Abstract

Active learning is widely used to reduce labeling effort and training time by repeatedly querying only the most beneficial samples from unlabeled data. In real-world problems where data cannot be stored indefinitely due to limited storage or privacy issues, the query selection and the model update should be performed as soon as a new data sample is observed. Various online active learning methods have been studied to deal with these challenges; however, there are difficulties in selecting representative query samples and updating the model efficiently without forgetting. In this study, we propose Message Passing Adaptive Resonance Theory (MPART) that learns the distribution and topology of input data online. Through message passing on the topological graph, MPART actively queries informative and representative samples, and continuously improves the classification performance using both labeled and unlabeled data. We evaluate our model in stream-based selective sampling scenarios with comparable query selection strategies, showing that MPART significantly outperforms competitive models.

1. Introduction

The recent success of deep learning in the field of visual object recognition and speech recognition is largely attributed to the massive amount of labeled data. However, most of the data samples in the real-world are not labeled, so it takes a lot of time and effort to label and use them for deep learning. In addition, in many countries, the collection and storage of medical data or data from personal robots are prohibited due to privacy concerns. This burden impedes the widespread use of deep learning in real-world applications such as medical AI, home appliances and robotics, and separates the best solutions from becoming the best real-world solutions.

From this point of view, active learning is a promising field of machine learning. The goal of active learning is to reduce the labeling cost by querying labels for only a small portion of the entire unlabeled dataset. The query selection is made according to which samples are most beneficial, i.e. both informative and representative (Settles, 2009). In most active learning algorithms, samples are selected from a large pool of data for querying (Settles, 2011; Gal et al., 2017; Sener & Savarese, 2017; Beluch et al., 2018). Once the queried data is labeled, it is accumulated in the training dataset and the entire model is trained again. This process is repeated, typically assuming that all the data can be stored and the model can access the data again at any time. In the aforementioned real-world problems, however, it is impossible to store a massive amount of data due to limited storage or privacy issues. In addition, it is highly inefficient to repeatedly train the entire model with large data and review all the unlabeled data for querying whenever a new label is acquired.

In contrast, the active learning paradigm in an online manner does not assume that data samples can be accessed repeatedly but instead, the input data is given as a continuous stream (Lughofer, 2017). This entails that uncertainty estimation of input samples and the decision of whether to query or not should be made online. The result of querying and training with one sample affects the uncertainty of subsequent input samples. Therefore, the model update should be done on-the-fly so that the new uncertainty distribution is estimated for the next query. In such scenarios, it is difficult to select a query sample that is representative as well as informative because all data samples cannot be reviewed at the same time. Moreover, since the labeling cost is expensive and the oracle is often unavailable, the number of queries may be limited (Hao et al., 2017; Zhang et al., 2018; Serrao & Spiliopoulou, 2018). Catastrophic forgetting is another issue to overcome when learning data online (Lee et al., 2017). This online active learning process, while more suitable for real-world problems, is therefore more challenging than offline active learning.
In this regard, we propose Message Passing Adaptive Resonance Theory (MPART) that addresses these problems in online active learning. MPART learns the distribution of the input data without forgetting based on ART (Grossberg, 1987) which keeps the existing knowledge stable when learning new data. By utilizing a novel topology learning method on the learned distribution, MPART forms a topological graph using both labeled and unlabeled data. In order to exploit the limited amount of labels acquired, our model uses a message passing method which compensates for the lack of class information by propagating the information within the topological graph. Based on the learned topology, MPART can also select beneficial samples to query. The entire learning process, depicted in Figure 1, is performed in an online manner. We evaluate our model in the online active learning environments, showing that MPART significantly outperforms competitive models.

The main contributions can be summarized as follows.

- We propose a novel method that learns the distribution of input data by continuously constructing a weighted topological graph based on the ART network.
- We propose a message passing method for the graph to estimate the class labels in a semi-supervised manner and to select beneficial samples for querying.
- We design an online active learning task where the frequency of query is limited and the total number of classes is unknown. We validate the proposed model with various query selection strategies and datasets. The results of the experiment show that our model significantly outperforms the competitive models in online active learning environments.

2. Related Work

Topology Learning. Topology learning aims to learn the structural properties and topological relations underlying the input data. Additional knowledge such as class label or uncertainty can be obtained from the structural relationship using the topological information. SOM (Kohonen, 1990), a type of artificial neural network, learns the topological properties of the input data using neighborhood function and competitive Hebbian learning (CHL). This method is useful for dimension reduction of high-dimensional input data, but has a disadvantage in that it requires a priori selection of a subspace. GNG (Fritzke, 1995) and its derived models such as SOINN (Shen & Hasegawa, 2006) and E-SOINN (Shen et al., 2007) are incremental network models which can learn the topological relations of input data using the CHL. There are also variants of ART networks (see Section 3 for details) which integrate new knowledge into its existing knowledge so that what has already been learned is not forgotten by the new learning. Fuzzy ART-GL (Isawa et al., 2007) uses Group Learning that creates connections between similar categories. TopoART (Tscherepanow, 2010) combines incremental clustering with topology learning, which enables stable online clustering of non-stationary input data.

