

RESEARCH ARTICLE

Active Learning for Ordinal Classification Based on Adaptive Diversity-Based Uncertainty Sampling

DENIU HE 

Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

e-mail: d170201005@stu.cqupt.edu.cn

This work was supported by the Chongqing Key Laboratory of Computational Intelligence.


ABSTRACT Although a plethora of research has been conducted on active learning, little research attention has been focused on active learning for ordinal classification. Traditional multi-class active learning methods are typically designed for nominal multi-class classification. Therefore, they usually perform unsatisfactorily in ordinal classification settings. In ordinal classification, the cost of misclassifying an instance into an adjacent class is naturally lower than that of misclassifying it into a more disparate class. This principle is called ordering information. However, traditional active learning methods typically do not consider this ordering information during query selection. This paper proposes a novel adaptive hybrid active learning method for ordinal classification by considering the ordering information. In the proposed method, an uncertainty measure is introduced to select the hard-to-predict instances distributed between adjacent classes. In addition, a diversity measure is incorporated with the uncertainty measure to alleviate the potential sampling redundancy. Finally, an expected cost minimization measure with ordering information is designed. This measure balances the contributions of the uncertainty and diversity measures and prompts the algorithm to select the instances most likely to decrease the misclassification cost of the model. Extensive experiments on eleven public ordinal datasets demonstrate the superiority of the proposed method over several state-of-the-art methods.

INDEX TERMS Active learning, ordinal classification, uncertainty sampling, diversity, ordering information.

I. INTRODUCTION

Ordinal classification, also called ordinal regression, is a particular case of multi-class classification problem where the instances are labeled by ordinal scales, i.e., there is a natural total ordering among the output variables [1]. For instance, in movie ratings, customers can specify preferences by selecting, for each movie, one of several rating levels, such as one through five “stars” [2]. The ratings have a meaningful order that distinguishes ordinal classification from nominal multi-class classification. Since ordering information among classes makes sense in many real-world situations, ordinal classification has a wide range of research fields, such as

credit rating in the banking industry [3], age estimation in the computer vision field [4], and medical treatment in the medical domain [5]. To train an effective ordinal classification model, it is critical that one have and rely on a sufficient amount of reliable labeled instances. However, labeling ordinal classification instances is time-consuming and expensive because it depends on user preferences and domain expertise. As a result, labeling a large number of instances is usually prohibitively expensive. In this circumstance, it would be desirable to induce an ordinal classifier using an active learning technique. Active learning [6], [7], [8] seeks to address the above issue by constructing algorithms that can guide the labeling of a small set of instances, such that the generalization ability of the classifier is maximized while minimizing the labeling cost. This study aims to develop an

The associate editor coordinating the review of this manuscript and approving it for publication was Wanqing Zhao .

effective active learning method to construct a satisfactory ordinal classifier with a limited query budget in a pool-based setting.

As a way to relieve the tedious work of manual annotation, active learning is crucial in various machine learning problems [6], including classification [7], [8], clustering [9], regression [10], [11], recommendation [12], and so forth. Although a large number of active learning algorithms have been developed in the literature over the past few decades, little attention has been dedicated to the problem of active learning for ordinal classification. Most existing active learning methods for classification are aimed at binary problems. The few active learning algorithms suitable for multi-class problems are typically designed for standard nominal multi-class classification and usually perform unsatisfactorily in ordinal classification settings.

For ordinal classification problems, the misclassification costs are not the same for different errors [1]. Specifically, the cost of misclassifying an instance as an adjacent class is naturally lower than that of misclassifying it as a more disparate class [1], [5]. This ordering information reflects the structure of multiple-level discrete ordinal labels and usually comes along with a V-shaped cost vector [13]. Many previous studies have confirmed that the ordering information benefits constructing an accurate ordinal prediction model [1], [14], [15]. Therefore, it is conjectured that the ordering information can guide the query selection in active learning for ordinal classification.

This paper proposes an adaptive hybrid active learning method by integrating an uncertainty measure and a diversity measure. These two measures are combined by considering the ordering information among classes. In the proposed method, the base learner is called the reduced logistic ordinal classification model, which is instantiated from a reduction-based ordinal classification framework [16] by leveraging the logistic loss function. The base learner follows the threshold-based ordinal classification scheme [1] and reduces the ordinal classification problem to an easy-to-handle binary classification problem. In addition, the mean square error estimate on the reduced logistic ordinal classification model can surrogate the V-shaped cost in ordinal classification and represent the ordering information among classes. Thus, the proposed method uses the mean square error estimate to introduce the ordering information into query selection.

In many previous multi-class active learning methods, informative (hard-to-predict) instances are usually identified by defining uncertainty measures with the one-versus-one or one-versus-rest schemes [8], [17], where a nominal multi-class classification setting is considered by default. But, this way is not well applicable to the active learning problem for ordinal classification. In ordinal classification data, the hard-to-predict instances are usually distributed in regions between adjacent classes. Therefore, this paper designs a margin-based uncertainty measure tailored to the reduced logistic ordinal classification model to select instances dis-

tributed between adjacent classes. Since multiple separating hyperplanes exist in the reduced logistic ordinal classification model, the potential unbalanced hyperplane-updating problem may occur during active learning. Therefore, a threshold-cyclic sampling mechanism is introduced to control the uncertainty measure to evenly allocate the query resources to the multiple separating hyperplanes to mitigate the potential unbalanced hyperplane-updating problem. Nevertheless, uncertainty sampling is susceptible to the problem of sampling redundancy since uncertain instances are likely to be similar to each other [18]. Therefore, in the proposed method, a diversity measure is incorporated with the uncertainty measure to make the currently selected instance differs from the already labeled instances. In the previous studies [17], [19], the two measures are usually combined by a trade-off parameter. The value of this parameter needs to be set by the user based on a priori knowledge. However, prior knowledge is not available in most situations. Besides, the best values of the trade-off parameter may vary along with the active learning process. Therefore, how to provide an adaptive trade-off between uncertainty and diversity measures remains a challenging issue.

It is known that ordinal classification typically focuses on decreasing the misclassification cost by taking into account the ordering information. Therefore, an expected cost minimization measure is designed to balance the uncertainty and diversity measures. This measure can impose the active learner to select the informative instances that are most likely to decrease the misclassification cost of the model.

The main contributions of this paper are summarized as follows.

- This paper is the first study of adaptive hybrid active learning for ordinal classification. The proposed method innovatively introduces the ordering information into query selection to guide the combination of an uncertainty measure and a diversity measure to select useful instances adaptively in each iteration.
- A margin-based uncertainty measure tailored to the commonly used threshold-based ordinal classification scheme is designed for selecting informative instances distributed between adjacent classes. Besides, a threshold-cyclic sampling mechanism is introduced along with the uncertainty measure to mitigate the potential unbalanced hyperplane-updating problem. In addition, a diversity measure is incorporated with the uncertainty measure to alleviate the potential sampling redundancy problem.
- An expected cost minimization measure that imbues the ordering information is designed. This measure provides an adaptive trade-off between the uncertainty and diversity measures, thereby prompting the algorithm to select unlabeled instances that are most likely to reduce the misclassification cost of the model.
- Extensive experiments on eleven ordinal datasets demonstrate that the proposed method is superior to several state-of-the-art baseline methods.

The remainder of this paper is organized as follows. Section II reviews the related works. Section III describes the details of the proposed method. The experimental setup and the experimental results are reported in Section IV. Finally, a brief conclusion is presented in Section V.

II. RELATED WORK

Active learning aims to induce a reliable prediction model while minimizing labeling costs. Since active learning interactively queries labels from annotators, it can be viewed as a case of human-in-the-loop [20]. The critical component of an active learning algorithm is the query selection strategy. Existing query selection strategies can be roughly divided into informativeness sampling and representativeness sampling.

Strategies that concentrate on assessing instances' informativeness include uncertainty sampling [21], [22], [23], query-by-committee [24], expected change [25], and so on. Uncertainty sampling selects the instances for which the current classifier is least certain [26]. For instance, the simple margin approach [26] selects instances closest to the decision boundary. This approach is typically applied to binary classification problems. In the context of multi-class classification, methods such as maximum entropy [22], margin sampling [21], and least confidence [23] have been suggested. The query-by-committee strategy typically trains a set of prediction models and selects unlabeled instances on which the models disagree the most [24]. However, the potential bias among multiple prediction models may limit the performance of the query-by-committee-based approaches [27]. Methods based on expected change include the expected model change [28], expected model output change [29], expected error reduction [25], and so on. These methods estimate the change caused by each unlabeled instance being assigned the possible labels and weight the change by its estimated probability [27]. Therefore, most of these methods are computationally expensive. Although uncertainty sampling is an earlier proposed active learning strategy, it is still widely used in practice due to its simple intrinsic mechanism and fast running speed. Lookman *et al.* [30] recently introduced an uncertainty sampling method to the problem of materials discovery. Kim and Kim [31] designed an uncertainty-based active learning method based on a Bayesian neural network for reliable fire detection systems. Recently, a theoretical study for the convergence of uncertainty sampling has been provided by Raj and Bach [32].

Active learning strategies that assess instances' representativeness include clustering-based active learning [33], [34], optimal experimental design [11], [35], [36], diversity sampling [10], and so on. Clustering-based active learning employs a particular clustering algorithm to explore the clustering structure of the data and selects instances that represent the intrinsic geometry of the data. Although clustering-based active learning methods are suitable for multi-class active learning problems, it remains unclear how these algorithms perform when the clustering result is not sufficiently accu-

rate [18]. Optimal experimental design is a set of well-known approaches for representativeness sampling. These methods typically select instances that minimize the variance of the model parameters by relying on a particular data reconstruction framework [35]. Diversity sampling approaches aim to select unlabeled instances that differ from already labeled instances. It is known that diversity measures alone usually produce misleading results [6]. Therefore, diversity measures are usually incorporated with other measures to provide complementary information. However, it is still challenging to balance the contribution of a diversity measure when combining it with other measures.

