Active learning through two-stage clustering

Min Wang
School of Electrical Engineering
Southwest Petroleum University
Chengdu, China, 610500
Email: wangmin80616@163.com

Ke Fu
School of Electrical Engineering
Southwest Petroleum University
Chengdu, China, 610500
Email: fukemail@163.com

Fan Min*
School of Computer Science
Southwest Petroleum University
Chengdu, China, 610500
Email: minfanphd@163.com
Telephone: +86-135 4068 5200

***Abstract*—Clustering-based approaches take advantage of the structure of the data to select representative instances. However, some algorithms are either inefficient or only applicable to some data. In this paper, we propose an effective and adaptive algorithm that will be called active learning through two-stage clustering (ALTA). The first stage is data preprocessing using the two-round-clustering algorithm. Let n be the number of instances and obtain $\sqrt{n}$ small blocks. For each block, the closest instance of the center is selected as the sample. The second stage is the active learning of sampling instances through density clustering. This stage consists of a number of iterations of density clustering, labeling and classification. In general, data preprocessing reduces the size of the data and the complexity of the algorithm. The combination of distance vector clustering and density clustering makes the algorithm more adaptive. Experiments are performed in comparison against the state-of-the-art active learning algorithms on nine datasets. Results demonstrate that the new algorithm has higher classification accuracy with the same number of labeled data.**

***Index Terms*—Active learning; data preprocessing; two-round-clustering; density clustering.**

I. Introduction

Active learning [1], [2] is widely used in real problems where labeling instances are expensive, such as text classification [3], information extraction [4], image classification [5], and speech recognition [6]. It aims to train an accurate classifier with the minimum labeling cost by actively selecting a few number of most informative instances for labeling. [2]. Initially, the training set is empty or small. The learner selects most critical instances to be labeled. These instances are added to the training set to update classifier. Repeat this process until the labeling cost is used up, or the classifier achieves the required accuracy.

Most existing active learning methods select informative instances by considering an uncertainty criterion [7], [8]. Exemplar approaches include query-by-committee [7], [9] and uncertainty sampling [8]. Query-by-committee [9] approach takes advantage of ensemble methods to acquire the critical instances. A set of classifiers is built to classify unlabeled instances, and instances with the largest confliction are viewed as the most informative. Uncertainty sampling [8], [10] select informativeness instances from a given pool for manually labeling. Informativeness measures the ability of an instance in reducing the uncertainty of a statistical model. The main weakness of these approaches is that they are unable to exploit the abundance of unlabeled data and making it prone to sample bias.

Clustering-based approaches [11] take advantage of the structure of the data to select representative instances. Dasgupta et al. [11] suggested a labeling procedure based on a hierarchical clustering of the training data. The authors show that assuming the learner has obtained a "good" hierarchical clustering, an unlabeled instance can be labeled almost correctly with relatively few label queries. Wang et al. [12] proposed the active learning through density clustering algorithm with three new features. With the clustering structure, sample bias is avoided by selecting representative instances. Additionally, there is no need for an extra classifier.

In this paper, we propose an effective and adaptive algorithm that will be called active learning through two-stage clustering (ALTA). On the one hand, ALTA can exploit the abundance of unlabeled data and avoid sample bias. On the other hand, ALTA uses a two-stage clustering algorithm that can minimizes the dependence on the performance of a single clustering algorithm. Figure 1 illustrates our new algorithm through a running example. Figure 1 (a) depicts a typical dataset Jain with 373 instances. The first stage is data preprocessing, as shown in Figure 1 (b) – (c). First, we divide the dataset into 19 small blocks. We get the block information array $b_{1\times19}$ and the block size array $\rho_{1\times19}$, as shown in Figure 1 (b). Second, we calculate the central instance of each block and form a sampling array $s_{1\times19}$, as shown in Figure 1 (c). The second stage is the active learning of 19 sampling instances through density clustering. Figure 1 (d) – (f) show the process of active learning, including clustering, instances selection, prediction and voting. First, we build the master-tree based on the density and distance of each sampling instance, as shown in Figure 1 (d). Second, we divide it into 2 blocks based on the master-tree, as shown in Figure 1 (e). In Block 1, we select instances 4 and 13 to teach. In Block 2, we select instances 14 and 12 to teach. According to the strategy, Block 2 is pure, and all the remaining instances are directly predicted. Block 1 is not pure and needs to be further divided. Third, Block 1 is further divided into Block 2.1, Block 2.2, as shown in Figure

This work was supported by the National Natural Science Foundation of China (61379089); the Natural Science Foundation of Sichuan Province (15ZB0056, 2017JY0190); and the State Administration of Work Safety project (Sichuan-0008-2016AQ). (Corresponding author: Fan Min.)

