Active Sampling for Open-Set Classification without Initial Annotation

Zhao-Yang Liu, Sheng-Jun Huang*
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing 211106, China
{zhaoyangliu, huangsj}@nuaa.edu.cn

Abstract

Open-set classification is a common problem in many real world tasks, where data is collected for known classes, and some novel classes occur at the test stage. In this paper, we focus on a more challenging case where the data examples collected for known classes are all unlabeled. Due to the high cost of label annotation, it is rather important to train a model with least labeled data for both accurate classification on known classes and effective detection of novel classes. Firstly, we propose an active learning method by incorporating structured sparsity with diversity to select representative examples for annotation. Then a latent low-rank representation is employed to simultaneously perform classification and novel class detection. Also, the method along with a fast optimization solution is extended to a multi-stage scenario, where classes occur and disappear in batches at each stage. Experimental results on multiple datasets validate the superiority of the proposed method with regard to different performance measures.

Introduction

In traditional supervised learning tasks, it is commonly assumed that the class labels are identical in the training phase and test phase. However, in many real applications, the label set expands as more novel classes occur during the test phase. For example, in face recognition problem, the model is trained with data collected for a prefixed set of people, and then is applied to real environment with many new persons (Zhang and Patel 2017); in automated genre identification of web pages, web page genres evolve, and the predefined genre palette may not cover all the genres existing in a large corpus during the test phase (Guru et al. 2016).

Such problems are formalized as a learning framework called open-set classification (Scheirer et al. 2013). In this framework, training examples are all collected from known classes, while test examples are from both the known classes and some other novel classes. The target of open-set classification is to train a model that on one hand can accurately classify examples of known classes, and on the other hand can successfully detect those from novel classes. Obviously, this is a much more challenging task than close-set classification.

There are some studies trying to solve this problem in different ways. For example, (Yu et al. 2017; Júnior et al. 2016) detects novel data by the distance difference between known data and novel class data, (Masud et al. 2010; Guru et al. 2016) uses clustering to filter novel class data. Most of these methods require a large set of annotated examples from known classes to train the classification model, which however, is usually unavailable in real cases. Actually, in real tasks, label annotation is usually expensive and time costly. Thus a more practical scenario is that we have a dataset collected from known classes, but all examples are unlabeled. For example, face images may automatically collected by detecting faces from a video of a known set of people, but not precisely annotated with person ID; and web pages of predefined genre set may be collected in batch by a spider, yet it is not annotated for each page. This situation leads to a more challenging task of learning with least labeled data.

Active learning is a primary approach for learning from limited labeled data with high annotation cost. It actively select the most important examples to query their labels, and try to train an effective model with least labeled data. In this paper, we propose an Active Sampling algorithm for Open-set Classification without Initial Annotation, and ASOCIA for short. Specifically, given no initial labeled data, it is not applicable to select the most important instances based on the model prediction. Instead, active sampling of representative examples with no need of label information is a better choice.

We extend the experimental design method in (Nie et al. 2013) to simultaneously consider the representativeness, robustness and diversity. After querying the labels of selected examples, a model is trained based on these representative examples, and is expected to simultaneously achieve accurate classification on known class and successful detection of novel classes in the test phase. To distinguish novel classes from known ones, and also to distinguish each other of the known classes, we employ a low-rank representation learning method (Liu and Yan 2011) to obtain discriminative features. Further, to speed up the method for potential large scale data from both known and unknown classes in the test

*This work was supported by National Key R&D Program of China (2018YFB1004300), NSFC (61503182, 61876081, 61732006).

phase, we introduce a fast solution based on incremental SVD (Berry, Dumais, and O'Brien 1995). We also extend the method to a more dynamic environment with multiple test stages, and at each stage some novel classes occur while some known classes disappear.

Experiments are performed on multiple datasets to validate the effectiveness of the proposed method on open-set classification. Results with regard to accuracy and F-1 measure show that our method achieves better performance on both the classification of known classes and detection of novel classes.

