Mitigating shortage of labeled data using clustering-based active learning with diversity exploration

Xuyang Yan XYAN@@AGGIES.NCAT.EDU
Dept. of ECE, North Carolina A & T State University, NC, USA.

Shabnam Nazmi SNAZMI@AGGIES.NCAT.EDU
Dept. of ECE, North Carolina A & T State University, NC, USA.

Biniam Gebru BTGEBRU@AGGIES.NCAT.EDU
Dept. of ECE, North Carolina A & T State University, NC, USA.

Mohd Anwar MANWAR@NCAT.EDU
Dept. of Computer Science, North Carolina A & T State University, NC, USA.

Abdollah Homaifar HOMAIFAR@@NCAT.EDU
Dept. of ECE, North Carolina A & T State University, NC, USA.

Mrinmoy Sarkar MSARKAR@AGGIES.NCAT.EDU
Dept. of ECE, North Carolina A & T State University, NC, USA.

Kishor Datta Gupta KGUPTA@CAU.EDU
Dept. of Computer Science, Clark Altanta University, GA, USA

Abstract

In this paper, we proposed a new clustering-based active learning framework, namely Active Learning using a Clustering-based Sampling (ALCS), to address the shortage of labeled data. ALCS employs a density-based clustering approach to explore the cluster structure from the data without requiring exhaustive parameter tuning. A bi-cluster boundary-based sample query procedure is introduced to improve the learning performance for classifying highly overlapped classes. Additionally, we developed an effective diversity exploration strategy to address the redundancy among queried samples. Our experimental results justified the efficacy of the ALCS approach.

Keywords: Active learning, Clustering, Diversity

1. Introduction

Active learning (AL) (Lewis and Gale, 1994) approaches have been proposed to address the scarcity of associated labels and reduce the annotation costs for predictive modeling. As an important branch of AL, clustering-based AL methods are proposed to explore the *representativeness* of samples and they have shown reasonable success (Huang et al., 2010; Wang et al., 2017a, 2018, 2020). In clustering-based AL approaches, samples are assumed to share the same class label within the same cluster so that AL is conducted by querying the representative samples from those clusters (Dasgupta and Hsu, 2008).

Challenges. Despite the success of the existing clustering-based AL methods, three major challenges are identified as follows: the performance of existing clustering-based AL methods strongly depends on the selection of clustering parameters; the existing boundary-based selection strategy primarily queries labels for the farthest samples in each cluster without considering neighboring clusters (Wang et al., 2018, 2020); limited efforts are made

on clustering-based AL methods to consider the diversity among queried samples (Wang et al., 2017b; Kee et al., 2018; Wang et al., 2020; Xiao et al., 2020).

In light of these challenges, a clustering-based AL method, namely **AL** utilizing a **C**lustering-based **S**ampling (ALCS),[1] is proposed. ALCS adopts a density-based clustering technique, namely fitness proportionate sharing clustering (FPS-clustering) (Yan et al., 2017), to relax the dependency on clustering parameter optimization. We develop a new bi-cluster boundary-based selection procedure to improve the learning performance of ALCS in datasets with highly overlapped classes. Furthermore, an effective diversity exploration strategy is introduced to reduce the redundancy among active queried samples.

2. Proposed methodology

In this section, the ALCS technique is discussed in terms of its two main components: (i) clustering; and (ii) distance-based instance selection with diversity exploration. Figure 1 provides an overview of the ALCS technique and details are discussed below.

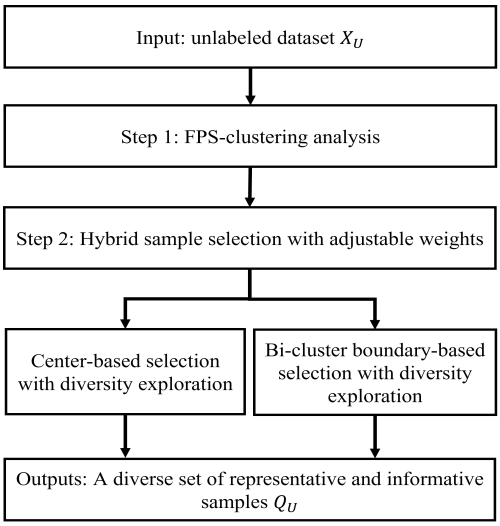


Figure 1: A workflow of the proposed clustering-based AL framework.