1 (f). In Block 2.2, we select instances 11 and 18 to teach. Block 2.1 is still not pure, but the number of teaching has been exhausted. For the remaining instances of Block 2.1, the class label is determined by voting. Finally, Figure 1 (g) shows the classification results. All instances in each block have the same label as the sample instance.

The ALTA algorithm has the following advantages. First, data preprocessing reduces the amount of data and improves the efficiency of the algorithm. Second, the first stage uses distance vector clustering, the second stage uses density clustering. The combination of the two makes the algorithm better adaptable. Third, the first stage can directly obtain the density of each small block as the key parameter of the second stage.

II. Preliminaries

In this section, we review the active learning problem, density clustering and the active learning through density clustering (ALEC) algorithm.

A. Problem definition

Active learning [6] is widely used in real problems where labeling instances are expensive. In practical applications, we consider such application scenarios. The user specifies the number of critical instances to query based on the budget. For example, the total budget for a classification task is \$100, where the price of each label is \$1, and the total number of queries for a task is 100 instances. We consider the following problem.

Definition 1: [13] A decision system S is the 3-tuple:

$$S = (U, C, d), \tag{1}$$

where U is a finite set of objects called the universe, C is the set of conditional attributes, and d is the decision attribute.

Problem 1: [12] **Active learning with fixed number of labels**

Input: The decision system $S = (U, C, d)$ where the value of d is unknown and the number of labels N is provided by the oracle.

Output: The training set $U_r \subset U$, the predicted labels for $U_t = U - U_r$.

Optimization objective: Maximize the prediction accuracy on U_t.

B. Density clustering

Rodriguez et al. [14] proposed the density clustering algorithm, which aims to detect non-spherical clusters and to automatically find the appropriate number of clusters. There are two leading criteria in this method: local density and minimum distance with higher density. The local density ρ_i of data point i is defined as

$$\rho_i = \sum_j \chi(d_{ij} - d_c), \tag{2}$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, d_c is a cutoff distance, d_{ij} is the distance between two instances. The local density ρ_i is equal to the number of instances that are closer than d_c to instance i.

δ is measured by computing the minimum distance between the instance i and any other instance with higher density, namely

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}). \tag{3}$$

For the point with highest density,

$$\delta_i = \max_j (d_{ij}). \tag{4}$$

C. Active learning through density clustering

Active learning deals with the classification task where labeling data is costly. Clustering-based approaches take advantage of the structure of the data to select representative instances. Figure 2 presents active learning through density clustering (ALEC) algorithms [12], which consists of six primary steps.

First, some initialization works are fulfilled. The number of clusters is set to $k = 2$. The set of instances labeled (taught) by the expert is $U_{\mathrm{I}} = \emptyset$. The set of instances classified by the active learner is $U_{\mathrm{II}} = \emptyset$. The set of unlabeled instances is $U_{\mathrm{III}} = U$. Second, the dataset is clustered using density clustering algorithm [14]. The number of clusters is k, which increases with each iteration. Third, with the cluster information, some instances are identified as critical ones. They are denoted as U'_{I} and labeled with expert. Instances in U'_{I} are moved from U_{III} to U_{I}. Fourth, with the existing class labels, some unlabeled instances are classified. They are denoted as U'_{II} and are moved from U_{III} to U_{II}. Fifth, if there are more labels available and some instances are unclassified, the number of clusters increases, and the loop continues. Sixth, if no more labels are available, namely, $U_{\mathrm{I}} \geq N$, the loop is terminated. All unclassified instances are classified directly.