Semi-supervised Learning. Semi-supervised learning uses a small amount of labeled data with a large amount of unlabeled data to improve the performance of the model. (Iscen et al., 2019) and (Douze et al., 2018) use the label propagation method to assign labels for unlabeled data using the nearby labeled data. GAM (Stretcu et al., 2019) uses an agreement model that calculates the probability of two nodes sharing the same label on the graph. EGNN
Adaptive Resonance Theory (ART), inspired by brain information processing mechanisms, is an unsupervised learning method for pattern recognition (Grossberg, 1987). The basic principle of ART is that object recognition is achieved by the interaction of ‘bottom-up’ sensory information and ‘top-down’ expectations. We refer to any neural network model based on ART as ART network. ART networks integrate new knowledge into the entire knowledge so that what has already been learned is not forgotten by new learning.

A basic ART network consists of two layers fully-connected to each other: an input layer and an output layer. As an input vector enters the input layer, it activates each node in the output layer according to the connection weight. Using the choice function with the activation values, the closest matching node is selected as a winner. Then, the fitness of the input to the winner node is calculated using the match function. If the fitness is greater than a vigilance parameter \( \rho \), the input vector is integrated into the winner, updating the associated weight. Otherwise, a new node is created with a weight set equal to the input vector. With these mechanisms, ART networks can categorize incoming data online without forgetting. There are various ART networks such as Fuzzy ART (Carpenter et al., 1991), Gaussian ART (Williamson, 1996) and Hypersphere ART (Anagnostopoulos & Georgioupolos, 2000). We use Fuzzy ART as our backbone model, which incorporates fuzzy logic to enhance generalizability.

Fuzzy ART will be described in Section 4.1 along with MPART’s topology learning method.

3. Background

Each related work shows a high potential to solve real-world problems in various domains, and motivates our idea of online active semi-supervised learning. In this section, we briefly introduce the ART networks and our motivations.

3.1. ART Networks

Online Active Learning. The goal of online active learning is not only to reduce annotation costs, but also to continuously expand existing knowledge by exploring new information. OASIS (Goldberg et al., 2011) is a Bayesian model that combines online active learning and semi-supervised learning methods. SOAL (Hao et al., 2017) and OA3 (Zhang et al., 2018) utilize second-order information for online binary classification under limited query budgets. While these methods are successful in binary classification, they cannot handle the increasing diversity of classes due to the assumption of a fixed number of classes. Other methods (Loy et al., 2012; Weigl et al., 2016) employ consensus of models to better estimate uncertainty. However, using multiple models incurs large computational costs, making them less scalable for complex tasks. The most closely related work to ours is (Shen et al., 2011), enhancing the SOINN to enable online active semi-supervised learning. The proposed A-SOINN, however, selects the queries among representative values of clusters, which might be difficult for the oracle to recognize. Moreover, since this model stores only a single label in a cluster, it cannot handle the mixture of classes within the cluster and is susceptible to noise. Our model solves these problems by querying the input itself and using a distribution for labels in each node.

3.2. Motivations

Due to the nature of ART, the ART networks can efficiently learn new data by updating only a part of the model. In addition, it does not need to predefine the structure and size of the model, which allows flexible self-organization of irregular and complex data distributions. These properties play an important role in online learning, so we chose an ART network as a backbone model of MPART.

Moreover, in online active learning scenarios where data cannot be stored, it is difficult to select a representative sample for querying because all data samples cannot be reviewed simultaneously. We address this problem by learning the topology of input data. There is also an ART network called TopoART that can learn the topology. However, TopoART, in which only the winner and the runner-up node are connected, cannot effectively represent the relationships between nodes. In TopoART, the strength of the edge continuously increases as the data is learned, so a proper normalization method is required to use the edge information for message passing on the topological graph. Therefore, we propose a novel method to learn the topology and utilize it for message passing to select queries and predict labels.

4. Methods

MPART learns the distribution and topology of input data to construct a topological graph online. Then, class prediction and uncertainty estimation of the input data are performed using a message passing method on the graph. The estimated uncertainty is used to select and query useful samples. The entire process of MPART is described in Algorithm 1. In
the following sections, we describe the topology learning, message passing, and active querying methods.