It is known that no single active learning strategy can consistently perform well on any dataset it encounters. Therefore, many researchers are dedicated to developing synthesis active learning approaches [11]. Brinker [19] first proposed to combine diversity and uncertainty in active learning. By integrating uncertainty and diversity measures, Wang *et al.* [17] proposed an active learning approach for multi-class classification. However, the above two methods depend on a fixed trade-off parameter value to balance the uncertainty and diversity measures. Since the parameter values need to be determined by the users, the applications of these approaches are restricted. Yang *et al.* [18] proposed a multi-class active learning approach by uncertainty sampling with diversity maximization. Although this method does not rely on trade-off parameters, it is only suitable for batch mode active learning problems. Additionally, this method only considers the diversity inside the current batch but ignores the diversity between the currently selected and already labeled instances.

Although great progress has been made in active learning research on classification, little effort has been devoted to the problem of active learning for ordinal classification. The active learning algorithm for ordinal classification was first studied by exploiting monotonicity constraints [37]. This approach, however, is only suitable for monotonic classification problems [38], which is a special case of generic ordinal classification problems [38]. Recently, Li *et al.* [36] proposed an A-optimal experimental design criterion for the active ordinal classification problem. However, one apparent shortcoming of this method is that it must calculate the inverse of a large matrix in each iteration for each unlabeled instance. When the data dimension is large, the prohibitive computational cost will limit its usability. Ge *et al.* [39] tried to solve the imbalanced ordinal classification problem by extending an uncertainty sampling criterion to a threshold-based ordinal classification model. One immediate problem with this method is that the uncertainty sampling criterion may suffer from sampling redundancy. Additionally, an unbalanced hyperplane-updating problem among multiple hyperplanes may occur. To the best of our knowledge, these two works are the only two active learning studies focused on ordinal classification. Note that these two methods failed to consider the ordering information in query selection and stick to a single query selection criterion in the active learning process, which limits their active learning performance in ordinal

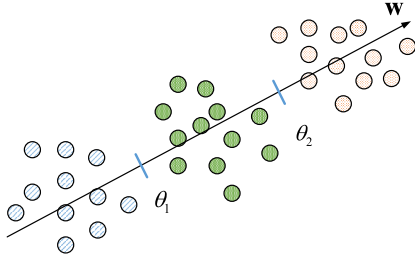


FIGURE 1. Toy example of a three-class ordinal dataset (denoted by blobs of different colors) under the threshold-based ordinal classification structure.

classification. The abovementioned situation encourages this study to develop a more effective active learning method that combines uncertainty and diversity measures by considering ordering information.

III. PROPOSED METHOD

This section provides the technical details of the proposed active learning method for ordinal classification. The proposed method will be described starting from the base learner instantiation.

A. REDUCED LOGISTIC ORDINAL CLASSIFICATION MODEL

This study designs the base learner based on the reduction-based ordinal classification framework proposed by Li and Lin [16] and a logistic function. The base learner is called the reduced logistic ordinal classification model (shorted as RLOC). The reduction-based ordinal classification framework is a general framework that reduces ordinal classification problems into binary classification problems and provides a unified view for many threshold-based ordinal classification models.

Let $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training set, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i -th feature vector, which corresponds to an ordinal label $y_i \in \mathcal{Y} = \{C_1, \dots, C_K\}$. There is total ordering among the classes, such as $C_1 < \dots < C_K$, where “ $<$ ” denotes a particular ordering relation, and K is the number of classes. Among the classes, C_k and C_{k+1} ($1 \leq k < K$) refer to as a pair of adjacent classes. Without loss of generality, the K ordinal scale labels are typically represented by K consecutive integers, such as $\{1, 2, \dots, K\}$. Generally, the threshold-based ordinal classification scheme performs prediction by learning $K - 1$ ordered thresholds: $\theta_1 \leq \theta_2 \leq \dots \leq \theta_{K-1}$, and $\theta_0 = -\infty$ and $\theta_K = +\infty$ are typically assumed. Then, for an unobserved instance \mathbf{x} , it is classified as C_k in the case that the prediction output $\mathbf{w}^T \mathbf{x}$ falls in the interval of $(\theta_{k-1}, \theta_k]$, where $\mathbf{w} \in \mathbb{R}^d$ is the learned weight vector, and $\mathbf{w}^T \mathbf{x}$ represents the projection of \mathbf{x} on \mathbf{w} . Figure 1 illustrates a three-class ordinal dataset under the threshold-based ordinal classification structure.

The reduction-based ordinal classification framework reduces the ordinal problem into multiple binary classification problems. These binary classification problems are

TABLE 1. Examples of extended binary instances generation for original training instances of different classes (the number of classes $K = 5$).

$\langle \mathbf{x}_i, y_i \rangle$	$y_i = 1$	$y_i = 2$	$y_i = 3$	$y_i = 4$	$y_i = 5$
$\mathbf{x}_i^1 = (\mathbf{x}_i, 1, 0, 0, 0)$	$y_i^1 = 0$	$y_i^2 = 1$	$y_i^3 = 1$	$y_i^4 = 1$	$y_i^5 = 1$
$\mathbf{x}_i^2 = (\mathbf{x}_i, 0, 1, 0, 0)$	$y_i^1 = 0$	$y_i^2 = 0$	$y_i^3 = 1$	$y_i^4 = 1$	$y_i^5 = 1$
$\mathbf{x}_i^3 = (\mathbf{x}_i, 0, 0, 1, 0)$	$y_i^1 = 0$	$y_i^2 = 0$	$y_i^3 = 0$	$y_i^4 = 1$	$y_i^5 = 1$
$\mathbf{x}_i^4 = (\mathbf{x}_i, 0, 0, 0, 1)$	$y_i^1 = 0$	$y_i^2 = 0$	$y_i^3 = 0$	$y_i^4 = 0$	$y_i^5 = 1$

solved jointly through a single binary classifier [16]. The reduction framework extends each original training instance $\langle \mathbf{x}_i, y_i \rangle$ into the following $K - 1$ binary training instances:

$$\begin{aligned} & \langle \mathbf{x}_i^k, y_i^k \rangle, \quad k = 1, \dots, K - 1, \\ & \mathbf{x}_i^k = (\mathbf{x}_i, \mathbf{e}_k) \in \mathbb{R}^{d+K-1}, \\ & y_i^k = 1 - I[y_i \leq k] \end{aligned} \quad (1)$$

where $\mathbf{e}_k \in \mathbb{R}^{K-1}$ is an extension vector with the k -th element as value 1 and the rest of the elements all as zeros. As shown in Eq. (1), $I[\cdot]$ is an indicator function that returns 1 if the inside condition is true; otherwise, zero is returned. Therefore, each extended instance \mathbf{x}_i^k is associated with a binary label $y_i^k \in \{0, 1\}$. Table 1 shows the examples of extended binary instances generation.

In the extended binary classification problem, the weight vector as follows

$$\bar{\mathbf{w}} = (\mathbf{w}, -\theta) \in \mathbb{R}^{d+K-1} \quad (2)$$

is learned to predict the output of \mathbf{x}_i^k , such that $\bar{\mathbf{w}} \mathbf{x}_i^k = (\mathbf{w}, -\theta)^T \mathbf{x}_i^k = \mathbf{w}^T \mathbf{x}_i - \theta_k$, where $\theta = [\theta_1, \theta_2, \dots, \theta_{K-1}]$. The projection $\bar{\mathbf{w}} \mathbf{x}_i^k$ can be interpreted as the distance from \mathbf{x}_i to the k -th decision hyperplane, and of course, it can be a negative value. Now, one can introduce the logistic function into the reduction framework and formulate the cumulative conditional probability for $y_i > k$, i.e., $\pi(\mathbf{x}_i^k) = P(y_i > k | \mathbf{x}_i)$, as

$$\begin{aligned} \pi(\mathbf{x}_i^k) &= \frac{\exp(\bar{\mathbf{w}}^T \mathbf{x}_i^k)}{1 + \exp(\bar{\mathbf{w}}^T \mathbf{x}_i^k)} \\ &= \frac{\exp(\mathbf{w}^T \mathbf{x}_i - \theta_k)}{1 + \exp(\mathbf{w}^T \mathbf{x}_i - \theta_k)}, \end{aligned} \quad (3)$$

where $k = 1, \dots, K - 1$. The logistic function is employed in consideration of the log-likelihood and its derivatives can be easily computed using calculations similar to those in standard logistic regression. Based on $\pi(\mathbf{x}_i^k)$, the predictive ordinal label of instance \mathbf{x}_i can be calculated as

$$\hat{y}_i = \sum_{k=1}^{K-1} I[\pi(\mathbf{x}_i^k) \geq 0.5], \quad (4)$$

where $I[\cdot]$ is an indicator function that returns 1 if the inner condition is true; otherwise, zero is returned. Eq. (4) can be viewed as a decoding procedure.

Given the training set \mathcal{L} , the l_2 regularized log-likelihood function for the extended binary classification problem can

be formulated as

$$l(\bar{\mathbf{w}}; \mathcal{L}) = \sum_{i=1}^n \sum_{k=1}^{K-1} (y_i^k \log \pi(\mathbf{x}_i^k) + (1 - y_i^k) \log(1 - \pi(\mathbf{x}_i^k))) - \frac{\lambda}{2} \|\bar{\mathbf{w}}\|_2^2. \quad (5)$$

The optimal values of the weight vector \mathbf{w} and the thresholds $[\theta_1, \theta_2, \dots, \theta_{K-1}]$ can be obtained jointly by maximizing the log-likelihood function in Eq. (5). This optimization problem can be solved by the Newton-Raphson maximization procedure. In addition, one can ensure that the thresholds $[\theta_1, \theta_2, \dots, \theta_{K-1}]$ are ordered.

Theorem 1: By maximizing the log-likelihood function defined in Eq. (5), the optimal solution (\mathbf{w}^, θ^*) satisfies $\theta_1^* \leq \theta_2^* \leq \dots \leq \theta_{K-1}^*$.*

Proof: For an optimal solution (\mathbf{w}^*, θ^*) , assume that $\theta_k^* > \theta_{k+1}^*$ for some k . One only need to prove that switching θ_k^* and θ_{k+1}^* would not decrease the objective value of Eq. (5). One can consider the change of the objective value in the following three cases.