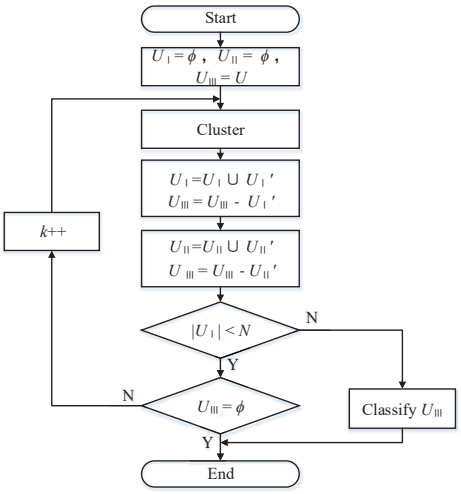
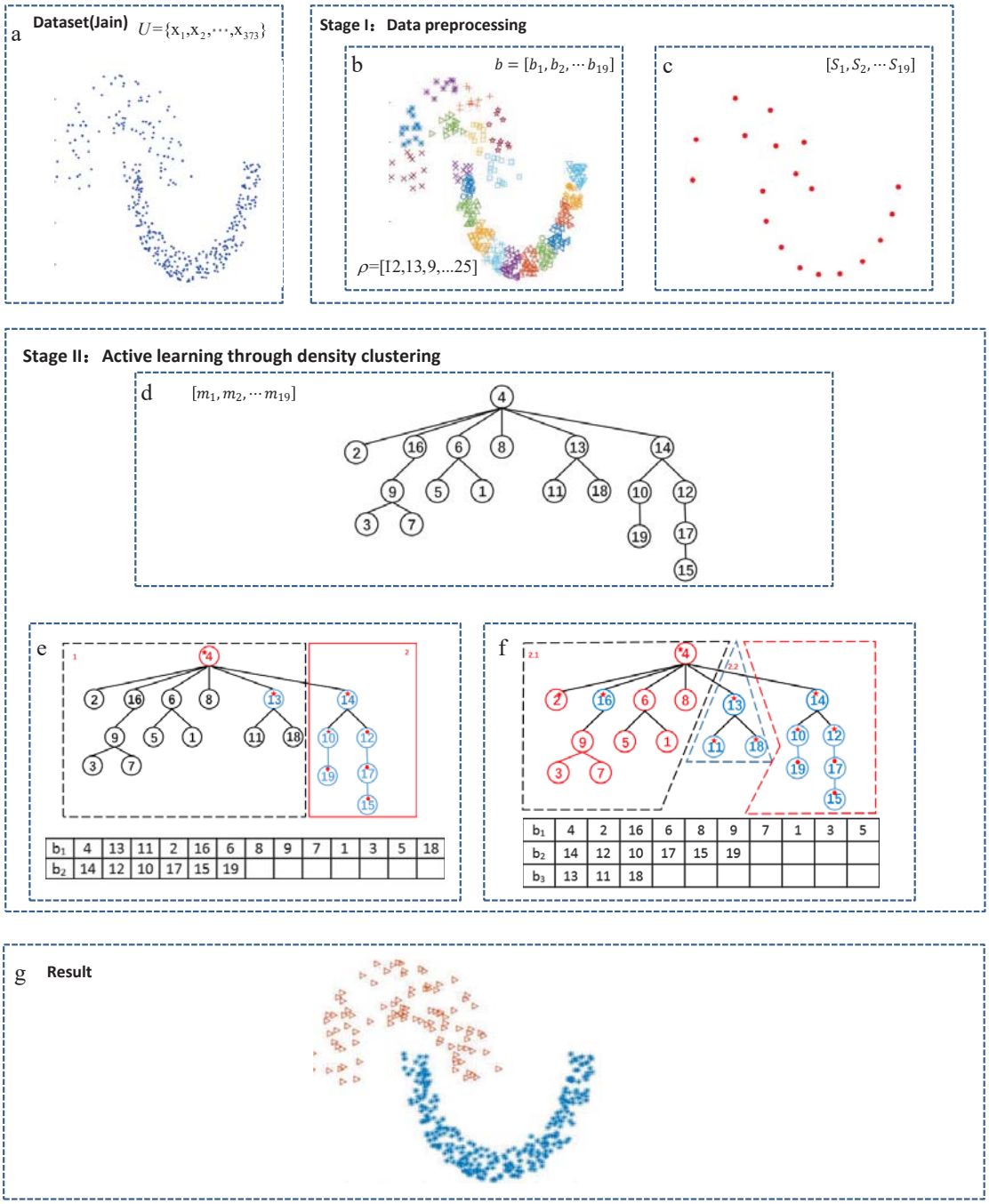


Fig. 2. Active learning through density clustering.



b_1	4	13	11	2	16	6	8	9	7	1	3	5	18
b_2	14	12	10	17	15	19							

b_1	4	2	16	6	8	9	7	1	3	5
b_2	14	12	10	17	15	19				
b_3	13	11	18							

Fig. 1. Running example of the ALTA algorithm.