4.1. Topology Learning

Node Formation. We first extract representation vector \( r_i \in [0,1]^r \) from the input data \( x_i \) using the pre-trained BYOL (Grill et al., 2020) and Parametric UMAP (Sanburg et al., 2020) (see Section 5.3 for details). As in Fuzzy ART, the representation vector \( r_i \) is complement coded to \( I_i = [r_i, \overline{r_i}] \) in order to avoid proliferation of prototypes (Carpenter et al., 1991). With this, we can measure the similarity between the input and the category node \( j \) using the match function \( M_j \) and the choice function \( T_j \) as follows.

\[
M_j(I_i) = \frac{\|I_i \wedge w_j\|_1}{\|I_i\|_1}, \quad T_j(I_i) = \frac{\|I_i \wedge w_j\|_1}{\alpha + \|w_j\|_1} \tag{1}
\]

Here, \( \wedge \) is the element-wise minimum operator, \( \| \cdot \|_1 \) is the L1 norm, and \( \alpha > 0 \) is a choice parameter. A node \( j \) becomes activated by \( I_i \) if \( M_j(I_i) \) is greater than or equal to a vigilance parameter \( \rho \in (0, 1) \). When there are multiple activated nodes, or co-activated nodes, one with the highest \( T_j(I_i) \) is chosen as a winner denoted by \( J_i \). In case only one node is activated, we also refer to the node as the winner node \( J_i \). We update \( w_{J_i} \) with a learning rate \( \beta \in (0, 1] \), and increase \( d_{J_i} \) by 1 as follows.

\[
w_{J_i}^\text{new} = \beta(I_i \wedge w_{J_i}^\text{old}) + (1 - \beta)w_{J_i}^\text{old}
\]

\[
d_{J_i}^\text{new} = d_{J_i}^\text{old} + 1 \tag{2}
\]

If none is activated, a new node \( J_i \) is created with an initial weight \( w_{J_i} = I_i \) and a winning count \( d_{J_i} = 1 \).

Edge Formation. In addition to the node formation, edges between the co-activated nodes are developed. Unlike SOINN or TopoART in which only the winner and the runner-up node are connected, we connect the winner to every co-activated node to better represent more complex topology of multiple nodes. For an edge connecting nodes \( i \) and \( j \), its count \( c_{ij} \) is set as the number of times \( i \) and \( j \) have been co-activated. The edge weight \( e_{ij} \) is defined as a ratio of \( c_{ij} \) to the sum of winning counts of incident nodes \( i \) and \( j \) as shown in Equation 3.

\[
e_{ij} = \frac{c_{ij}}{d_i + d_j} \quad \tag{3}
\]

The weight \( e_{ij} \) well describes the correlation between \( i \) and \( j \). If the two nodes have never been co-activated, \( e_{ij} \) is 0. As they get co-activated more frequently, \( e_{ij} \) grows larger; it reaches the maximum value of 1 if \( i \) and \( j \) have been co-activated whenever one of them was the winner. Since \( e_{ij} \) is bounded to \([0, 1]\), it can be used to adjust the degree of message passing without normalization. The edge formation does not interrupt the node formation, so the properties of the underlying ART network are also preserved.

Algorithm 1 The MPART algorithm

\[
V \leftarrow \{\}, C \leftarrow \{\}
\]

for \( x_t \) in input data stream do

\[
r_t \leftarrow \text{DimensionReduction}(x_t)
I_t \leftarrow [r_t, \overline{r_t}]
A \leftarrow \{\}
\]

for \( j \) in \( 1, \ldots, |V| \) do

\[
M_j \leftarrow \frac{\|I_t \wedge w_j\|_1}{\|I_t\|_1}
T_j \leftarrow \frac{\|I_t \wedge w_j\|_1}{\alpha + \|w_j\|_1}
\]

if \( M_j \geq \rho \) then

\[
A \leftarrow A \cup \{j\}
\]

end if

end for

if \( A \) is empty then

\[
J_t \leftarrow |V| + 1, \quad V \leftarrow V \cup \{J_t\}
\]

end if

\[
c_{J_t} \leftarrow 0, \quad c_{vJ_t} \leftarrow 0 \quad \forall v \in V - \{J_t\}
\]

\[
q_{J_t}(y) \leftarrow 0 \quad \forall y \in C
\]

\[
w_{J_t} \leftarrow I_t, \quad d_{J_t} \leftarrow 1
\]

else

\[
J_t \leftarrow \arg \max_{j \in A(T_j)}
\]

\[
c_{J_t} + c_{vJ_t} + 1, \quad c_{vJ_t} \leftarrow c_{vJ_t} + 1 \quad \forall v \in A - \{J_t\}
\]

\[
w_{J_t} \leftarrow \beta(I_t \wedge w_{J_t}) + (1 - \beta)w_{J_t}
\]

\[
d_{J_t} \leftarrow d_{J_t} + 1
\]

end if

\[
q^{(L)}_{J_t}(d_{J_t}) \leftarrow \text{MessagePassing}(J_t, c, d, q)
\]

\[
p_c, \hat{y} \leftarrow \text{NodeClassification}(q^{(L)}_{J_t})
\]

\[
s_t \leftarrow \text{UncertaintyEstimation}(p_c, q^{(L)}_{J_t}, d_{J_t})
\]

if \( s_t \) satisfies query condition then

\[
y_t \leftarrow \text{QueryLabel}(x_t)
\]