Our main contributions are summarized as follows:

- The ASOCIA framework is proposed for a novel and challenging setting of open-set classification without initial annotation.
- A new strategy incorporating structure sparsity and diversity is proposed for active selection of representative examples. Also, the discriminative low-rank representation with a fast solution is introduced for classifying known classes and detecting novel classes.
- Experimental study is performed to validate the effectiveness of the proposed method on both the active sampling and model performance.

The rest of the paper is organized as follows. In the next section, related studies from different aspects are summarized and discussed. Then we propose the ASOCIA framework with detailed introduction on active sampling and low-rank representation learning. After that, experimental results are presented, followed by the conclusion.

Related work

Open-Set Classfication

Semi-supervised methods make use of both the labeled data and unlabeled test data which contain novel class data to train the model. In LACU (Da, Yu, and Zhou 2014), the augment risk is introduced to adjust the separator closer to the labeled region. While in (Guru et al. 2016; Masud et al. 2010), clustering technique is used to construct the boundary for filtering examples of novel class.

Open-set classification has attracted many research interests. In (Scheirer et al. 2013), a open risk is introduced into the supervised classification model. After that, probability models (Scheirer, Jain, and Boult 2014; Zhang and Patel 2017) are proposed based on the open risk concept. The EVT approach (Scheirer et al. 2011) is adopted to split the score list of test data and divide them into novel or known data. The methods in (Júnior et al. 2016) and (Bouguelia, Belaid, and Belaid 2014) detect novel class by the distance of test data to labeled training data. In (Yu et al. 2017), authors adopt adversarial learning to generate pseudo negative data which are close to each known class.

Outlier detection techniques are also used for open-set classification by treating the examples from novel classes as outliers. The method in (Mu, Ting, and Zhou 2017) uses iForest (Liu, Ting, and Zhou 2008) to detect anomaly data which contain novel class data. The method in (Mu et al. 2017) uses matrix sketch technique to store main known class information and compute inner products between sketch matrix rows and test data to recognize novel class. Due to the use of matrix sketch technique, this method may need lots of labeled data.

In addition, other problems such as zero-shot learning (Xian et al. 2016), the attribute-incremental learning (Vapnik, Vashist, and Pavlovitch 2009), the class incremental learning (Kuzborskij, Orabona, and Caputo 2013) are also related to the open-set classification problem. While most methods mentioned above utilize many labeled data without considering limited annotation cost, and in real world, a new item often starts with data collection and annotation, no plenty of data available.

Active Learning

Active learning is a primary approach to deal with limited labeled data. It selects the most important examples to query their labels from the oracle. Different criteria have been proposed to estimate how important an example is for improving the classification model (Huang, Jin, and Zhou 2010; Huang and Zhou 2013).

In our problem setting, we need to select a batch of examples from the unlabeled dataset for once. And experimental design methods fit the data selection situation. In (Yu, Bi and Tresp 2006), authors propose the TED method for transductive experimental design, which tends to select data representative to those yet unexplored data. Based on the idea of data construction, (Yu et al. 2008; Shi and Shen 2016) transforms the TED as a convex problem and can get a global optimal solution. Another method ANLR (Hu et al. 2013) further improves the result by local reconstruction with only neighbors. (Nie et al. 2013) proposes the RRSS where the $L_{2,1}$ norm is adopted to constrain the data construction loss and the relationship matrix of training data.

Low-Rank Representation

In many studies (Narayanan and Mitter 2010; Donoho and Grimes 2003), a common assumption is that high-dimensional data lies in a low-dimensional subspace and it is reasonable in reality to structural data such as images, texts and digital audio files. So the data could be compressed from high dimension to low dimension. LRR (Liu, Lin, and Yu 2010) can be seen as a compressed sensing technique, which tries to minimize the rank of the relationship matrix. (Liu, Lin, and Yu 2010) solves the problem with a strong assumptions that the training data of each class are sufficient and the noises of data are at low level. (Liu and Yan 2011) ease the problem by introducing the effects of hidden data.