2.1 Clustering

To effectively alleviate the exhaustive parameter tuning issue, the FPS-clustering algorithm is employed to discover the cluster information as the first step of the ALCS technique. The FPS-clustering algorithm takes the unlabeled dataset X_U as the input and then outputs a set of clusters and the corresponding cluster information Ω, which is expressed as follows:

$$\Omega = \{(C_i, \mathbf{d}(C_i))|i = 1, ..., c\}, \tag{1}$$

$$\mathbf{d}(C_i) = \{d(\mathbf{x}_i^j, C_i)|j = 1, ..., |C_i|\}. \tag{2}$$

1. This work has been published in the Springer Applied Intelligence: https://doi.org/10.1007/s10489-021-03139-y.

Algorithm 1 The hybrid sample selection strategy

Input: Ω, n_q, ρ_i
Parameter: Q_{center}, $Q_{boundary}$, and n_{q_i}
Output: The set of queried samples Q_U

$Q_U = \emptyset$
for i = 1 to n_C **do**
    Calculate the number of queries for cluster i: $n_{q_i} = \lfloor \frac{|C_i|}{n_U} \times n_q \rceil$
    Perform the center-based query with diversity exploration to obtain Q_{center} using equation 3.
    Perform the bi-cluster boundary-based query with diversity exploration to obtain $Q_{boundary}$ using equation 6
    $Q_U = Q_U \cup \{Q_{center} \cup Q_{boundary}\}$
end for
return Q_U

Here, C_i refers to the center of the i^{th} cluster, c denotes the total number of discovered clusters, and $d(\mathbf{x}_i^j, C_i)$ is the distance from sample $\mathbf{x}_i^j$ to its respective cluster center C_i. The cardinality $|C_i|$ denotes the number of samples that belong to C_i.

2.2 Distance-based sample selection

Hybrid sample selection strategy. After the clustering procedure, ALCS employs a novel hybrid sample selection strategy for active label query. Let the number of queried samples from the i_{th} cluster be n_{q_i}. The bi-cluster boundary-based selection step takes $\lfloor n_{q_i} \times \rho_i \rceil$ samples from cluster i as $Q_{boundary}$ where ρ_i denotes the sampling weight from the boundary of two adjacent clusters. Accordingly, the center-based selection policy chooses the remaining $\lfloor n_{q_i} \times (1 - \rho_i) \rceil$ samples from the center region as Q_{center}. The value of ρ_i ranges from zero to one. Algorithm 1 summarizes the hybrid sample selection procedure.

Center-based sample selection. For center-based selection, the query priority of each sample is computed in terms of **C**luster **R**epresentativeness (CR). Let $CR(*)$ and $P(*)$ be the cluster representativeness and query priority functions for clustered samples, respectively. For center-based selection, the query priority of x_i^j is calculated below.

$$P(x_i^j) = CR(x_i^j), \tag{3}$$

and

$$CR(x_i^j) = \frac{1}{1 + e^{d(x_i^j, C_i)}}. \tag{4}$$

Where $d(x_i^j, C_i)$ refers to the distance from x_i^j to C_i. From equation 4, the representativeness of each sample is inversely proportional to its distance to C_i and samples that are close to the cluster center have higher representativeness.