First, for an instance \mathbf{x}_i with $y_i = k + 1$. In this case, $y_i^k = 1$ and $y_i^{k+1} = 0$, switching the thresholds changes the objective value by

$$\Delta l = \log \frac{1 + \exp(\theta_k^* - \mathbf{w}^{*T} \mathbf{x}_i)}{1 + \exp(\theta_{k+1}^* - \mathbf{w}^{*T} \mathbf{x}_i)} + \log \frac{1 + \exp(\mathbf{w}^{*T} \mathbf{x}_i - \theta_{k+1}^*)}{1 + \exp(\mathbf{w}^{*T} \mathbf{x}_i - \theta_k^*)}. \quad (6)$$

Because $\theta_k^* - \mathbf{w}^T \mathbf{x} > \theta_{k+1}^* - \mathbf{w}^T \mathbf{x}$, it is not difficult to find that the change Δl is non-negative.

Second, for an instance \mathbf{x}_i with $y_i < k + 1$. In this case, $y_i^k = 0$ and $y_i^{k+1} = 0$, switching the thresholds changes the objective value by

$$\Delta l = \log \frac{1 + \exp(\mathbf{w}^{*T} \mathbf{x}_i - \theta_{k+1}^*)}{1 + \exp(\mathbf{w}^{*T} \mathbf{x}_i - \theta_k^*)} + \log \frac{1 + \exp(\mathbf{w}^{*T} \mathbf{x}_i - \theta_k^*)}{1 + \exp(\mathbf{w}^{*T} \mathbf{x}_i - \theta_{k+1}^*)} = 0. \quad (7)$$

Third, for an instance \mathbf{x}_i with $y_i > k + 1$. In this case, $y_i^k = 1$ and $y_i^{k+1} = 1$. Analogously, switching the thresholds does not change the objective value in this case.

Since the Δl for the instances of the three cases are all non-negative, the claim is justified. \square

The conclusion in Theorem 1 will be used later. The time complexity of the RLOC model is $\mathcal{O}(\xi n(K-1)(d+K-1)^2)$, where ξ is the number of iterations required for the optimization procedure to converge. The structure of the RLOC model is similar to the all-threshold logistic model in [2]. But, the thresholds $[\theta_1, \theta_2, \dots, \theta_{K-1}]$ in our model can be obtained jointly as a part of the extended weight vector $\bar{\mathbf{w}}$ in the binary logistic regression problem. Therefore, the RLOC model inherits the theoretical rigors of the logistic regression model.

B. ADAPTIVE DIVERSITY-BASED UNCERTAINTY MEASURE

This subsection focuses on designing a robust diversity-based uncertainty sampling method based on the RLOC model. There are $K - 1$ parallel separating hyperplanes in the RLOC model. Intuitively, the hard-to-predict instances in ordinal data are usually located between adjacent classes and are typically close to one of the separating hyperplanes. Therefore, this paper extends the simple margin criterion [26] to the multiple hyperplane setting.

Let $\mathcal{U} = \{\mathbf{x}_i\}_{i=1}^m$ be the unlabeled instance pool. Given an unobserved instance $\mathbf{x}_i \in \mathcal{U}$, the absolute distance from it to the k -th separating hyperplane in the RLOC model can be computed as

$$\begin{aligned} \Delta(\mathbf{x}_i, \theta_k) &= |\mathbf{w}^T \mathbf{x}_i - \theta_k| \\ &= |\bar{\mathbf{w}}^T \mathbf{x}_i^k|, \\ &k = 1, \dots, K - 1. \end{aligned} \quad (8)$$

It is clear that the less the value of $\Delta(\mathbf{x}_i, \theta_k)$, the more uncertain of \mathbf{x}_i with respect to the k -th separating hyperplane. Therefore, the proposed method employs $\Delta(\mathbf{x}_i, \theta_k)$ to assess the uncertainty of $\mathbf{x}_i \in \mathcal{U}$.

Since there are $K - 1$ hyperplanes in the RLOC model, a threshold-cyclic sampling mechanism is introduced to prevent the potential unbalanced hyperplane-updating problem. According to this mechanism, the query resources are equally allocated to the $K - 1$ separating hyperplanes in a round-robin manner. Thus, every successive $K - 1$ query instances are sequentially selected concerning the $K - 1$ different hyperplanes. Accordingly, the unlabeled candidate instance set with respect to the k -th hyperplane can be represented as

$$\mathcal{U}_k = \{\mathbf{x}_i \in \mathcal{U} | \Delta(\mathbf{x}_i, \theta_k) < \Delta(\mathbf{x}_i, \theta_h), h \neq k\}. \quad (9)$$

The instances in \mathcal{U}_k are closer to the k -th separating hyperplane than to the other hyperplanes.

To alleviate the potential sampling redundancy, it makes sense to incorporate a diversity measure with the above uncertainty measure. The diversity measure as follows

$$Div(\mathbf{x}_i, \mathcal{L}) = \min_{\mathbf{x}_j \in \mathcal{L}} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2, \quad \mathbf{x}_i \in \mathcal{U}_k, \quad (10)$$

is adopted in the proposed method. The larger the value of $Div(\mathbf{x}_i, \mathcal{L})$, the more difference of \mathbf{x}_i from the labeled instances. Therefore, the proposed method combines the uncertainty measure and the diversity measure as follows

$$DU(\mathbf{x}_i; \theta_k) = \frac{Div(\mathbf{x}_i, \mathcal{L})^\alpha}{(1 + \Delta(\mathbf{x}_i, \theta_k))^{(1-\alpha)}}, \quad k = 1, \dots, K - 1, \quad (11)$$

where $0 \leq \alpha \leq 1$ is a trade-off parameter that controls the contributions of the uncertainty and diversity measures. When $\alpha = 0$, the measure $DU(\mathbf{x}_i; \theta_k)$ degenerates to an uncertainty sampling measure. Note that the optimal value of α depends on the data. However, in an active learning setting, there are not enough labeled instances to divide out a validation set for parameter tuning.

To solve the above problem, the proposed method uses the ordering information to guide the cooperation of the uncertainty and diversity measures. The ordering information among labels can be represented as a V-shaped misclassification cost vector \mathbf{c} as follows [13]

$$\begin{cases} \mathbf{c}[y, k-1] > \mathbf{c}[y, k], & \text{for } 2 \leq k < y; \\ \mathbf{c}[y, k+1] > \mathbf{c}[y, k], & \text{for } y < k \leq K-1. \end{cases} \quad (12)$$

where the $\mathbf{c}[y, k]$ represents the cost of misclassifying a y class instance as a k class instance. In what follows, an expected cost minimization measure that imbues the ordering information is designed to guide the cooperation of the two measures. The expected cost minimization selects the instance for which, if labeled, is most likely to minimize the misclassification cost of the RLOC model on all the unlabeled instances in \mathcal{U} . In each iteration, the α in Eq. (11) is set as a series of given values $[0.0, 0.1, 0.2, \dots, 1.0]$. Thus, an informative candidate instance set S is obtained. After that, the query instance is selected from S based on the expected cost minimization measure. This is equivalent to approximately determining the value of parameter α .

The proposed method uses the mean square error in the extended binary problem to surrogate the misclassification cost of each original instance. Therefore, the misclassification cost for $\mathbf{x}_i \in \mathcal{U}$ is defined as

$$MSE(\mathbf{x}_i) = \frac{1}{K-1} \sum_{k=1}^{K-1} (y_i^k - \pi(\mathbf{x}_i^k))^2. \quad (13)$$

It is not difficult to prove that $MSE(\mathbf{x}_i)$ can represent the ordering information among the ordinal labels.

Theorem 2: Given a particular instance \mathbf{x}_i , the mean square error over its extended binary instances in the RLOC model is V-shaped.

Proof: According to Theorem 1, one can prove that $\theta_1 \leq \dots \leq \theta_{K-1}$. Therefore, the prediction outputs of \mathbf{x}_i in the extended binary problem maintain that $\pi(\mathbf{x}_i^1) \geq \dots \geq \pi(\mathbf{x}_i^{K-1})$. If \mathbf{x}_i is predicted to belong to class h , one has

$$\begin{cases} \pi(\mathbf{x}_i^t) > 0.5, & \text{for } 1 \leq t < h \leq K-1; \\ \pi(\mathbf{x}_i^t) < 0.5, & \text{for } 1 \leq h < t \leq K-1. \end{cases} \quad (14)$$

Suppose $1 \leq h < t \leq K-2$ and h is the misclassified label. Therefore, According to Eq. (3) and Eq. (4), one can obtain $\pi(\mathbf{x}_i^t) < 0.5$. In the case that the true label is t , according to Eq. (13), the mean square error for the $K-1$ extended binary instances is

$$\begin{aligned} & MSE(\mathbf{x}_i|h, t) \\ &= \frac{1}{K-1} \left(\sum_{k=1}^{K-1} \pi(\mathbf{x}_i^k)^2 + (t-1) - 2 \sum_{k=1}^{t-1} (\pi(\mathbf{x}_i^k)) \right). \end{aligned} \quad (15)$$

In the case that the true label is $t+1$, according to Eq. (13), the mean square error for the $K-1$ extended binary instances is

$$MSE(\mathbf{x}_i|h, t+1)$$

$$= \frac{1}{K-1} \left(\sum_{k=1}^{K-1} \pi(\mathbf{x}_i^k)^2 + (t) - 2 \sum_{k=1}^t (\pi(\mathbf{x}_i^k)) \right). \quad (16)$$

Thus, one can obtain

$$\begin{aligned} & MSE(\mathbf{x}_i|h; t+1) - MSE(\mathbf{x}_i|h; t) \\ &= \frac{1}{K-1} (1 - 2\pi(\mathbf{x}_i^t)) > 0. \end{aligned} \quad (17)$$