\[
q_{J_t}(y_t) \leftarrow q_{J_t}(y_t) + 1
\]

\[
C \leftarrow C \cup \{y_t\}
\]

end if

end for

4.2. Message Passing

Nodes formed in MPART possess additional information vectors such as class probabilities or uncertainty measures for better inference and querying. To compensate for the lack of information, we use message passing to aggregate the information within the topological graph. Before describing the procedure, we first define a notion of neighbors. Given nodes \( i \) and \( j \), if the shortest path from \( i \) to \( j \) exists and its length is \( l \), then we say \( j \) is an \( l \)-hop neighbor of \( i \). Any node is a 0-hop neighbor to itself. A set of all \( l \)-hop neighbors of \( i \) is denoted as \( N_i^{(l)} \); a set of direct neighbors \( N_i^{(1)} \) will also be written as \( N_i \). We further denote an \( l \)-hop neighborhood \( N_i^{(0:1)} \) as a union of \( N_i^{(0)}, \ldots, N_i^{(l)} \).

To illustrate message passing, let \( X \) denote an information vector (or scalar) of interest such as winning count \( d \) or label density \( q \) (will be described later). A single-layer message passing on the target node \( i \) updates \( X_i \) by adding \( X_j \)
of every neighboring node \( j \in \mathcal{N}_i \), weighted by \( e_{ij} \) and discounted by a propagation rate \( \delta \in [0, 1] \). Multi-layer message passing is achieved by performing the single-layer message passing recursively, aggregating the information from broader neighboring area. The recursive procedure for \( L \)-layer message passing targeted at the winner \( J_t \) can be formalized as Equation 4, where \( X_i^{(l)} \) denotes the \( l \)-layer aggregated information and \( X_i^{(0)} \) is set as \( X_i \).

\[
X_i^{(l)} = X_i^{(l-1)} + \delta \sum_{j \in \mathcal{N}_i} e_{ij} X_j^{(l-1)}, \forall i \in \mathcal{N}_{J_t}^{(0:L-1)}
\]

We finally obtain the \( L \)-layer aggregated information \( X_{J_t}^{(L)} \) by applying this update for \( l \) from 1 to \( L \). Note that we only have to consider nodes in \( \mathcal{N}_{J_t}^{(0:L-1)} \) when performing the \( l \)-th layer update since the information outside this set will not be propagated to \( J_t \) within the remaining \( (L-l) \) updates. This reduces the computational cost compared to when reviewing all the nodes in the graph.

**Node Classification.** MPART infers the class of input \( x_t \) by aggregating the class information near the winner \( J_t \). The class information \( q_i \) of a node \( i \), called the label density, is a distribution over a set of known class labels \( C \) and indicates how probable the node belongs to each class. We use distributions rather than single values to cope with situations where one node represents samples of different classes.

Before receiving any label, each \( q \) is set to an empty distribution, i.e. an empty vector. As a new label is received and added to the label set \( C \), every \( q \) is expanded by one dimension (zero-valued) which will be responsible for that label. For new nodes, label densities are set to zero vectors of dimension \( |C| \). When a label \( y_t \) is provided with an input \( x_t \), the corresponding density value in the winner \( J_t \), i.e. \( q_{J_t}(y_t) \), increases by 1. This increment also applies if the label was obtained via active querying. Since labels are provided rarely, we perform \( L \)-layer message passing for \( q \) to alleviate the deficiency. The aggregated label density \( q_{J_t}^{(L)} \) contains class information of neighboring nodes, so it is representative and robust to noise. We obtain the class probability distribution \( p_t \) by normalizing it as Equation 5.

\[
p_t(y) = \frac{q_{J_t}^{(L)}(y)}{\sum_{y' \in C} q_{J_t}^{(L)}(y')}
\]

If \( q_{J_t}^{(L)} \) is a zero vector, we set \( p_t \) to a uniform distribution. The label of the input is inferred as one with the highest probability. This distribution is also used to estimate the label uncertainty of a node, as described in the next section.

### 4.3. Active Querying

We combine the uncertainty-based sampling with density-weighted method for query selection. The uncertainty is measured per node using the label density and weighted by distribution density of data to produce the selection score. This score can be used with various selection strategies, including those suggested in Section 5.2.