Bi-cluster boundary-based sample selection. We propose an effective bi-clusters boundary-based selection strategy to identify the most uncertain samples using the distance to their assigned cluster center and neighboring cluster center. This strategy utilizes the law of cosines to query the most informative samples from the cross-boundary region with two adjacent cluster centers. Assume the candidate bi-boundary sample in the i^{th} cluster is $x_{CB_i}^j$ and the candidate set is $CB = \{x_{CB_i}^j | j = 1, ..., \frac{|C_i|}{2}\}$. The two adjacent cluster centers are denoted as NC_1 and NC_2. The query priority of $x_{CB_i}^j$ is calculated

using the **C**luster **U**ncertainty (CU), which is expressed as follows:

$$P(x^j_{CB_i}) = CU(x^j_{CB_i}), \tag{5}$$

and

$$CU(x^j_{CB_i}) = \frac{1}{1 + e^{\frac{d_1+d_2}{d_{ref_1}+d_{ref_2}}}}. \tag{6}$$

Where d_{ref_1} and d_{ref_2} denote the distance from C_i to its two neighboring cluster centers. Here, d_1 refers to the distance from $x^j_{CB_i}$ to NC_1 and d_2 refers to the distance from $x^j_{CB_i}$ to NC_2, respectively. Detailed derivations can be found in (Yan et al., 2022). According to equation 6, a sample is considered to have higher uncertainty when it has a larger CU.

Diversity exploration. We developed a diversity exploration strategy based on fitness proportionate niching (FPN) (Workineh and Homaifar, 2012) to guide the search for informative and representative samples. Let X_{C_i} be a set of samples that belongs to C_i, the query priority function is expressed as follows.

$$P(X_{C_i}) = \begin{cases} CR(X_{C_i}), & \text{query from centers;} \\ CU(X_{C_i}), & \text{query from boundaries.} \end{cases} \tag{7}$$

Based on the priority function, the proposed diversity exploration procedure aims to decompose X_{C_i} into a number of small niches and query a set of diverse samples from different niches. During the sample selection procedure, the sample with the highest query priority is inserted into the queried sample set initially. Then, a niche can be formed by a set of samples in the neighborhood of this sample and a priority sharing strategy is employed to decrease the query priorities of other samples in the niche. The average distance for all k-nearest-neighbor graphs within a cluster is used as the neighborhood radius. As a rule of thumb, we set the value of k to be the square root of cluster size. Assume n^j_i and $X_{n^j_i}$ denote the j^{th} niche in C_i and a set of samples belong to n^j_i, respectively. Equation 8 describes the priority sharing function.

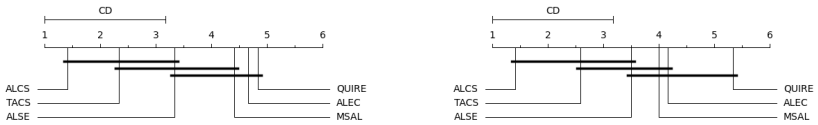
$$P(X_{n^j_i}) = \frac{P(X_{n^j_i})}{\sum P(X_{n^j_i})}, x_i \in n^j_i. \tag{8}$$

From equation 8, samples from the same niche will have relatively low priorities during the next sample query stage. Consequently, it guarantees to query more diverse samples from each cluster.

3. Experiments and results

Datasets and compared methods. Twelve benchmark datasets from (Dheeru and Karra Taniskidou, 2017) are used in the experiments. We compared ALCS with five state-of-the-art clustering-based AL methods, including QUIRE (Huang et al., 2010), ALEC (Wang et al., 2017a), active learning through multi-standard optimization (MSAL) (Wang et al., 2019), active learning through label error statistical (ALSE) (Wang et al., 2020), and three-way active learning through clustering selection (TACS) (Min et al., 2020). The implementation of the ALCS method, using python, is available at a Github repository .[2]

2. https://github.com/XuyangAbert/ALCS



(a) *Acc* (b) F_{mac}

Figure 2: Comparison of ALCS against other clustering-based AL methods with the Nemenyi test with $\alpha = 0.05$.