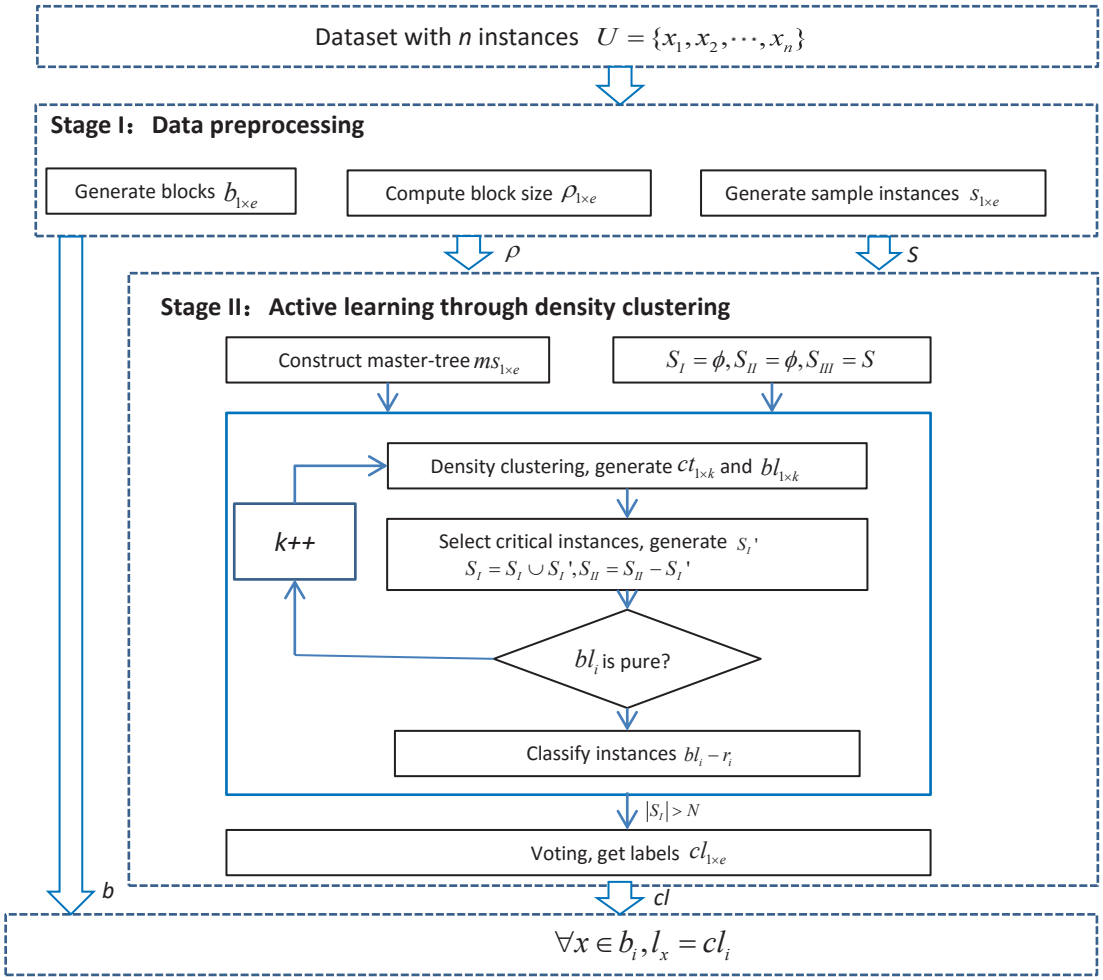


Fig. 3. Framework of the ALTA algorithm.

III. The Proposed Algorithm

A. Algorithm description

Figure 3 shows our ALTA framework. Stage I is data preprocessing. This stage selects the most representative sampling instances, decreasing the data size directly from n to e. First, the two-round-clustering algorithm is used to generate e small blocks $b_{1\times e}$. The number of instances in each block is calculated and the block size array ρ is obtained.

$$\rho_i = |b_i|, \tag{5}$$

where $|b_i|$ is the size of block i. Second, calculate the true center instance of each block and obtain a sample array $s_{1\times e}$.

Two-round-clustering mainly includes three steps. Step 1. Initialize e centers $c_{1\times e}$ using uniform sampling,

$$c_i = x_{i*\sqrt{n}}, i = 0 \cdots \sqrt{n} - 1. \tag{6}$$

Step 2. Each instance belongs to the block of the nearest center. Step 3. Each virtual center takes the mean values of all instances of its block. Step 4. Calculate the closest instance to the virtual center, that is, the sampling instance. The loop terminates after two iterations.