Suppose $2 \leq t < h \leq K$ and h is the misclassified label. According to Eq. (3) and Eq. (4), one can obtain $\pi(\mathbf{x}_i^{t-1}) > 0.5$. Analogously, one can obtain

$$\begin{aligned} & MSE(\mathbf{x}_i|h; t-1) - MSE(\mathbf{x}_i|h; t) \\ &= \frac{1}{K-1} (2\pi(\mathbf{x}_i^{t-1}) - 1) > 0. \end{aligned} \quad (18)$$

Therefore, the mean square error over the extended binary instances in the RLOC model is V-shaped, and it makes sense to surrogate the ordering information among the ordinal labels. \square

According to the works in [40] and [41], the mean square error in a logistic regression model can be estimated as the sum of a squared bias and a variance. Consequently, the expected cost of the RLOC model on \mathcal{U} can be formulated as

$$\begin{aligned} EC(\mathcal{U}; \mathcal{L}) &= \sum_{i=1}^{|\mathcal{U}|} \sum_{k=1}^{K-1} \mathbb{E}_{\mathcal{L}}[MSE(\mathbf{x}_i^k)] \\ &= \sum_{i=1}^{|\mathcal{U}|} \sum_{k=1}^{K-1} \mathbb{E}_{\mathcal{L}}[\pi(\mathbf{x}_i^k) - \hat{\pi}(\mathbf{x}_i^k)]^2 \\ &= \sum_{i=1}^{|\mathcal{U}|} \sum_{k=1}^{K-1} (\mathbb{E}_{\mathcal{L}}[\pi(\mathbf{x}_i^k) - \hat{\pi}(\mathbf{x}_i^k)])^2 \\ &\quad + \mathbb{E}_{\mathcal{L}}[\hat{\pi}(\mathbf{x}_i^k) - \mathbb{E}_{\mathcal{L}}[\hat{\pi}(\mathbf{x}_i^k)]]^2, \end{aligned} \quad (19)$$

where $\hat{\pi}(\mathbf{x}_i^k)$ is the predictive output of the extended binary instance \mathbf{x}_i^k based on the RLOC model given the training set \mathcal{L} , $\pi(\mathbf{x}_i^k)$ indicates the expected real output of \mathbf{x}_i^k , and $\mathbb{E}_{\mathcal{L}}[\cdot]$ represents the expectation over the labeled set \mathcal{L} . In Eq. (19), the term $\mathbb{E}_{\mathcal{L}}[\pi(\mathbf{x}_i^k) - \hat{\pi}(\mathbf{x}_i^k)]$ means the bias of the estimation, and the term $\mathbb{E}_{\mathcal{L}}[\hat{\pi}(\mathbf{x}_i^k) - \mathbb{E}_{\mathcal{L}}[\hat{\pi}(\mathbf{x}_i^k)]]^2$ indicates the variance of the estimation. For simplicity, denote by $Bias(\hat{\pi}(\mathbf{x}_i^k))$ and $Var(\hat{\pi}(\mathbf{x}_i^k))$ the bias and variance of $\hat{\pi}(\mathbf{x}_i^k)$, respectively.

Given an extended binary instance \mathbf{x}_i^k , by taking $z_i^k = \bar{\mathbf{w}}^T \mathbf{x}_i^k$ as a variable, one can obtain the first-order Taylor series expansion of the function $\pi(\mathbf{x}_i^k)$ about z_i^k as

$$\begin{aligned} \hat{\pi}(\mathbf{x}_i^k) &= \pi(\mathbf{x}_i^k) + \pi(\mathbf{x}_i^k)(1 - \pi(\mathbf{x}_i^k))(\hat{\mathbf{w}} - \bar{\mathbf{w}})^T \mathbf{x}_i^k \\ &\quad + o\left(\left\| \hat{\mathbf{w}}^T \mathbf{x}_i^k - \bar{\mathbf{w}}^T \mathbf{x}_i^k \right\|\right), \end{aligned} \quad (20)$$

where $\hat{\mathbf{w}}$ is the first-order approximation for the extended weight vector given the training set \mathcal{L} , and $\bar{\mathbf{w}}$ is the expected real extended weight vector. Thus, the first-order approximation for the bias of $\hat{\pi}(\mathbf{x}_i^k)$ is formulated as

$$Bias(\hat{\pi}(\mathbf{x}_i^k)) = \mathbb{E}_{\mathcal{L}}[\pi(\mathbf{x}_i^k) - \hat{\pi}(\mathbf{x}_i^k)]$$

$$\begin{aligned}
 &= \mathbb{E}_{\mathcal{L}}[\pi(\mathbf{x}_i^k)(1 - \pi(\mathbf{x}_i^k))(\hat{\mathbf{w}} - \bar{\mathbf{w}})^T \mathbf{x}_i^k] \\
 &= \pi(\mathbf{x}_i^k)(1 - \pi(\mathbf{x}_i^k))(\mathbb{E}[\hat{\mathbf{w}}] - \bar{\mathbf{w}})^T \mathbf{x}_i^k \\
 &= \pi(\mathbf{x}_i^k)(1 - \pi(\mathbf{x}_i^k))(\text{Bias}(\hat{\mathbf{w}}))^T \mathbf{x}_i^k, \quad (21)
 \end{aligned}$$

and the variance of $\hat{\pi}(\mathbf{x}_i^k)$ is approximated as

$$\begin{aligned}
 \text{Var}(\hat{\pi}(\mathbf{x}_i^k)) &= \mathbb{E}_{\mathcal{L}}[\hat{\pi}(\mathbf{x}_i^k) - \mathbb{E}_{\mathcal{L}}[\hat{\pi}(\mathbf{x}_i^k)]]^2 \\
 &= \pi(\mathbf{x}_i^k)^2(1 - \pi(\mathbf{x}_i^k))^2 \mathbf{x}_i^k{}^T \text{Cov}(\hat{\mathbf{w}}) \mathbf{x}_i^k, \quad (22)
 \end{aligned}$$

where $\text{Bias}(\hat{\mathbf{w}})$ and $\text{Cov}(\hat{\mathbf{w}})$ are the bias and covariance matrix of $\hat{\mathbf{w}}$, respectively, and both of them can be derived approximately from the Taylor series expansion for the log-likelihood function $l(\bar{\mathbf{w}}; \mathcal{L})$ about $\bar{\mathbf{w}}$.

The negative Hessian matrix of the log-likelihood function in Eq. (5) is

$$F(\bar{\mathbf{w}}; \mathcal{L}) = \sum_{i=1}^{|\mathcal{L}|} \sum_{k=1}^{K-1} \pi(\mathbf{x}_i^k)(1 - \pi(\mathbf{x}_i^k)) \mathbf{x}_i^k \mathbf{x}_i^k{}^T + \lambda I, \quad (23)$$

which is also referred to as the observed Fisher information matrix, where $I \in \mathbb{R}^{(d+K-1) \times (d+K-1)}$ is an identity matrix. According to the work in [42], the approximate bias of $\hat{\mathbf{w}}$ is

$$\text{Bias}(\hat{\mathbf{w}}) = -\lambda F(\bar{\mathbf{w}}; \mathcal{L})^{-1} \bar{\mathbf{w}}, \quad (24)$$

and the approximate covariance matrix of $\hat{\mathbf{w}}$ is

$$\text{Cov}(\hat{\mathbf{w}}) = F(\bar{\mathbf{w}}; \mathcal{L})^{-1} (F(\bar{\mathbf{w}}; \mathcal{L}) - \lambda I) F(\bar{\mathbf{w}}; \mathcal{L})^{-1}. \quad (25)$$

Therefore, the expected misclassification cost of the RLOC model on \mathcal{U} given the training set \mathcal{L} can be calculated by Eq. (26), as shown at the bottom of the next page.

To determine which candidate instance in S can minimize the expected misclassification cost of the RLOC model based on Eq. (26), one can estimate the expected cost for each candidate instance analogous to the calculation of expected error in reference [25]. But, in this way, one needs to re-train the model by adding each candidate instance with its possible labels into the training set in the manner of one-step-look-ahead. This is prohibitively computationally expensive. Besides, the probability estimate is usually inaccurate in the active learning situation where there are very few labeled instances in \mathcal{L} . Therefore, this paper seeks to calculate the expected cost for each candidate instance in an alternative way. It is clear that the estimation of the expected cost depends on the Fisher information matrix $F(\bar{\mathbf{w}}; \mathcal{L})$. Since the Fisher information matrix does not rely on the target variable, one can update it for each candidate instance in an unsupervised way as follows

$$\begin{aligned}
 F(\bar{\mathbf{w}}; \mathcal{L}, \mathbf{x}_j) \\
 = F(\bar{\mathbf{w}}; \mathcal{L}) + \sum_{k=1}^{K-1} \pi(\mathbf{x}_j^k)(1 - \pi(\mathbf{x}_j^k)) \mathbf{x}_j^k \mathbf{x}_j^k{}^T, \quad \mathbf{x}_j \in S. \quad (27)
 \end{aligned}$$

The above method avoids frequent model retraining and thus saves a lot of computational overhead. Subsequently, the expected cost for each candidate instance $\mathbf{x}_j \in S$ can be

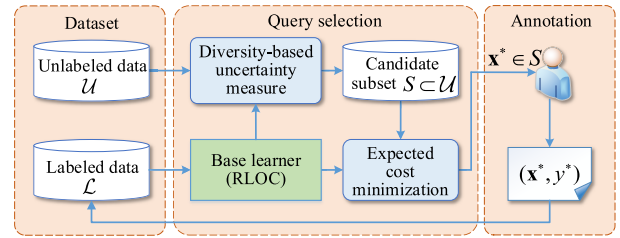


FIGURE 2. Workflow of the adaptive diversity-based uncertainty sampling method.

computed by Eq. (28), as shown at the bottom of the next page. Then, the query instance is determined as

$$\mathbf{x}^* = \arg \min_{\mathbf{x}_j \in S} EC(\mathcal{U}; \mathcal{L}, \mathbf{x}_j). \quad (29)$$

Suppose $|\mathcal{U}| \gg d + K - 1$. Thus, for each candidate instance $\mathbf{x}_j \in S$, the time complexity of computing the expected cost is $EC(\mathcal{U}; \mathcal{L}, \mathbf{x}_j)$ is $\mathcal{O}(|\mathcal{U}|(K-1)(d+K-1)^2 + (d+K-1)^3) = \mathcal{O}(|\mathcal{U}|(d+K-1)^2)$.