Result discussions. Table 1 compares the performance of ALCS with five clustering-based AL approaches using 10% labeled data. It is observed that ALCS provides better performance in most datasets, and it has the highest average ranks for both *Acc* and F_{mac}. In Australian, Aggregation, Spambase, Waveforms, Electricity, Penbased, GasSensor, and MNIST datasets, ALCS outperforms the other five clustering-based AL methods on both Acc and Fmac metrics. These results imply the efficacy of ALCS in handling datasets with highly overlapped classes. The Nemenyi post-hoc test is performed with a significance level of 0.05 and results are shown in Figure 2. Figure 2 displays that ALCS is statistically better than ALEC, QUIRE, and MSAL methods in terms of *Acc* and F_{mac}. On the other hand, ALCS presented statistically comparable performance with TACS and ALSE.

Table 1: Performance comparison of ALCS and five clustering-based AL methods.

Dataset	Metrics	ALEC	QUIRE	MSAL	ALSE	TACS	ALCS
R15	*Acc*	84.58(6)	**99.26(1)**	99.14(2)	86.27(5)	98.45(4)	99.07(3)
	F_{mac}	84.09(6)	**99.21(1)**	98.27(3)	83.94(5)	97.66(4)	99.06(2)
Australian	*Acc*	80.80(5)	81.29(4)	68.78(6)	81.38(3)	82.08(2)	**83.31(1)**
	F_{mac}	79.71(6)	80.87(3)	68.69(6)	80.82(4)	80.92(2)	**83.06(1)**
Aggregation	*Acc*	91.06(5)	71.01(6)	91.25(4)	91.91(3)	92.74(2)	**99.43(1)**
	F_{mac}	76.86(5)	44.21(6)	76.92(4)	77.63(3)	78.15(2)	**99.09(1)**
Vehicle	*Acc*	46.11(6)	53.23(3)	48.92(4)	46.39(5)	53.45(2)	**54.74(1)**
	F_{mac}	54.66(3)	49.52(6)	**55.12(1)**	52.37(5)	54.83(2)	54.30(4)
Spambase	*Acc*	76.48(4)	75.73(5)	75.32(6)	76.57(3)	79.58(2)	**81.54(1)**
	F_{mac}	75.85(3)	74.79(6)	75.87(5)	75.46(4)	80.28(2)	**80.92(1)**
Waveforms	*Acc*	75.42(5)	75.87(4)	75.32(6)	76.89(3)	78.17(1)	76.66(2)
	F_{mac}	75.84(3)	74.91(6)	75.47(4)	75.12(5)	76.52(2)	**76.62(1)**
Electricity	*Acc*	82.81(5)	82.48(6)	83.01(3)	83.22(2)	82.88(4)	**85.34(1)**
	F_{mac}	80.47(3)	79.89(6)	80.44(5)	80.83(2)	80.56(4)	**83.76(1)**
DLA0.01	*Acc*	86.27(5)	72.14(6)	92.48(4)	93.18(3)	**99.22(1)**	93.61(2)
	F_{mac}	86.28(5)	72.51(6)	86.98(4)	87.15(3)	**97.98(1)**	88.26(2)
Penbased	*Acc*	87.94(5)	82.74(6)	89.48(3)	88.13(4)	91.24(2)	**94.80(1)**
	F_{mac}	86.98(5)	72.68(6)	88.04(4)	89.01(3)	91.03(2)	**94.76(1)**
GasSensor	*Acc*	64.94(5)	64.40(6)	65.79(4)	66.44(3)	66.88(2)	**72.81(1)**
	F_{mac}	61.95(5)	60.60(6)	62.84(4)	63.74(3)	64.25(2)	**71.55(1)**
DCCC	*Acc*	**76.88(1)**	75.15(5)	74.85(6)	75.26(4)	75.45(3)	76.43(2)
	F_{mac}	54.16(4)	44.21(6)	49.56(5)	57.35(2)	54.95(3)	**60.84(1)**
MNIST	*Acc*	87.58(4)	84.52(6)	87.12(5)	88.45(2)	87.75(3)	**91.83(1)**
	F_{mac}	87.15(2)	83.48(6)	86.54(4)	86.81(3)	84.07(5)	**91.79(1)**
Avg. ranks	*Acc*	4.67	4.83	4.42	3.33	2.33	**1.42**
	F_{mac}	4.17	5.33	4.08	3.50	2.58	**1.41**