**Uncertainty Estimation.** We use two kinds of uncertainty measures as in (Kim et al., 2020). The first uncertainty \( u_e \) indicates the insufficiency of class information accumulated in each node. It is determined by the sum of label density values over all known classes. As more labels are input or more times of message passing are performed, \( u_e \) decreases as the sum of densities increases. For an input sample \( x_t \), \( u_e \) is defined as that of the winner \( J_t \) as in Equation 6, with a sensitivity constant \( k_e > 0 \) and \( \tanh \) for clamping.

\[
u_e = 1 - \tanh\left( k_e \sum_{y \in C} q_{J_t}^{(L)}(y) \right)
\]

The second uncertainty \( u_a \) measures the class impurity, i.e. how many classes are mixed within a node. We can use the normalized entropy of class probability, while setting \( u_a \) to 0 when no more than one label is observed.

\[
u_a = \begin{cases} 
\frac{-\sum_{y \in C} p_t(y) \log p_t(y)}{\log(|C|)}, & \text{if } |C| > 1 \\
0, & \text{otherwise}
\end{cases}
\]

The two quantities, \( u_e \) and \( u_a \), can be viewed as epistemic and aleatoric uncertainties respectively, in that \( u_e \) results from insufficient observation while \( u_a \) arises due to ambiguity in the input data. These uncertainties are complementary when used for query selection, since \( u_e \) plays an important role in the early stage of learning while \( u_a \) is more crucial after enough number of labels have been acquired. We combine them with a weight \( \tau \in [0, 1] \) to get a query selection score \( u_t \) as shown in Equation 8. Figure 2 illustrates that the distribution of queries selected using \( u_t \) tends to be sparse.
We set up an online active learning task to imitate real-world work, the query budget is set as the maximum number of inputs. Specifically, the number of queries is limited to within a period of \( W \) consecutive inputs. TheVisualization of result on the CIFAR-10 dataset for 1/500 query frequency and \( B = 3 \) using ‘Explorer’ strategy. The color of the dots represent the label of input data, and the intensity of the topological graph represents the density of nodes and edges. The distribution of the queried samples is spread evenly.

The basic goal of this task is to perform multi-class classification, where the number of classes is unknown in advance. The input samples are provided sequentially as a stream, i.e. one after another, and cannot be stored or reused. The whole training dataset is initially unlabeled and the model can inquire the oracle about the labels of a small number of inputs. Specifically, the number of queries is limited to a fixed budget \( B \) within a period of \( W \) consecutive inputs. We denote such constraint as a query frequency \( B/W \). The final classification performance is evaluated on the hold-out test dataset.

For experiments, we used four kinds of datasets with different distributions: Mouse retina transcriptomes (Macosko et al., 2015; Poličar et al., 2019), Fashion MNIST (Xiao et al., 2017), EMNIST Letters (Cohen et al., 2017), and CIFAR-10 (Krizhevsky et al., 2009). The datasets consist of 12, 10, 26, and 10 classes, respectively. The budget \( B \) was set to 1 or 2, and the period \( W \) was set to 100, 500, 1,000 or 2,000. For model training, we only used 10,000 randomly sampled data from the training split per trial. This is to push the situation to an extreme; with 1/500 or less query frequency, the model will not be able to receive labels for some classes unless it successfully selects and queries all the samples from different classes.

5.2. Query Selection Strategy

Random. A random query is selected from a sequence of inputs. This is a baseline and works as an ablated version of active selection strategies. It is equivalent to randomly providing labels for some inputs, so MPART learns in an online semi-supervised manner when using this strategy.
**Memory.** This can only be used at \(1/W\) query frequency, assuming the model has a memory that can store at most one sample. During each query period, one sample with the maximum density-weighted query selection score \(s_t\) is stored and queried at the end of the period. By using this strategy, the model can select the most beneficial sample.

**Explorer.** ‘Explorer’ assumes the most stringent situation where the learning agent cannot store any input sample. In this situation, the learning agent selects \(B\) samples online for each query period \(W\), and cannot change the sample once selected. Therefore, the chance of selecting an beneficial sample decreases as the exploration gets longer because of the fixed query selection period. In this strategy, the uncertainty distribution of the data explored so far is continuously estimated to solve the exploration-exploitation dilemma (Berger-Tal et al., 2014). The random variable of the uncertainty distribution \(S_n\) is assumed to follow the normal distribution \(S_n \sim \mathcal{N}(\mu_n, \sigma_n^2)\), where \(n\) is the number of accumulated samples since the previous query. The parameters \(\mu_n\) and \(\sigma_n^2\) are calculated as the sample mean and variance of \(s_t\) as in Equation 10.

\[
\begin{align*}
\mu_n &= (1 - n^{-1})\mu_{n-1} + n^{-1}s_t \\
\sigma_n^2 &= (1 - n^{-1})\sigma_{n-1}^2 + n^{-1}(\mu_n - s_t)^2
\end{align*}
\] (10)

An input sample is considered beneficial if its density-based query selection score is likely to be greater than that of most unseen samples. To be more concrete, we select the \(t\)-th sample if Equation 11 is satisfied. Here, \(\text{Binom}\) denotes a probability mass function of a binomial distribution, \(F_{S_n}(\cdot)\) is the cumulative density function of \(S_n\), \(t_p\) is the time from the start of the current period, and \(b\) is the number of remaining query budget in the current period. If \(\sigma_n = 0\), we set \(F_{S_n}(s_t)\) to 0.5.

\[
\sum_{m=0}^{b-1} \text{Binom}(m; W - t_p, 1 - F_{S_n}(s_t)) > 0.5
\] (11)

When a query is selected, \(b\) is reduced by 1, and both \(\mu_0\) and \(\sigma_0\) are reset to 0. If \(b = 0\), no query is performed until the end of the period, and \(b\) is reset to \(B\) when a new period starts. With the ‘Explorer’ strategy, the model can efficiently and effectively choose the useful samples by predicting the beneficialness of the unseen ones.