In (Liu and Yan 2011) the data matrix is decomposed to principal feature and salient feature which are further used to perform sub-space segmentation and classification. While here we are interested in the affiliation matrix in principal feature and need to calculate the affiliation matrix for each test instance, so the computation cost is large. In (Zhang, Lin, and Zhang 2013), complete solutions are provided, based on the idea, we propose a fast solution by introducing the incremental SVD decomposition.

The ASOCIA Framework

Problem Formulation

In traditional supervised learning, model is trained on a labeled set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with n examples, where $y_i \in Y$ is the class label of the i-th instance, and $Y = \{1, 2, \cdots, K\}$ is the close-set of K class labels. At the test phase, each instance belongs to one of the K classes in Y. The task is to learn a model $f(\mathbf{x}) : \mathcal{X} \rightarrow Y$ to classify test instances into one of the K classes.

In a classical open-set classification task, the training data $X = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consists of n examples from K known classes $Y = \{1, 2, \cdots, K\}$. While in the test phase, the test set consists of instances from the open-set classes $Y = \{1, 2, \cdots, K, K+1, \cdots, K+M\}$, where the M novel classes $K+1, \cdots, K+M$ are unseen during the training phase. The task is to learn a model $f(\mathbf{x}) : \mathcal{X} \rightarrow \{1, 2, \cdots, K, novel\}$, where the $novel$ represents all novel classes.

In this paper, we consider a special case of open-set classification with no initial annotation. At the beginning of the training phase, we are given a dataset $X = \{\mathbf{x}_i\}_{i=1}^n$ with n instances. Each instance belongs to one of the K classes in $Y = \{1, 2, \cdots, K\}$. However, the ground-truth annotation y_i is not available for all instances $\mathbf{x}_i \in X$. We need to actively sample a batch of important instances from X, query their class labels, and then train a model $f(\mathbf{x}) : \mathcal{X} \rightarrow \{1, 2, \cdots, K, novel\}$ to perform classification of known classes as well as detection of $novel$ classes.

Active Selection of Representative Examples

In the ASOCIA framework, we have no initially labeled data even for the known classes, and need to actively select a batch of most important examples from the unlabeled pool to annotate. Without a classification model to estimate the uncertainty or informativeness of an unlabeled instance, it is more practical to perform active sampling based on representativeness. Among the active learning methods, experimental design (Yu, Bi, and Tresp 2006; Nie et al. 2013) has shown effective performance for representative sampling.

In (Yu, Bi, and Tresp 2006), a transductive experimental design (TED) method is proposed to select the examples that can best represent the whole data using a linear representation. Formally, given a dataset $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ with n instances of d-dimensional feature vectors, TED tries to select a set B of m examples from X with the following objective function.

$$\min_{B,W} \sum_{i=1}^{n} \left(\|\mathbf{x}_i - B\mathbf{w}_i\|_2^2 + \gamma \|\mathbf{w}_i\|_2^2 \right) \tag{1}$$
$$s.t. \quad B \subset X, |B| = m, W = [\mathbf{w}_1, \cdots, \mathbf{w}_n] \in \mathbb{R}^{m \times n}.$$

Here $\mathbf{w}_i$ is the linear weight vector reflecting the relations between the selected examples and the instance $\mathbf{x}_i$.