4. Concluding remarks

In this paper, we presented a novel active learning framework using clustering-based sampling to handle the shortage of prior label information. It utilizes the FPS-clustering proce-

dure to explore the structure of unlabeled data without exhaustive parameter tuning. A new distance-based sample selection procedure with an effective diversity exploration strategy was developed to enhance the quality of queried labels. Experimental results established that ALCS provided better or comparable performance than the five clustering-based AL approaches without tuning the clustering parameters.

Merits, Limitations & Future work. As a new clustering-based AL framework, ALCS effectively handles the dependency on clustering parameters and offers a promising solution to improve the diversity among queried labels. Moreover, the bi-cluster boundary selection strategy is designed to enhance the learning performance in datasets with highly overlapped classes. Limitations of the ALCS can be summarized from two aspects: (i) the imbalance among different class distributions is not considered in ALCS; and (ii) ALCS is currently limited to offline AL problems; Therefore, our future work will focus on addressing these two limitations of the ALCS framework.

Acknowledgments

We would like to acknowledge support for this project from the Air Force Research Laboratory and the Office of the Secretary of Defense under agreement number (FA8750-15-2-0116). This work is also partially supported by National Science Foundation and NASA University Leadership Initiative (ULI) under grant number (2000320) and (80NSSC20M0161).

References

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215, 2008.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. In *Advances in neural information processing systems*, pages 892–900, 2010.

Seho Kee, Enrique Del Castillo, and George Runger. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences*, 454:401–418, 2018.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

Fan Min, Shi-Ming Zhang, Davide Ciucci, and Min Wang. Three-way active learning through clustering selection. *International Journal of Machine Learning and Cybernetics*, 11(5):1033–1046, 2020.

Min Wang, Fan Min, Zhi-Heng Zhang, and Yan-Xue Wu. Active learning through density clustering. *Expert systems with applications*, 85:305–317, 2017a.

Min Wang, Ke Fu, and Fan Min. Active learning through two-stage clustering. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7. IEEE, 2018.

Min Wang, Ying-Yi Zhang, and Fan Min. Active learning through multi-standard optimization. *IEEE Access*, 7:56772–56784, 2019.

Min Wang, Ke Fu, Fan Min, and Xiuyi Jia. Active learning through label error statistical methods. *Knowledge-Based Systems*, 189:105140, 2020.

Ran Wang, Xi-Zhao Wang, Sam Kwong, and Chen Xu. Incorporating diversity and informativeness in multiple-instance active learning. *IEEE transactions on fuzzy systems*, 25 (6):1460–1475, 2017b.

Abrham Workineh and Abdollah Homaifar. Fitness proportionate niching: Maintaining diversity in a rugged fitness landscape. In *Proceedings of the International Conference on Genetic and Evolutionary Methods (GEM)*, pages 1–7. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2012.

Yanshan Xiao, Zheng Chang, and Bo Liu. An efficient active learning method for multi-task learning. *Knowledge-Based Systems*, 190:105137, 2020.

Xuyang Yan, Abdollah Homaifar, Shabnam Nazmi, and Mohammad Razeghi-Jahromi. A novel clustering algorithm based on fitness proportionate sharing. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, pages 1960–1965. IEEE, 2017.

Xuyang Yan, Shabnam Nazmi, Biniam Gebru, Mohd Anwar, Abdollah Homaifar, Mrinmoy Sarkar, and Kishor Datta Gupta. A clustering-based active learning method to query informative and representative samples. *Applied Intelligence*, pages 1–18, 2022.