### 5.3. Feature Extraction & Dimension Reduction

We projected the input data into 4-dimensional embedding space using the pre-trained Parametric UMAP (Sainburg et al., 2020) for each dataset. The embedding space was mapped to \([0, 1]^4\) using min-max normalization on training data. Prior to projection, we extracted 2048-dimensional feature vectors from CIFAR-10 using the BYOL (Grill et al., 2020) pre-trained with the ImageNet dataset (Deng et al., 2009). Parametric UMAP is suitable for use in semi-supervised learning by capturing structure in unlabeled data to provide the embeddings for unseen data online which can be clustered properly. We trained Parametric UMAP using 30% of the training data for each dataset, and the rest was used for MPART training. Note that no labeled data was used until this step.

### 6. Results and Discussion

#### 6.1. Comparison with Competitive Models

The performance of the competitive models LPART (Kim et al., 2020) and A-SOINN (Shen et al., 2011) was compared to that of MPART with various query selection strategy, shown in Table 1 and Figure 4. The fully supervised settings including multi-layer perceptron (MLP) model were used as references, which was trained using all labeled data for each dataset. The MLP is consisted of 3 layers with 128 neurons per layer and we reported the highest test accuracy while training up to 200 epochs. The same ART-related parameter values were used in MPART and LPART, while A-SOINN parameters were adjusted to achieve its best performance.
Message Passing Adaptive Resonance Theory

Table 1. Comparison of classification accuracy (mean ± std) between our model (MPART) and the competitive models according to the query selection strategy. We also report classification accuracy of MPART depending on whether density-weighted query selection score (DS) is applied or not. The number of layers \( L = 3 \) for message passing of MPART was used. (unit : %)