The problem in Eq. 1 is NP-hard, and is solved with greed optimization (Yu, Bi, and Tresp 2006). Later, a more robust method is proposed in (Nie et al. 2013) by introducing structured sparsity. Specifically, this method dose not directly select a subset B from X. Instead, all examples are used to represent each instance $\mathbf{x}_i$ with the weight vector $\mathbf{w}_i$ of n dimensions. The objective functions is as follows.

$$\min_{W} \|(X - XW)^\top\|_{2,1} + \gamma \|W\|_{2,1}, \tag{2}$$

where the second term with $\ell_{2,1}$-norm on W aims to achieve structured sparsity. On one hand, the loss function will be less sensitive to outliers compared to that in Eq. 1; and on the other hand, the $\ell_{2,1}$-norm leads to a row-sparse solution of W. After the optimization of the above problem, the representative examples are selected according to the row-sum values of absolute W. A larger sum value of $|\mathbf{w}_i|$ implies that $\mathbf{x}_i$ contributes more to represent other examples, thus is more representative and should be selected to annotate.

When we sort the examples by the row-sum values of the absolute W, there could be some similar examples among the top ranked examples, which may contain redundant information. Annotating such redundant examples will contribute less to the model training, and thus leads to waste of annotation cost. To solve this problem, we propose to incorporate diversity into the objective function when optimizing the weight vectors. Specifically, we have:

$$\min_{W} \|\left(X - XW\right)^\top\|_{2,1} + \gamma \|W\|_{2,1} + \lambda \|WW^\top\|_F^2, \tag{3}$$

where γ and λ are two trade-off parameters. The third term minimizes the correlations among different rows of W, and thus is expected to enhance the diversity of selected examples. In summary, the objective of Eq. 3 is to select examples that can well represent the whole dataset via linear combination, have structured sparsity and high diversity. Such selected examples are expected to be most helpful to train an effective model.

Next we will discuss the optimization of Eq. 3. We adopt an approach similar in (Nie et al. 2013) to solve this convex problem. By setting the derivative of Eq. 3 to zero, we have:

$$X^\top XWU - X^\top XU + \gamma VW + 4\lambda WW^\top W = 0. \tag{4}$$

Both U and V are diagonal matrix, whose elements are computed according to (Nie et al. 2010):

$$U_{i,i} = \frac{1}{\|\mathbf{x}_i - X\mathbf{w}_i\|}; \quad V_{i,i} = \frac{1}{\|\mathbf{w}^i\|}.$$

where $\mathbf{w}_i$ and $\mathbf{w}^i$ represent i-th column and i-th row respectively. By further denoting $M = WW^\top$, after calculated U and V, fix $M = WW^T$, then we can update W and M alternately, repeat the procedure until convergence condition satisfied.

The algorithm for active sampling is summarized in Algorithm 1. Note in the experiments of multiple test phases, we also use this algorithm to select additional examples for annotation at each stage.

Low-Rank Representation Learning

In open-set classification, we need to simultaneously perform classification on known classes and detection of novel classes. A discriminative feature representation of the data examples is crucial for achieve good performance on this task. Recently, low-rank representation (LRR) based feature

Algorithm 1 Active Sampling

```
Input: The data $X \in \mathbb{R}^{d \times n}$; parameters $\gamma$ and $\lambda$
Output: The selected $m$ representative examples;
Initialize $W \in \mathbb{R}^{n \times n}$, and $U, V$ as diagonal matrices;
repeat
    Calculate diagonal elements: $U_{i,i} = \frac{1}{\|\mathbf{x}_i - X\mathbf{w}_i\|}$;
    Calculate diagonal elements: $V_{i,i} = \frac{1}{\|\mathbf{w}^i\|}$ ;
    Calculate $M = WW^\top$;
    for $i = 1$ to $n$ do
        Calculate each column of $W$ as
        $\mathbf{w}_i = U_{i,i}\big(U_{i,i}X^\top X + \gamma V + 4\lambda M\big)^{-1}X^\top \mathbf{x}_i$;
        Update $M$ by $M = WW^\top$;
    end for
until (satisfy the convergence condition)
Calculate the row-sum values of $|W|$;
Return the $m$ examples with largest row-sum values.
```

learning has achieved great success in various applications (Liu and Yan 2011; Zhou, Lin, and Zhang 2016). The basic assumption of LRR is that data from the same class should be distributed in the same low-dimensional subspace. While the dimension of the subspace corresponds to the rank of the representation matrix, LRR tries to find the lowest-rank representation that can represent the data examples with linear combinations of given dictionary.