C. ALGORITHM AND COMPLEXITY ANALYSES

The algorithmic procedure of the proposed active learning method is presented in Algorithm 1. The workflow of the proposed method is diagrammed in Figure 2, which facilitates the understanding of the algorithm procedure.

Algorithm 1 Adaptive Diversity-Based Uncertainty Sampling for Ordinal Classification (ADUS)

Require: Initial training set \mathcal{L} ; unlabeled instance pool \mathcal{U} ; query budget B ; the number of classes K .

Ensure: The expanded training set \mathcal{L} , and the trained RLOC model.

- 1: Train the RLOC model with \mathcal{L} ; $b \leftarrow B$;
- 2: **while** $b > 0$ **do**
- 3: $k \leftarrow 1$;
- 4: **while** $k \leq K - 1$ and $b > 0$ **do**
- 5: Retrain the RLOC model on \mathcal{L} ;
- 6: Collect the instance set \mathcal{U}_k based on Eq. (9);
- 7: Obtain the candidate subset $S \subseteq \mathcal{U}_k$ based on the diversity-based uncertainty measure in Eq. (11) with $\alpha = [0.0, 0.1, 0.2, \dots, 1.0]$;
- 8: Obtain the most informative instance $\mathbf{x}^* = \arg \min_{\mathbf{x}_j \in S} EC(\mathcal{U}; \mathcal{L}, \mathbf{x}_j)$ based on Eq. (28);
- 9: Inquire \mathbf{x}^* 's label y^* from the annotator; $b \leftarrow b - 1$;
- 10: $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\mathbf{x}^*\}$; $\mathcal{L} \leftarrow \mathcal{L} \cup (\mathbf{x}^*, y^*)$; $k \leftarrow k + 1$;
- 11: **end while**
- 12: **end while**

Suppose there are $n = |\mathcal{L}|$ training instances and $m = |\mathcal{U}|$ unlabeled instances, and $m \gg n$. Therefore, in the current iteration, retrain the RLOC model in line 5 requires $\mathcal{O}(\xi n(K-1)(d+K-1))$ time. In line 6, to collect the instance set \mathcal{U}_k , one shall first calculate $\Delta(\mathbf{x}_i, \theta_k)$ for all the unlabeled instances based on Eq. (8), which requires $\mathcal{O}(m(K-1)(d+K-1))$ time. Then, \mathcal{U}_k is obtained by Eq. (9), which costs $\mathcal{O}(m(K-1))$ time. In the worst situation, there are m instances in \mathcal{U}_k . Thus, in line 7, obtaining the candidate subset S needs the computational time of $\mathcal{O}(m)$. Suppose $m \gg |S|$. In line 8,

to obtain the most informative candidate instance, one needs first to update the Fisher information matrix and compute its inverse, which requires $\mathcal{O}(n(K-1)(d+K-1)^2)$ time and $\mathcal{O}((d+K-1)^3)$ time, respectively. Computing $\pi(\mathbf{x}_i^k)$ for all the unlabeled extended binary instances requires $\mathcal{O}(m(K-1)(d+K-1))$ time. Execute Eq. (28) for the candidate instances in S requires $\mathcal{O}(m(K-1)(d+K-1)^2)$ time.

Suppose $m \gg d+K-1$, the time complexity for each query selection in the worst situation is $\mathcal{O}(m(d+K-1)^2)$.

IV. EXPERIMENT

In this section, the performance of the proposed method is examined by comparing it with several state-of-the-art baseline methods on eleven public datasets. All the experiments were conducted on a Windows 10 64-bit operating system with 32GB RAM and an Intel(R) Core(TM) i7-8700 CPU@3.2GHz processor. The programming language is Python. The source codes are publicly available at <https://github.com/DeniuHe/ADUS>.

A. DATASETS AND EXPERIMENTAL SETUP

Table 2 summarizes the details of the eleven used datasets. HDI [43] is the human development index data that contains 179 countries classified into four rating levels. ARWU2020 is the assessment data of the top 990 world universities in 2020 released by ShanghaiRanking Consultancy. The datasets Obesity and PowerPlant are from the UCI dataset repository. The other seven datasets are from reference [1]. The datasets ARWU2020 and PowerPlant were originally regression data, which have been discretized into ordinal data by an equal frequency bin operation [1]. Before the experiments, all the datasets were pre-processed by the following z-score standardization:

$$x_{ij} = \frac{x_{ij} - \text{mean}(x_j)}{\text{std}(x_j)}, \quad (30)$$

where x_{ij} denotes the j -th feature value of instance \mathbf{x}_i , and $\text{mean}(x_j)$ and $\text{std}(x_j)$ are the mean value and the standard deviation of the j -th feature, respectively.

To verify the effectiveness of the proposed method (denoted as **ADUS**), the experiment compares it with the following ten state-of-the-art baseline methods and one designed method.

- **Random** is the random sampling method.
- **USME** is an uncertainty sampling method instantiated based on the RLOC model and the maximum entropy strategy [22].
- **USLC** is an uncertainty sampling method instantiated based on the RLOC model and the least confidence strategy [23].
- **USMS** is an uncertainty sampling method instantiated based on the RLOC model and the margin-based sampling strategy [21].
- **FISTA** [35] is a transductive experimental design method based on an exclusive sparsity norm.
- **ALCE** [44] is a multi-class active learning algorithm based on a cost embedding approach.
- **McPAL** [45] is a multi-class probabilistic active learning method that selects unlabeled instances with maximal probabilistic gain.
- **MCSVMA** [8] is an SVM-based multi-class active learning method that selects unlabeled instances by considering the criteria of rejection, compatibility, and uncertainty.
- **ALOR** [39] is an active ordinal classification method based on a reduced SVM model. This method selects the unlabeled instance with the smallest distance from the nearest decision boundary.
- **LogitA** [36] is an A-optimal experimental design method for ordinal classification, which tends to select representative unlabeled instances.
- **ADUS_n** is a designed method similar to ADUS, except that there is no threshold-cyclic instance selection mechanism. This method is included in the comparison to verify whether the threshold-cyclic instance selection mechanism benefits the ADUS method.

In the experiments, each dataset is divided into an unlabeled pool (80% of the data) and a testing set (20% of the

$$\begin{aligned} EC(\mathcal{U}; \mathcal{L}) &= \sum_{i=1}^{|\mathcal{U}|} \sum_{k=1}^{K-1} \text{Bias}(\pi(\mathbf{x}_i^k))^2 + \text{Var}(\pi(\mathbf{x}_i^k)) \\ &= \sum_{i=1}^{|\mathcal{U}|} \sum_{k=1}^{K-1} \left(\pi(\mathbf{x}_i^k)(1 - \pi(\mathbf{x}_i^k))(\lambda F(\bar{\mathbf{w}}; \mathcal{L})^{-1} \bar{\mathbf{w}})^T \mathbf{x}_i^k \right)^2 \\ &\quad + \pi(\mathbf{x}_i^k)^2(1 - \pi(\mathbf{x}_i^k))^2 \mathbf{x}_i^{kT} F(\bar{\mathbf{w}}; \mathcal{L})^{-1} (F(\bar{\mathbf{w}}; \mathcal{L}) - \lambda I) F(\bar{\mathbf{w}}; \mathcal{L})^{-1} \mathbf{x}_i^k. \end{aligned} \quad (26)$$

$$\begin{aligned} EC(\mathcal{U}; \mathcal{L}, \mathbf{x}_j) &= \sum_{i=1}^{|\mathcal{U}|} \sum_{k=1}^{K-1} \left(\pi(\mathbf{x}_i^k)(1 - \pi(\mathbf{x}_i^k))(\lambda F(\bar{\mathbf{w}}; \mathcal{L}, \mathbf{x}_j)^{-1} \bar{\mathbf{w}})^T \mathbf{x}_i^k \right)^2 \\ &\quad + \pi(\mathbf{x}_i^k)^2(1 - \pi(\mathbf{x}_i^k))^2 \mathbf{x}_i^{kT} F(\bar{\mathbf{w}}; \mathcal{L}, \mathbf{x}_j)^{-1} (F(\bar{\mathbf{w}}; \mathcal{L}, \mathbf{x}_j) - \lambda I) F(\bar{\mathbf{w}}; \mathcal{L}, \mathbf{x}_j)^{-1} \mathbf{x}_i^k. \end{aligned} \quad (28)$$

TABLE 2. Information of the used datasets.

No.	Datasets	#Instances	#Features	#Classes	Distribution
1	Balance-scale [1]	625	4	3	[288, 49, 288]
2	HDI [43]	179	3	4	[50, 75, 35, 19]
3	Car [1]	1728	21	4	[1210, 384, 69, 65]
4	ARWU2020 ¹	990	6	5	[198, 198, 198, 198, 198]
5	Bank-5bin [1]	8192	8	5	[1639, 1639, 1638, 1638, 1638]
6	Computer-5bin [1]	8192	12	5	[1639, 1639, 1638, 1638, 1638]
7	Automobile [1]	205	71	6	[3, 22, 67, 54, 32, 27]
8	Obesity ²	2111	16	7	[272, 287, 290, 290, 351, 297, 324]
9	Bank-10bin [1]	8192	8	10	[820, 820, 819, 819, 819, 819, 819, 819, 819, 819]
10	Computer-10bin [1]	8192	12	10	[820, 820, 819, 819, 819, 819, 819, 819, 819, 819]
11	PowerPlant ²	9568	4	10	[956, 956, 957, 957, 957, 957, 957, 957, 957, 957]

¹ <http://www.shanghairanking.com/>

² <https://archive.ics.uci.edu/ml/index.php>

data) through five-fold stratified cross-validation six times. Thus, there are a total of 30 runs for each dataset. The initial training set contains instances randomly selected one instance from each class in the unlabeled pool. For all the compared methods, the query budget is set as $25 \times K$, where K is the number of classes. At each iteration, an unlabeled instance is selected to query its label, and the RLOC model is retrained. Meanwhile, the ordinal classification performance of the retrained model is evaluated on the testing set. The active learning process stops when the query budget is exhausted. Finally, the average results of the 30 runs are reported.