<table>
<thead>
<tr>
<th>Query Selection Frequency</th>
<th>Model</th>
<th>Mouse retina transcripts</th>
<th>Fashion MNIST</th>
<th>EMNIST LETTERS</th>
<th>CIFAR-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully Supervised</td>
<td>MLP</td>
<td>94.0±1.7</td>
<td>71.9±1.6</td>
<td>50.9±2.0</td>
<td>75.1±1.7</td>
</tr>
<tr>
<td></td>
<td>LPART</td>
<td>97.2±0.1</td>
<td>73.2±0.2</td>
<td>60.2±0.4</td>
<td>74.2±0.3</td>
</tr>
<tr>
<td></td>
<td>A-SOINN</td>
<td>94.2±1.5</td>
<td>66.5±1.5</td>
<td>47.6±1.6</td>
<td>65.5±1.7</td>
</tr>
<tr>
<td></td>
<td>MPART</td>
<td>97.2±0.1</td>
<td>73.3±0.2</td>
<td>60.7±0.3</td>
<td>74.3±0.3</td>
</tr>
<tr>
<td>1 / 1000</td>
<td>LPART</td>
<td>82.5±7.5</td>
<td>41.1±6.4</td>
<td>14.7±2.6</td>
<td>23.8±5.3</td>
</tr>
<tr>
<td></td>
<td>A-SOINN</td>
<td>88.1±13.0</td>
<td>45.2±7.0</td>
<td>16.9±4.1</td>
<td>37.7±6.8</td>
</tr>
<tr>
<td></td>
<td>MPART-RANDOM</td>
<td>81.0±7.0</td>
<td>42.8±8.5</td>
<td>17.1±3.1</td>
<td>40.9±8.1</td>
</tr>
<tr>
<td></td>
<td>MPART-MEMORY</td>
<td>93.3±3.3</td>
<td>53.6±4.8</td>
<td>26.1±2.8</td>
<td>55.5±4.6</td>
</tr>
<tr>
<td></td>
<td>MPART-EXPLORER</td>
<td>91.8±4.0</td>
<td>53.5±5.4</td>
<td>25.4±2.4</td>
<td>55.4±3.7</td>
</tr>
<tr>
<td></td>
<td>MPART-MEMORY (w/o DS)</td>
<td>90.5±5.6</td>
<td>51.4±5.6</td>
<td>19.8±3.5</td>
<td>44.9±7.0</td>
</tr>
<tr>
<td></td>
<td>MPART-EXPLORER (w/o DS)</td>
<td>87.2±8.1</td>
<td>49.9±6.9</td>
<td>18.2±3.4</td>
<td>44.1±6.7</td>
</tr>
<tr>
<td>1 / 500</td>
<td>LPART</td>
<td>89.0±6.3</td>
<td>54.6±5.9</td>
<td>21.0±3.4</td>
<td>34.0±4.7</td>
</tr>
<tr>
<td></td>
<td>A-SOINN</td>
<td>91.0±5.4</td>
<td>50.3±6.0</td>
<td>25.6±4.8</td>
<td>44.2±7.0</td>
</tr>
<tr>
<td></td>
<td>MPART-RANDOM</td>
<td>88.3±4.4</td>
<td>54.3±7.6</td>
<td>23.4±3.8</td>
<td>50.9±5.2</td>
</tr>
<tr>
<td></td>
<td>MPART-MEMORY</td>
<td>94.5±1.1</td>
<td>62.0±4.4</td>
<td>35.6±2.5</td>
<td>59.7±3.4</td>
</tr>
<tr>
<td></td>
<td>MPART-EXPLORER</td>
<td>94.6±0.9</td>
<td>59.4±5.2</td>
<td>36.0±2.9</td>
<td>59.2±3.3</td>
</tr>
<tr>
<td></td>
<td>MPART-MEMORY (w/o DS)</td>
<td>91.9±3.4</td>
<td>51.7±5.8</td>
<td>28.7±3.3</td>
<td>52.2±5.4</td>
</tr>
<tr>
<td></td>
<td>MPART-EXPLORER (w/o DS)</td>
<td>92.1±3.4</td>
<td>53.9±4.8</td>
<td>29.2±3.8</td>
<td>51.6±6.6</td>
</tr>
<tr>
<td>1 / 100</td>
<td>LPART</td>
<td>94.6±1.5</td>
<td>64.4±2.0</td>
<td>38.0±2.3</td>
<td>62.0±2.9</td>
</tr>
<tr>
<td></td>
<td>A-SOINN</td>
<td>91.5±7.0</td>
<td>55.7±6.8</td>
<td>29.9±4.9</td>
<td>50.9±7.9</td>
</tr>
<tr>
<td></td>
<td>MPART-RANDOM</td>
<td>94.8±1.2</td>
<td>66.8±2.0</td>
<td>43.6±3.2</td>
<td>63.4±2.6</td>
</tr>
<tr>
<td></td>
<td>MPART-MEMORY</td>
<td>95.9±0.5</td>
<td>67.7±1.6</td>
<td>47.9±1.6</td>
<td>67.4±1.5</td>
</tr>
<tr>
<td></td>
<td>MPART-EXPLORER</td>
<td>95.9±0.9</td>
<td>67.7±1.3</td>
<td>47.5±1.5</td>
<td>66.8±1.4</td>
</tr>
<tr>
<td></td>
<td>MPART-MEMORY (w/o DS)</td>
<td>95.0±1.6</td>
<td>63.2±2.9</td>
<td>39.7±2.5</td>
<td>60.6±2.7</td>
</tr>
<tr>
<td></td>
<td>MPART-EXPLORER (w/o DS)</td>
<td>94.4±1.5</td>
<td>65.7±2.1</td>
<td>41.9±2.0</td>
<td>62.3±2.4</td>
</tr>
</tbody>
</table>

for fair comparison. A-SOINN needs to query the prototype of the most dense node, which is a weighted sum of encoded representations and does not correspond to any input samples. Therefore, in A-SOINN, the prototype of the node was not directly queried, but the most recent input sample that activates the node was queried. In all experimental settings except with MLP, the MPART showed the highest accuracy.

Comparison with LPART. MPART with ‘Random’ strategy showed higher performance than LPART, which performs online semi-supervised learning. This indicates that although both models are based on the same backbone model, Fuzzy ART, MPART’s message passing method is more effective than LPART’s label propagation method. MPART’s message passing differs from LPART’s label propagation in several aspects. The critical difference is that LPART does not sufficiently learn and use topology information of input data. LPART permanently updates the label density \( q \) when label propagation is performed, whereas MPART updates \( q^0 \) using only true labels and infers the class through message passing. The MPART’s message passing method with multi-layers prevents information propagation of one node from permanently affecting other nodes, and when a new label is acquired, the information can be immediately reflected in the inference of all nodes. In addition, MPART queries representative samples by estimating the distribution density of input data using the topology learned online, whereas in the case of LPART, it is difficult to select a representative sample because it cannot effectively estimate the distribution of the input data.