Stage II is the active learning through density clustering of all sampling instances. This stage mainly includes three steps. First, some initialization works are fulfilled. The set of instances labeled (taught) by the expert is $S_{\mathrm{I}} = \emptyset$. The set of instances classified by the active learner is $S_{\mathrm{II}} = \emptyset$. The set of unlabeled instances is $S_{\mathrm{III}} = S$. The number of clusters is set to $k = 2$. Obtain local density of sampling instance ρ and compute minimal distance δ to construct the master tree ms. These will form the basic structure of the subsequent critical instance selection and classification.

Second, select critical instances and classify the pure blocks. This is the main loop corresponding to clustering, labeling, and classifying iteratively. The sample dataset is clustered using density clustering algorithm [14]. We generate the center array $ct = [ct_1, \ldots, ct_k]$ and block information table $bl = [bl_1, \ldots, bl_k]$. With the importance measure, some instances are identified as critical ones. Wang et al. [12] define the

importance measure

$$\gamma = \rho \cdot \delta. \tag{7}$$

They are denoted as S'_{I} and labeled by the expert. Let the set of labeled instances in block bl_i be r_i. bl_i is viewed pure iff 1) $|r_i| \geq |\alpha|$ and 2) r_i is pure, α is calculated according to N. In this case instances in $bl_i - r_i$ are classified to have the same label. The newly classified instances is U'_{II}. The loop terminates when no more labels are available ($|U_I| \geq N$), or all blocks are pure ($U_{\mathrm{III}} = \emptyset$).

Third, classify impure blocks. We will classify the instances if $U_{\mathrm{III}} \neq \emptyset$. That is, there are still some instances unlabeled and unclassified. Standard voting is employed for them.

TABLE I
NOTATIONS AND VARIABLES USED IN FIGURE 3.

Notation	Meaning	Comments
U	All instances	$U = \{x_1, \ldots, x_n\}$
e	$e = \lfloor \sqrt{n} \rfloor$	
S	All sample instances	$S = \{x_1, \ldots, x_e\}$
N	The number of labels provided	See Problem 1
k	Current number of clusters	Initialized as 2
S_{I}	Instances labeled by the oracle	Initialized as $\emptyset$
S'_{I}	Instances labeled in current iteration	
S_{II}	Instances classified by the active learner	Initialized as $\emptyset$
S'_{II}	Instances classified in current iteration	
S_{III}	Unlabeled instances	Initialized as S
b	Block information in Stage I	e clusters
bl	Block information in Stage II	k clusters
ρ_i	The local density of x_i	See Eq. (5)
δ_i	The distance of x_i	See Eq. (3)
ms_i	The master index of x_i	
c_i	The ith center of two-round-clustering, $c_i \in U$	Currently e centers
ct_i	The ith center of density clustering, $ct_i \in S$	Currently k centers
r_i	Labeled instances in bl_i	
cl_i	The label of sampling instance	$cl = [cl_1, cl_2, \ldots, cl_e]$
l_i	The label of x_i	$l = [l_1, l_2, \ldots, l_n]$

B. Complexity analysis

The time complexity of the algorithm are quantified in this section.

Proposition 1: Let m and n be the number of attributes and instances, respectively. For Algorithm ALTA, the time complexity is $O(mn^{\frac{3}{2}})$.

Proof: Table II lists the time complexity. Stage I takes $O(mn^{\frac{3}{2}})$ of time. Stage II takes $O(mn)$ of time. The final step for label sharing takes $O(n)$ of time. Therefore, the time complexity of Algorithm ALTA is

$$O(mn^{\frac{3}{2}}) + O(mn) + O(n) = O(mn^{\frac{3}{2}}). \tag{8}$$

This completes the proof. ■

TABLE II
TIME COMPLEXITY OF THE ALTA ALGORITHM.

Stage	Description	Complexity
Stage I. Data preprocessing	Generate small blocks	$O(mn^{\frac{3}{2}})$
Stage I. Data preprocessing	Compute sampling instances	$O(mn)$
Stage II. Active learning	Build master tree	$O(me^2) = O(mn)$
Stage II. Active learning	Cluster	$O(me^2) = O(mn)$
Stage II. Active learning	Select critical instances	$O(me^2) = O(mn)$
Stage II. Active learning	Classify when ($\lvert U_{\mathrm{I}} \rvert < N$)	$O(e) = O(n^{\frac{1}{2}})$
Stage II. Active learning	Classify when ($\lvert U_{\mathrm{I}} \rvert >= N$)	$O(e) = O(n^{\frac{1}{2}})$
Total		$O(mn^{\frac{3}{2}})$ + $2O(n^{\frac{1}{2}})$ + $4O(mn)$ = $O(mn^{\frac{3}{2}})$

IV. EXPERIMENTS

We conducted experiments to analyze the effectiveness of the ALTA algorithm and answer the following questions:

1) Is the ALTA algorithm more accurate than other supervised classification algorithms, such as C4.5, Naïve Bayes, and Bagging?
2) Is the ALTA algorithm more accurate than state-of-the-art active learning algorithms, including QBC, ALEC and MAED?
3) Is the ALTA algorithm efficient?