The metrics Mean Zero-one Error (MZE), Mean Absolute Error (MAE), and Macro F1 score (F1) are employed in the experiments. MZE and MAE are commonly used to evaluate the performance of ordinal classification [1]. Macro F1 score is a common metric used to evaluate the performance of multi-class classification. In addition, the metric Area Under Learning Curve (AULC) [46] is employed to quantify the overall performance of the active learning algorithms. The trapezoidal approximation rule [46] is used to calculate the area under the learning curve of MZE (AULC-MZE), the area under the learning curve of MAE (AULC-MAE), and the area under the learning curve of F1 (AULC-F1). In general, the lower the values of AULC-MZE and AULC-MAE, the better the performance of an active learning algorithm. In contrast, for the metric AULC-F1, the larger the value, the better the performance of an active learning algorithm.

B. EXPERIMENTAL RESULTS

The learning curves of different methods on the metrics MZE, MAE, and F1 are plotted in Figures 3, 4, and 5, respectively. From the learning curves in the three figures, one can observe that the proposed method is among the top performers in the active learning process on most datasets. Since there are multiple methods in the comparison, some of the learning curves in the three figures are inevitably crossed and overlaid.

For quantitative comparison, the evaluation results of different methods on AULC-MZE, AULC-MAE, and AULC-F1 are reported in Tables 3, 4, and 5, respectively. The best results

are marked in boldface in the three tables, and the “AvgRank” denotes the average rank of the compared method based on the evaluation results on all the datasets. To detect whether a compared method performs significantly differently from ADUS on each dataset in terms of the three metrics, the experiment performed the Wilcoxon signed-rank test [47] between ADUS and the compared methods at a confidence level of 0.05. The marker “*” in the three tables indicates that there is a statistically significant difference between ADUS and the compared method on the corresponding dataset. The win/tie/loss counts of ADUS versus the compared methods based on the Wilcoxon signed-rank test are summarized in Table 6 to show the significant analysis results intuitively. The results in Tables 3, 4, and 5 show that the proposed method performs superior to the compared method on most datasets in terms of the three metrics. The significant analysis results in Table 6 show that there are statistically significant differences between ADUS and the ten baseline methods on most datasets, and the ADUS achieves zero loss in the Wilcoxon signed-rank test. It can be observed that there is no significant difference between ADUS and ADUS_n on some data sets, but the overall performance of ADUS is better than that of ADUS_n.

In order to verify the effectiveness of the adaptive mechanism in the proposed method, the experiment compares ADUS against the non-adaptive diversity-based uncertainty sampling method (denoted as **DUS- α**) that degenerates from ADUS. The α in “DUS- α ” denotes the trade-off parameter. In the comparison, the parameter α in DUS- α is set as [0.0, 0.1, 0.2, \dots , 1.0]. The average ranks of the different methods on the metrics of AULC-MZE, AULC-MAE, and AULC-F1 are recorded and plotted in Fig. 6. One can observe that ADUS is superior to the method DUS- α with different fixed parameter values.

C. DISCUSSION

The experimental results in subsection IV-B illustrate that the proposed method ADUS has advantages over the compared methods in active learning for ordinal classification. Multiple factors result in the outstanding performance of the proposed method:

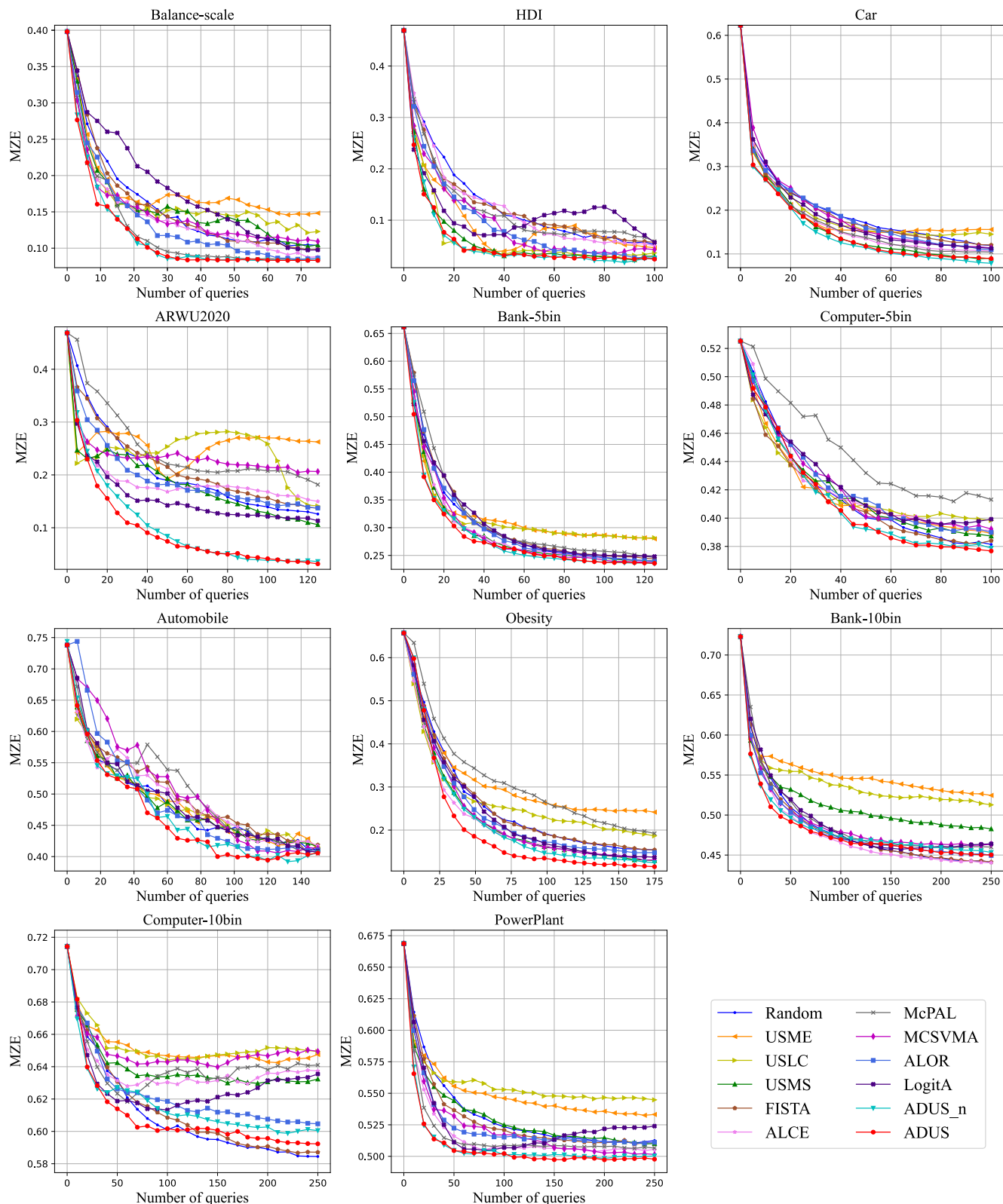


FIGURE 3. Learning curves of MZE for the twelve compared methods.

1) The hard-to-predict instances in ordinal data are usually located between adjacent classes. Labeling these instances is beneficial for improving the performance

of an ordinal classifier. The uncertainty sampling measure is designed by extending the idea of margin sampling to the RLOC model, which allows the proposed

