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–7.
7. D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 7, pp. 2218–2232, 2009.
8. S. Patra and L. Bruzzone, "A cluster-assumption based batch mode active learning technique," *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1042–1048, 2012.
9. L. Shi, Y. Zhao, and J. Tang, "Batch mode active learning for networked data," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 33, 2012.
10. G. Chmaj, K. Walkowiak, M. Tarnawski, and M. Kucharzak, "Heuristic algorithms for optimization of task allocation and result distribution in peer-to-peer computing systems," *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 3, pp. 733–748, 2012.
11. G. Chmaj and S. Latifi, "Decentralization of a multi data source distributed processing system using a distributed hash table." *International Journal of Communications, Network & System Sciences*, vol. 6, no. 10, 2013.
12. D. DeBarr and H. Wechsler, "Spam detection using clustering, random forests, and active learning," in *Sixth Conference on Email and Anti-Spam. Mountain View, California*, 2009.
13. L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
14. J. Gall and V. Lempitsky, "Class-specific hough forests for object detection," in *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013, pp. 143–157.
15. A. Yao, J. Gall, and L. Van Gool, "A hough transform-based voting framework for action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2061–2068.
16. C. Marsala and M. Detyniecki, "High scale video mining with forests of fuzzy decision trees," in *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*. ACM, 2008, pp. 413–418.
17. Random forest packages. [Online]. Available: http://cran.r-project.org/web/packages/
18. University of oulu texture database. [Online]. Available: http://www.outex.oulu.fi/temp/
19. Hand-labeled darpa lagr datasets. [Online]. Available: http://www.mikeprocopio.com/labeledlagrdata.html
20. M. Pietikäinen, T. Nurmela, T. Mäenpää, and M. Turtinen, "View-based recognition of real-world textures," *Pattern Recognition*, vol. 37, no. 2, pp. 313–323, 2004.
21. T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.
22. M. J. Procopio, J. Mulligan, and G. Grudic, "Learning terrain segmentation with classifier ensembles for autonomous robot navigation in unstructured environments," *Journal of Field Robotics*, vol. 26, no. 2, pp. 145–175, 2009.