The computations were performed on a Windows 10 64-bit operating system with 8GB RAM and Intel (R) Core 2Quad CPU Q9500@2.83GHz processors, using Java software. The ALTA source code is available at *www.fansmale.com/software.html*

A. Datasets

TABLE III
DATASET INFORMATION.

ID	Name	Source	$\lvert U \rvert$	$\lvert C \rvert$	$\lvert V_d \rvert$
1	Spiral	Synthetic	312	2	3
2	Flame	Synthetic	240	2	2
3	Aggregation	Synthetic	788	2	7
4	Iris	UCI	150	4	3
5	Seeds	UCI	210	7	3
6	Ionosphere	UCI	351	2	34
7	Sonar	UCI	208	2	61
8	Solar-flare	UCI	323	2	12
9	DLA	DWP	165633	17	5

B. Comparison with supervised classification algorithms

We compare the ALTA algorithm with three well-known supervised classification algorithms: Bagging [15], C4.5 [16], Naïve Bayes (NB) [17]. We assume that all labels are hidden and can be revealed by the oracle upon request. The performance of the classifier is evaluated by the prediction accuracy.

$$acc = \frac{|U_t| - w}{|U_t|} \times 100\%, \tag{9}$$

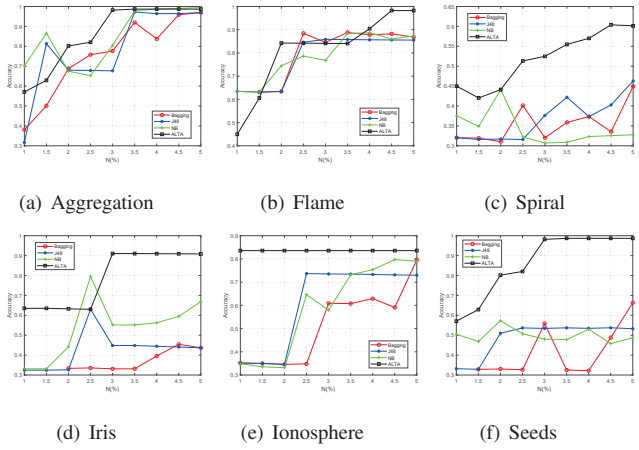
where $|U_t|$ is the size of the testing set, and w is the number of misclassified instances.

Table IV compares the accuracy of ALTA with three supervised classifiers. We set $N = 0.05|U|$ for the eight datasets. The mean rank of the algorithms is listed statistically. The classification accuracy of the best algorithm is highlighted in boldface. ALTA generally outperformed existing supervised classification algorithms. It had the highest accuracy on the seven datasets. Among all the algorithms, we have the highest classification accuracy and the lowest mean rank. Mean accuracy of ALTA is 0.8607 and the mean rank is 1.11.

TABLE IV
THE ACCURACY OF ALTA AND SUPERVISED CLASSIFICATION ALGORITHMS.

Accuracy	Bagging	J48	NB	ALTA
Aggregation	0.9691	0.9691	**0.9947**	0.9869
Flame	0.8684	0.8553	0.8728	**0.9825**
Ionosphere	0.7958	0.7297	0.7898	**0.8357**
Iris	0.4366	0.4366	0.6690	**0.9085**
Seeds	0.6633	0.5327	0.4874	**0.9869**
Solar-flare	0.9772	0.9707	0.9055	**0.9779**
Sonar	0.5253	0.5202	0.5455	**0.6061**
Spiral	0.4493	0.4628	0.3277	**0.6014**
Mean accuracy	0.7107	0.6847	0.6991	**0.8607**
Mean rank	2.72	3.39	2.78	**1.11**

Figure 4 illustrates the accuracy of ALTA and the classical supervised classifiers Bagging, C4.5, and Naïve Bayes (NB). These algorithms were tested using Weka's built-in codes. For the eight datasets, N ranges from $0.01|U|$ to $0.05|U|$. Compared with the counterparts, the ALTA algorithm has the following advantages. First, in 6 of the datasets, ALTA reach maximal accuracy faster than the other algorithms. Second, ALTA is stable with increasing N. In the 6 datasets, the accuracy increase steadily with increasing N. In contrast, the other three algorithms fluctuate on the Spiral and Seeds datasets.