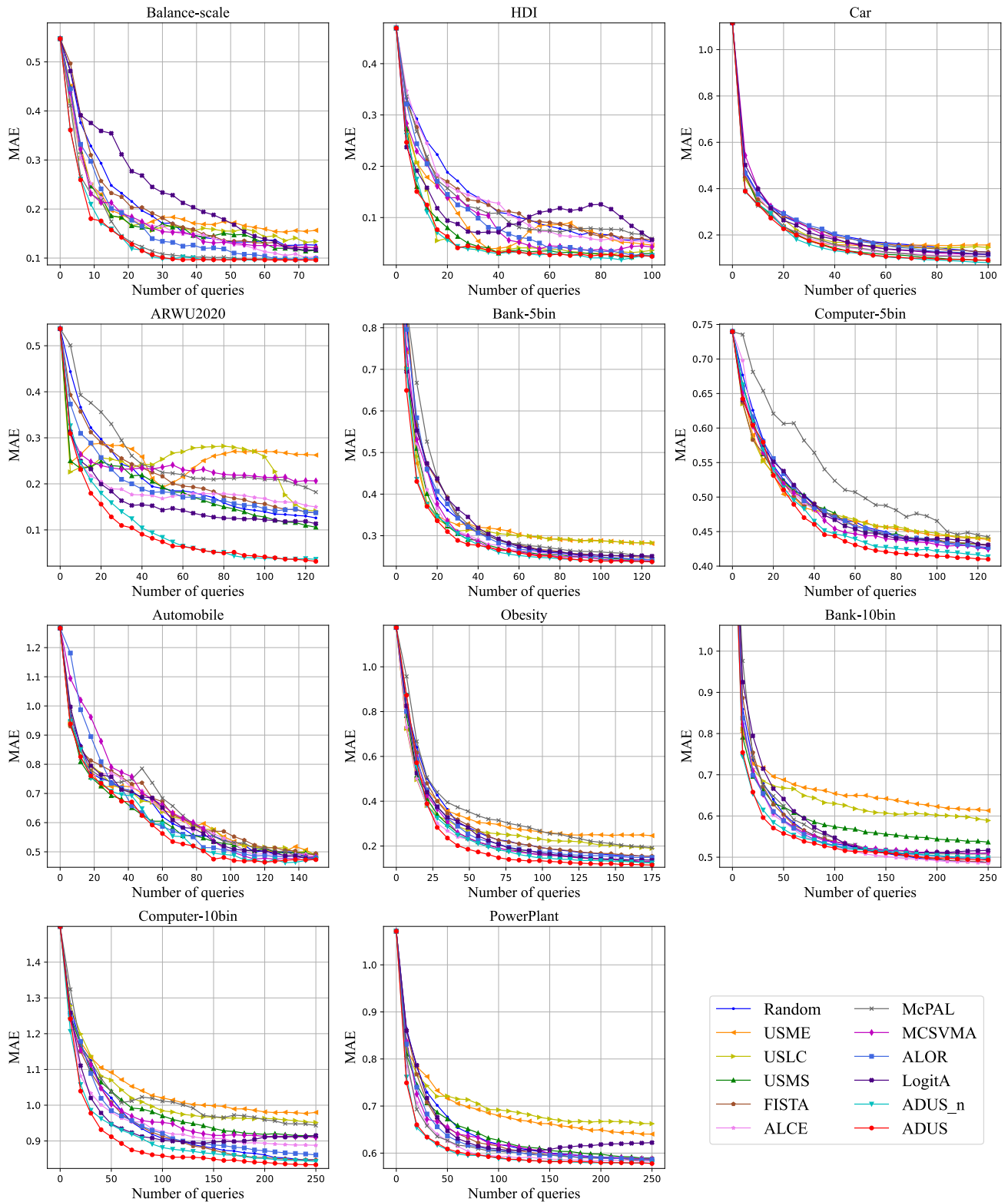


FIGURE 4. Learning curves of MAE for the twelve compared methods.

method to select instances distributed between adjacent classes.

2) A threshold-cyclic instance selection mechanism is introduced to evenly allocate the query resources to

the multiple separating hyperplanes, thus mitigating the potential unbalanced hyperplane-updating problem.

3) A diversity measure is combined with the uncertainty measure to alleviate the sampling redundancy

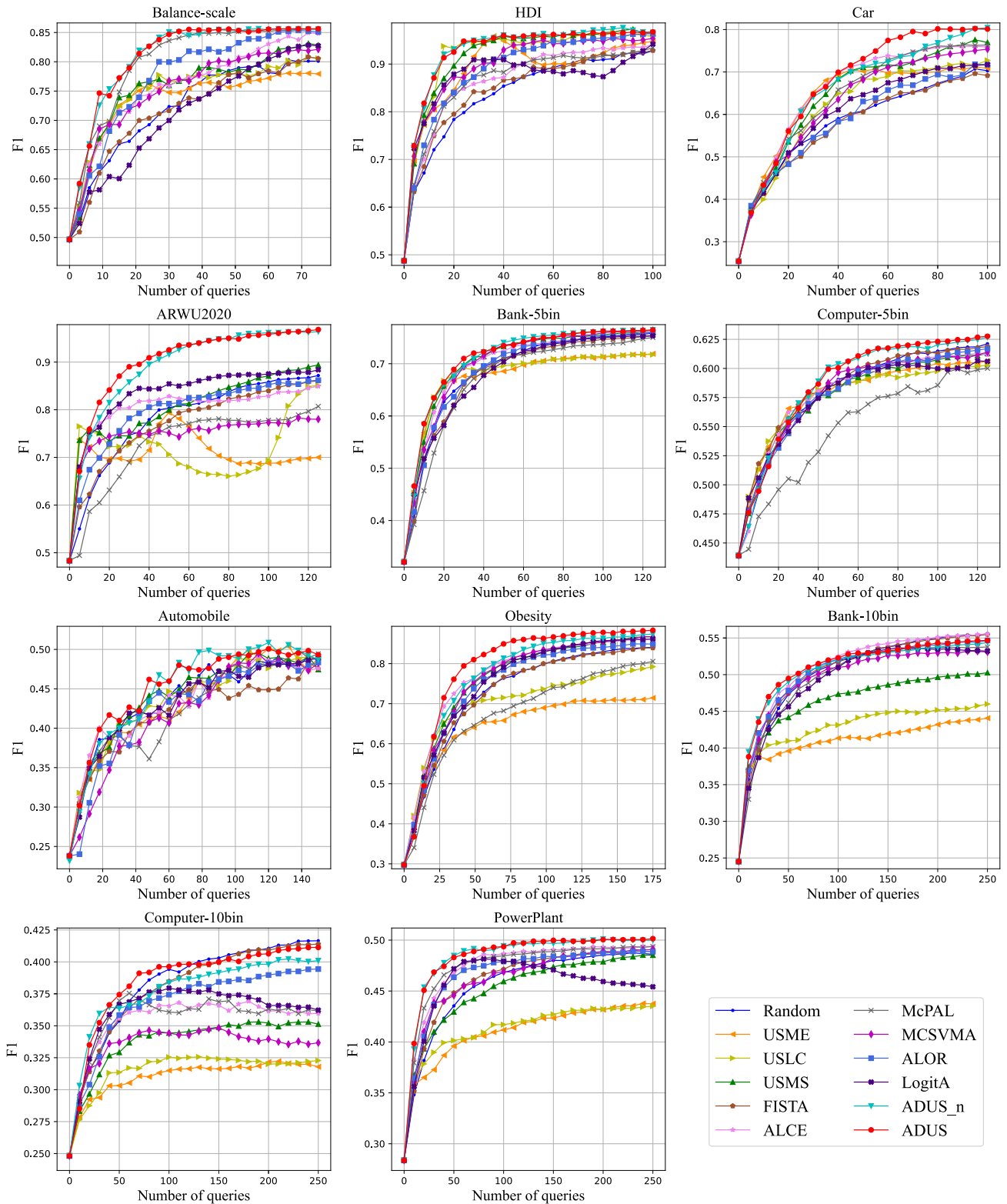


FIGURE 5. Learning curves of F1 for the twelve compared methods.

problem. In particular, an expected cost minimization measure that imbues the ordering information is designed to provide an adaptive trade-off between the uncertainty and diversity measures in each itera-

tion. This adaptive mechanism encourages the active learner to select informative instances that are most likely to decrease the misclassification cost of the model.

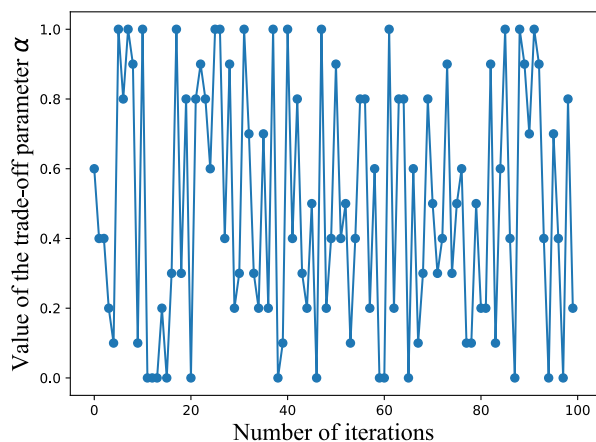


FIGURE 7. Values of α correspond to the selected instances in an active learning process by ADUS for the dataset Car. One can observe that the value of α is varying in different iterations.

convergence trend. Additionally, from the average ranks on AULC-MZE, AULC-MAE, and AULC-F1, one can find that the overall performance of LogitA is inferior to ALOR. These illustrate that instances' informativeness is necessary to be considered in active learning for ordinal classification. FISTA, ALCE, McPAL, and MCSVMA are active learning methods for nominal multi-class classification. Therefore, the performances of these methods are not as good as the proposed method. USME, USLC, and USME are three typical uncertainty sampling methods instantiated based on the RLOC model. Compared with USME and USLC, the method USMS prefers to select the instances between adjacent classes in ordinal data. One can see from the results in Tables 3, 4, and 5 that USMS performs superior to USME and USLC. This demonstrates that extending the margin-based sampling to the RLOC model in our method is reasonable. Additionally, since each of the three uncertainty sampling methods relies only on a single uncertainty measure, none of them performs outstandingly in the experiments. Compared to the proposed method, the designed method ADUS_n lacks the threshold-cyclic sampling mechanism. Although ADUS_n performs slightly better than ADUS on some datasets, there is no statistically significant difference. From Table 6, one can see that the overall performance of ADUS is better than that of ADUS_n. The above result indicates that the threshold-cyclic sampling mechanism is helpful in most cases.

The results depicted in Figure 6 show that the overall performance of ADUS is superior to that of the non-adaptive method DUS- α . This comparison further demonstrates the effectiveness of the adaptive mechanism. When $\alpha=0.0$, the DUS- α degenerates to an uncertainty sampling method. One can observe that DUS- $\alpha=0.0$ does not perform outstandingly in this comparison. This result indicates that a diversity measure is necessary when performing uncertainty sampling. As aforementioned, the optimal value of the trade-off parameter may vary in different iterations during the active learning process. The expected cost minimization measure provides an

adaptive trade-off between the uncertainty and diversity measures in each iteration. Figure 7 shows the trade-off parameter values in an active learning process by ADUS for the dataset Car. One can see that the value of α is adaptively adjusted in each iteration. This explains the outstanding performance of the proposed method.

V. CONCLUSION

This paper proposes an effective adaptive hybrid active learning method for ordinal classification. In contrast to nominal multi-class classification, ordinal classification usually focuses on decreasing the misclassification cost of the model by taking the ordering information into account. In addition, informative instances for ordinal classification typically lie within regions between adjacent classes. This paper designs a margin-based uncertainty measure tailored to the commonly used threshold-based ordinal classification scheme to select the hard-to-predict instances distributed between adjacent classes. Besides, a threshold-cyclic sampling mechanism is introduced along with the uncertainty measure to mitigate the potential unbalanced hyperplane-updating problem. Additionally, the proposed method incorporates a diversity measure with the uncertainty measure to enable the currently selected instance differs from the already labeled instances. In particular, an expected cost minimization measure that imbues the ordering information is designed to balance the uncertainty measure with the diversity measure adaptively. Thus, the proposed method can select instances most likely to reduce the model's misclassification cost. Finally, the proposed method was compared against several state-of-the-art baseline methods. The experimental results demonstrate the effectiveness of the proposed method. Compared with the most related method ALOR, the proposed method achieves an improvement of 14.53%, 16.20%, and 5.87% on the metrics AULC-MZE, AULC-MAE, and AULC-F1, respectively, when the query budget is twenty-five times the number of classes.