Comparison with A-SOINN. A-SOINN employs an explicit removal of edges and nodes to reduce the noise and maintain a reasonable model size. The removal procedure heavily depends on the deletion period \( \lambda \), which should ideally be determined according to the input statistics. Since such information is not known a priori, a wrong choice of \( \lambda \) severely degrades the model performance. Also, each node in A-SOINN holds only one label which is fixed since the first assignment either with actual label or by propagation from other nodes. Therefore, the order of labeling highly influences the inference results, especially at deci-
Table 2. Classification accuracy (mean ± std) according to MPART’s query frequency and number of layers L for message passing. The ‘Explorer’ strategy was used for query selection of MPART. (unit : %)

<table>
<thead>
<tr>
<th>Query Selection Frequency</th>
<th>Number of Layers</th>
<th>Mouse retina transcriptomes</th>
<th>Fashion MNIST</th>
<th>EMNIST Letters</th>
<th>CIFAR-10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L = 5</td>
<td>90.0±7.1</td>
<td>52.6±5.7</td>
<td>25.6±2.8</td>
<td>52.9±4.6</td>
</tr>
<tr>
<td></td>
<td>L = 3</td>
<td>92.3±4.0</td>
<td>56.5±3.9</td>
<td>26.5±2.3</td>
<td>53.9±4.4</td>
</tr>
<tr>
<td></td>
<td>L = 1</td>
<td>66.9±4.7</td>
<td>44.0±5.2</td>
<td>14.3±1.6</td>
<td>27.3±2.2</td>
</tr>
<tr>
<td>w/o MP</td>
<td>31.1±4.9</td>
<td>15.2±0.9</td>
<td>4.9±0.3</td>
<td>11.3±0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L = 5</td>
<td>94.1±2.2</td>
<td>60.1±4.6</td>
<td>33.4±3.2</td>
<td>58.1±4.0</td>
</tr>
<tr>
<td></td>
<td>L = 3</td>
<td>93.1±3.5</td>
<td>60.0±4.4</td>
<td>36.0±2.8</td>
<td>59.5±3.8</td>
</tr>
<tr>
<td></td>
<td>L = 1</td>
<td>86.8±2.6</td>
<td>59.2±3.9</td>
<td>21.4±1.2</td>
<td>38.3±2.0</td>
</tr>
<tr>
<td>w/o MP</td>
<td>31.0±2.3</td>
<td>19.6±1.6</td>
<td>6.0±0.2</td>
<td>12.5±0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L = 5</td>
<td>96.1±0.5</td>
<td>67.6±1.5</td>
<td>47.6±1.6</td>
<td>67.4±1.6</td>
</tr>
<tr>
<td></td>
<td>L = 3</td>
<td>95.8±0.7</td>
<td>67.3±1.5</td>
<td>48.0±1.7</td>
<td>67.2±1.3</td>
</tr>
<tr>
<td></td>
<td>L = 1</td>
<td>95.0±1.1</td>
<td>66.6±1.4</td>
<td>40.3±1.5</td>
<td>59.1±1.1</td>
</tr>
<tr>
<td>w/o MP</td>
<td>57.8±1.5</td>
<td>43.6±2.2</td>
<td>12.2±0.6</td>
<td>20.5±0.6</td>
<td></td>
</tr>
</tbody>
</table>

6.2. Ablation Study

Table 1 summarizes the results of performance evaluation of our model on four datasets according to query selection frequencies and strategies. In the w/o DS setting, the query selection score $u_t$ of Equation 8 was used instead of $s_t$ of Equation 9. In all experimental settings, the model that applies density-weighted query selection score showed significantly higher performance than the one without (w/o) DS. When comparing the performance with respect to the query selection strategy, the accuracy of the ‘Random’ strategy was generally low, and the ‘Memory’ strategy and the ‘Explorer’ strategy showed almost similar performance. This is because the ‘Explorer’ strategy properly estimates the uncertainty distribution of input data and efficiently selects the representative samples based on the remaining query opportunities. In addition, the lower the query frequency, the greater the difference in performance depending on the query selection strategy, indicating that the strategy can have a significant impact on performance in situations where labeled data is extremely scarce. Table 2 shows the performance evaluation of MPART according to the number of layers L for message passing. As the number of layers increased, the classification performance generally increased, but the results of using 3 and 5 layers were almost the same. On the other hand, when message passing is not applied (w/o MP), there is a significant performance penalty.

7. Conclusions

We propose Message Passing Adaptive Resonance Theory (MPART) for online active semi-supervised learning. MPART learns the distribution and the topology of the input data online, infers the class of unlabeled data, and selects the informative and representative samples through message passing between nodes on the topological graph. By evaluating our method on datasets including EMNIST Letters and CIFAR-10, we show that it outperforms the competitive models. In an online learning environment where data storage is limited and data labeling is expensive, waste of useful samples is inevitable. The proposed model fully utilizes the underlying structure of input data so that it can minimize the waste. This approach also reduces the need to create large datasets in advance in order to apply machine learning to various industries. We believe MPART offers new opportunities for machine learning techniques to be widely used in real-world applications.

Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (2019-0-01371-BabyMind/35%, 2015-0-00310-SW.StarLab/5%, 2017-0-01772-VTT/5%, 2018-0-0622-RMI/5%) grant funded by the Korean government.
Message Passing Adaptive Resonance Theory

References


Settles, B. Active learning literature survey. 2009.

Settles, B. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, pp. 1–18, 2011.