(a) Aggregation (b) Flame (c) Spiral

(d) Iris (e) Ionosphere (f) Seeds

Fig. 4. The accuracy of ALTA and supervised classification algorithms

C. Comparison with active learning algorithms

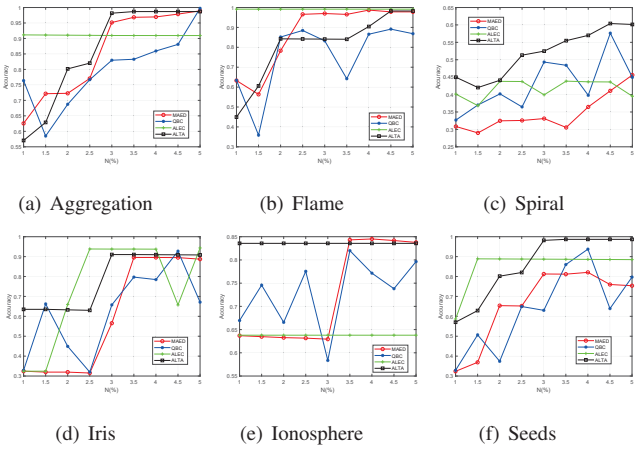
We compare the ALTA algorithm with three state-of-the-art active learning algorithms, including the uncertainty sampling algorithm MAED [18], QBC [9] and clustering-based active learning ALEC [12] algorithm.

Table V shows the accuracy of the ALTA algorithm and the three active learning algorithms. Based on the number of training instances $N = 0.05|U|$, we find the classification accuracy of each method. The mean rank of ALTA on these datasets is the lowest. The mean classification accuracy of ALTA is 0.8607. There are a few cases in which the ALTA algorithm is worse than some of the active learning algorithms. For example,the accuracy of the ALEC algorithm in the Flame dataset is better than ALTA. However, as N increases, the classification accuracy of both gradually reaches the same value.

TABLE V
THE ACCURACY OF ALTA AND THREE ACTIVE LEARNING ALGORITHMS.

Accuracy	MEAD	QBC	ALEC	ALTA
Aggregation	0.9893	**0.9973**	0.9094	0.9869
Flame	0.9781	0.8684	**0.9915**	0.9824
Ionosphere	**0.8378**	0.7964	0.6379	0.8357
Iris	0.8873	0.6713	**0.94366**	0.9085
Seeds	0.7538	0.7972	0.8850	**0.9869**
Solar-flare	0.9771	**0.9805**	0.9780	0.9778
Sonar	0.5076	**0.6313**	0.5606	0.6061
Spiral	0.4561	0.4496	0.6011	**0.6014**
Mean accuracy	0.7984	0.7740	0.7930	**0.8607**
Mean rank	3.00	2.63	2.38	**2.00**

Figure 5 compares the accuracy of the ALTA algorithm with the MAED, QBC and ALEC algorithms. For the six datasets, N ranges from $0.01|U|$ to $0.05|U|$. We observe that in the four datasets Aggregation, Spiral, Ionosphere and Seeds, the ALTA algorithm achieves the highest accuracy faster than the other three algorithms. For example, for the Seeds dataset, the classification accuracy reaches 0.9868 by labeling only 5% of the instances. The accuracy of MEAD, QBC and ALEC were 0.7538, 0.7972, 0.885, respectively.



(a) Aggregation (b) Flame (c) Spiral

(d) Iris (e) Ionosphere (f) Seeds

Fig. 5. The accuracy of ALTA and active learning algorithms

D. Algorithm efficiency

We use DLA dataset to quantify the efficiency of the ALTA algorithm. The number of DLA instances is 165,633. Figure 6 shows the relationship between the training dataset size N and runtime. We randomly sample 10%, 20% ... 100% instances for testing. Experiments demonstrate that the ALTA algorithm is efficient. It is two orders of magnitude faster than the ALEC and can handle millions of data sets in an acceptable test time.

E. Discussion

We are now able to answer the questions proposed at the beginning of this work.