REFERENCES

- [1] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 127–146, Jan. 2016.
- [2] J. D. M. Rennie and N. Srebro, "Loss functions for preference levels: Regression with discrete ordered labels," in *Proc. IJCAI Multidisciplinary Workshop Adv. Preference Handling*, vol. 2167. Norwell, MA, USA: Kluwer, 2005, pp. 180–186.
- [3] G. Manthoulis, M. Doumpos, C. Zopounidis, and E. Galarotis, "An ordinal classification framework for bank failure prediction: Methodology and empirical evidence for U.S. banks," *Eur. J. Oper. Res.*, vol. 282, no. 2, pp. 786–801, Apr. 2020.
- [4] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec. 2020.
- [5] Z. Ma and J. Ahn, "Feature-weighted ordinal classification for predicting drug response in multiple myeloma," *Bioinformatics*, vol. 37, no. 19, pp. 3270–3276, Oct. 2021.
- [6] P. Kumar and A. Gupta, "Active learning query strategies for classification, regression, and clustering: A survey," *J. Comput. Sci. Technol.*, vol. 35, no. 4, pp. 913–945, Jul. 2020.

- [7] A. J. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class active learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* Miami, FL, USA, Jun. 2009, pp. 2372–2379.
- [8] H. Guo and W. Wang, "An active learning-based SVM multi-class classification model," *Pattern Recognit.*, vol. 48, no. 5, pp. 1577–1597, May 2015.
- [9] H. Yu, X. Wang, G. Wang, and X. Zeng, "An active three-way clustering method via low-rank matrices for multi-view data," *Inf. Sci.*, vol. 507, pp. 823–839, Jan. 2020.
- [10] D. Wu, C.-T. Lin, and J. Huang, "Active learning for regression using greedy sampling," *Inf. Sci.*, vol. 474, pp. 90–105, Feb. 2019.
- [11] S. H. Park and S. B. Kim, "Active semi-supervised learning with multiple complementary information," *Expert Syst. Appl.*, vol. 126, pp. 30–40, Jul. 2019.
- [12] Z. Zhang, X. Jin, L. Li, G. Ding, and Q. Yang, "Multi-domain active learning for recommendation," in *Proc. 13th AAAI Conf. Artif. Intell.*, D. Schuurmans and M. P. Wellman, Eds. Phoenix, AZ, USA: AAAI Press, Feb. 2016, pp. 2358–2364.
- [13] H.-T. Lin and L. Li, "Reduction from cost-sensitive ordinal ranking to weighted binary classification," *Neural Comput.*, vol. 24, no. 5, pp. 1329–1367, 2012.
- [14] H. Yan, "Cost-sensitive ordinal regression for fully automatic facial beauty assessment," *Neurocomputing*, vol. 129, pp. 334–342, Apr. 2014.
- [15] Y. Shi, P. Li, H. Yuan, J. Miao, and L. Niu, "Fast kernel extreme learning machine for ordinal regression," *Knowl.-Based Syst.*, vol. 177, pp. 44–54, Aug. 2019.
- [16] L. Li and H.-T. Lin, "Ordinal regression by extended binary classification," in *Proc. NIPS*, vol. 19, 2006, pp. 865–872.
- [17] Z. Wang, X. Fang, X. Tang, and C. Wu, "Multi-class active learning by integrating uncertainty and diversity," *IEEE Access*, vol. 6, pp. 22794–22803, 2018.
- [18] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.
- [19] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proc. 20th Int. Conf. Mach. Learn.* T. Fawcett and N. Mishra, Eds. Washington, DC, USA: AAAI Press, Aug. 2003, pp. 59–66.
- [20] A. Holzinger, "Interactive machine learning for health informatics: When do we need the human-in-the-loop?" *Brain Informat.*, vol. 3, no. 2, pp. 119–131, Jun. 2016.
- [21] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden Markov models for information extraction," in *Proc. Int. Symp. Intell. Data Anal.* in Lecture Notes in Computer Science, vol. 2189, F. Hoffmann, D. J. Hand, N. M. Adams, D. H. Fisher, and G. Guimarães, Eds. Cascais, Portugal: Springer, Sep. 2001, pp. 309–318.
- [22] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "Entropy-based active learning with support vector machines for content-based image retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jun. 2004, pp. 85–88.
- [23] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *Proc. 20th Nat. Conf. Artif. Intell. 17th Innov. Appl. Artif. Intell. Conf.*, M. M. Veloso and S. Kambhampati, Eds. Pittsburgh, PA, USA: AAAI Press, Jul. 2005, pp. 746–751.
- [24] J. Vandoni, E. Aldea, and S. L. Hégarat-Masclé, "Evidential query-by-committee active learning for pedestrian detection in high-density crowds," *Int. J. Approx. Reasoning*, vol. 104, pp. 166–184, Jan. 2019.
- [25] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, C. E. Brodley and A. P. Danyluk, Eds. Williamstown, MA, USA: Williams College, Jul. 2001, pp. 441–448.
- [26] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Mar. 2001.
- [27] Y. Xue and M. Hauskrecht, "Active learning of multi-class classification models from ordered class sets," in *Proc. 33rd AAAI Conf. Artif. Intell., AAAI 31st Innov. Appl. Artif. Intell. Conf. IAAI, 9th AAAI Symp. Educ. Adv. Artif. Intell., (EAAI)* Honolulu, HI, USA, Feb. 2019, pp. 5589–5596.
- [28] W. Cai, Y. Zhang, Y. Zhang, S. Zhou, W. Wang, Z. Chen, and C. H. Q. Ding, "Active learning for classification with maximum model change," *ACM Trans. Inf. Syst.*, vol. 36, no. 2, pp. 15:1–15:28, 2017.
- [29] C. Käding, A. Freytag, E. Rodner, A. Perino, and J. Denzler, "Large-scale active learning with approximations of expected model output changes," in *Proc. German Conf. Pattern Recognit.*, vol. 9796, B. Rosenhahn and B. Andres, Eds. Hannover, Germany: Springer, Sep. 2016, pp. 179–191.
- [30] T. Lookman, P. V. Balachandran, D. Xue, and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," *NPJ Comput. Mater.*, vol. 5, no. 1, pp. 1–17, Dec. 2019.
- [31] Y.-J. Kim and W.-T. Kim, "Uncertainty assessment-based active learning for reliable fire detection systems," *IEEE Access*, vol. 10, pp. 74722–74732, 2022.
- [32] A. Raj and F. R. Bach, "Convergence of uncertainty sampling for active learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 162, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds. Baltimore, MD, USA, 2022, pp. 18310–18331.
- [33] M. Wang, F. Min, Y.-X. Wu, and Z.-H. Zhang, "Active learning through density clustering," *Expert Syst. Appl.*, vol. 85, pp. 305–317, Nov. 2017.
- [34] D. He, H. Yu, G. Wang, and J. Li, "A two-stage clustering-based cold-start method for active learning," *Intell. Data Anal.*, vol. 25, no. 5, pp. 1169–1185, Sep. 2021.
- [35] X. Wang, Y. Huang, J. Liu, and H. Huang, "New balanced active learning model and optimization algorithm," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2826–2832.
- [36] J. Li, Z. Chen, Z. Wang, and Y.-C.-I. Chang, "Active learning in multiple-class classification problems via individualized binary models," *Comput. Statist. Data Anal.*, vol. 145, May 2020, Art. no. 106911.
- [37] P. Soons and A. Feelders, "Exploiting monotonicity constraints in active learning for ordinal classification," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 659–667.
- [38] P. A. Gutiérrez and S. García, "Current prospects on ordinal and monotonic classification," *Prog. Artif. Intell.*, vol. 5, no. 3, pp. 171–179, Aug. 2016.
- [39] J. Ge, H. Chen, D. Zhang, X. Hou, and L. Yuan, "Active learning for imbalanced ordinal regression," *IEEE Access*, vol. 8, pp. 180608–180617, 2020.
- [40] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: An evaluation," *Mach. Learn.*, vol. 68, no. 3, pp. 235–265, 2007.
- [41] J. Wang and E. Park, "Active learning for penalized logistic regression via sequential experimental design," *Neurocomputing*, vol. 222, pp. 183–190, Jan. 2017.
- [42] S. L. Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Appl. Statist.*, vol. 41, no. 1, pp. 191–201, 1992.
- [43] X. Liu, H. Yu, G. Wang, and L. Guo, "A multi-criteria ordered clustering algorithm based on PROMETHEE," in *Proc. Develop. Artif. Intell. Technol. Comput. Robot.*, Oct. 2020, pp. 43–51.
- [44] K.-H. Huang and H.-T. Lin, "A novel uncertainty sampling algorithm for cost-sensitive multiclass active learning," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 925–930.
- [45] D. Kottke, G. Kreml, D. Lang, J. Teschner, and M. Spiliopoulou, "Multi-class probabilistic active learning," in *Proc. ECAI 22nd Eur. Conf. Artif. Intell.*, vol. 285, Hague, The Netherlands: IOS Press, 2016, pp. 586–594.
- [46] O. Reyes, A. H. Altalhi, and S. Ventura, "Statistical comparisons of active learning strategies over multiple datasets," *Knowl. Based Syst.*, vol. 145, pp. 274–288, Apr. 2018.
- [47] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 6, pp. 80–83, Dec. 1945.



DENIU HE received the B.S. degree in network engineering from the Science and Technology College, North China Electric Power University, Baoding, China, in 2009, and the M.S. degree in computer science and technology from Guangxi Minzu University, Nanning, China, in 2012. He is currently pursuing the Ph.D. degree with the Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications. His current research interests include active learning, ordinal classification, and intelligent decision-making.

...