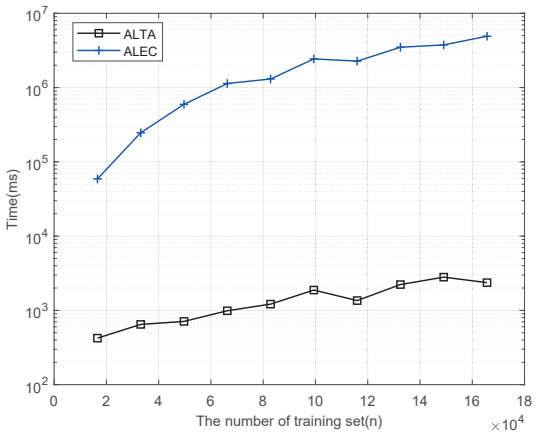


Fig. 6. Runtime as a function of the training set size n.

1) ALTA algorithm is accurate and in some cases it has better performance than the supervised learning algorithm.
2) ALTA algorithm is accurate, and in some cases it has better performance than the state-of-the-art active learning algorithms. It achieves high accuracy while requiring only a small number of labeled instances.
3) ALTA is efficient and scalable. It has lower time complexity than the ALEC algorithm.

The performance of clustering-based active learning algorithms is affected by the quality of clustering. ALTA uses a two-stage clustering algorithm that minimizes the dependence on the performance of a single clustering algorithm. In the first stage, a two-round clustering algorithm suitable for spherical datasets is adopted, and in the second stage, a density clustering algorithm suitable for non-spherical datasets is used to effectively improve the adaptability of the algorithm.

V. Conclusion and Further Work

In this paper, we propose the active learning through two-stage clustering (ALTA) algorithm, which is effective and adaptive. The new algorithm explore the sampling mechanism, reduces the size of the data, and greatly improves the efficiency of the algorithm. The time complexity of the algorithm is $O(mn^{\frac{3}{2}})$, which is lower than $O(mn^2)$ for ALEC. Experiments on the 8 datasets demonstrate that the new algorithm is more accurate than state-of-the-art algorithms. It is two or more orders of magnitude faster than ALEC.

References

[1] D. A. Cohn, Z.-B. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Computer Science*, vol. 4, no. 1, pp. 705–712, 1996.

[2] B. Settles, "Active learning literature survey," *University of Wisconsin-madison*, vol. 39, no. 2, pp. 127–131, 2010.

[3] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, no. 1, pp. 45–66, 2002.

[4] C. A. Thompson, M. E. Califf, and R. J. Mooney, "Active learning for natural language parsing and information extraction," in *Sixteenth International Conference on Machine Learning*, 1999, pp. 406–414.

[5] C. Zhang and T. Chen, "An active learning framework for content-based information retrieval," *Multimedia IEEE Transactions on*, vol. 4, no. 2, pp. 260–268, 2002.

[6] D. Yu, B. Varadarajan, L. Deng, and A. Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433–444, 2010.

[7] G. B. Ran, A. Navot, and N. Tishby, "Kernel query by committee (kqbc)," *Bachrach*, 2004.

[8] Lewis, D. David, Gale, and A. William, "A sequential algorithm for training text classifiers," *Acm Sigir Forum*, vol. 29, no. 2, pp. 3–12, 1994.

[9] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," *Proceding of the Fifth Workshop on Computational Learning Theory*, vol. 284, pp. 287–294, 1992.

[10] L. Sun, J.-C. Xu, and Y. Tian, "Feature selection using rough entropy-based uncertainty measures in incomplete decision systems," *Knowledge-Based Systems*, vol. 36, no. 6, pp. 206–216.

[11] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *International Conference on Machine learning*. ACM, 2008, pp. 208–215.

[12] M. Wang, F. Min, Y.-X. Wu, and Z.-H. Zhang, "Active learning through density clustering," *Expert Systems with Applications*, vol. 85, pp. 305–317, 2017.

[13] Y.-Y. Yao, "A partition model of granular computing," *Lecture Notes in Computer Science*, vol. 3100, pp. 232–253, 2004.

[14] A. Rodriguez and A. Laio, "Machine learning clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[15] J. R. Quinlan, "Bagging, boosting, and c4.5," in *AAAI*, 1996, pp. 725–730.

[16] Z.-Y. Xiang and L. Zhang, "Research on an optimized c4.5 algorithm based on rough set theory," in *International Conference on Management of E-Commerce and E-Government*, 2012, pp. 272–274.

[17] I. Rish, "An empirical study of the Naïve Bayes classifier," *Journal of Universal Computer Science*, vol. 1, no. 2, p. 127, 2001.

[18] D. Cai and X.-F. He, "Manifold adaptive experimental design for text categorization," *IEEE Transactions on Knowledge & Data Engineering*, vol. 24, no. 4, pp. 707–719, 2012.