Given the data matrix $X \in \mathbb{R}^{d \times n}$, the original LRR minimizes the following objective:

$$\min_{Z} \operatorname{Rank}(Z) \quad s.t. \quad X = AZ, \tag{5}$$

where A is the dictionary. To efficiently solve this problem, some alternative approaches with nuclear norm are proposed (Cai, Candes, and Shen 2010).

In our setting, very limited labeled data is available, and thus favors the methods that are more robust and require less examples. Latent LRR (Liu and Yan 2011) is a representative approach applicable to less data. It tries to exploit hidden data, and decomposes the data matrix X into two parts: a low-rank part XZ for principle features and a low-rank part LX for salient features, as formalized in the following equation.

$$\begin{aligned} &\min_{Z,L} \|Z\|_* + \|L\|_* \\ &s.t. \quad X = AZ + LX. \end{aligned} \tag{6}$$

The Algorithm

After solving the optimization problem in Eq. 6, the matrix Z captures the affiliation between data examples. Figure 1 visualizes the affiliation matrix of ExtendedYaleB dataset with 6 known classes. It can be observed that data from the same class have strong correlations. This affiliation matrix thus could be utilized for classification as well as novel class detection. Specifically, denote by $X_l \in \mathbb{R}^{d \times n}$ the set of representative examples selected via active sampling, and we have known their labels $\mathbf{y}_l$. The corresponding affiliation matrix is denoted by Z_l. In the test phase, a new instance $\mathbf{x}_o$ from open set is to be classified into one of the known classes or a novel class. If we add a test instance $\mathbf{x}_o$ into X_l, then we can have a new affiliation matrix (Zhang, Lin, and Zhang 2013):

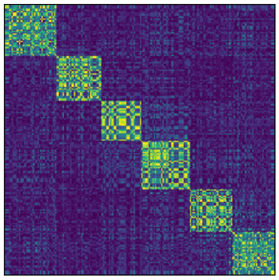


Figure 1: Visualization of the affiliation matrix Z with 6 known classes on ExtendedYaleB dataset

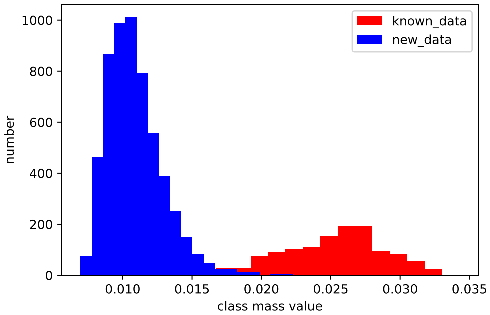


Figure 2: true score(class mass value) distribution of test data

$$\hat{Z}^* = \hat{V}\hat{S}\hat{V}^\top,$$

where $\hat{S}$ is a block diagonal matrix and $\hat{V}^\top$ comes from $SVD(X_{l+o}) = \hat{U}\hat{\Sigma}\hat{V}^\top$, $\hat{Z}^*$ has one more row and one more column than Z_l. We then delete the last column, and denote by $\mathbf{z}_o$ the absolute value of the last row of $\hat{Z}^*$. $\mathbf{z}_o$ describes how the test example $\mathbf{x}_o$ is affiliated with the labeled data X_l. Then we calculate the score for each of the K known classes as:

$$C_o^k = \frac{\sum_{i=1}^{m} I(\mathbf{y}_l(i) = k) \cdot \mathbf{z}_o(i)}{\sum_{i=1}^{m} I(y_i = k)}, \tag{7}$$

where $I(\cdot)$ is the identity function. C_o^k estimates how likely $\mathbf{x}_o$ belongs to class k.

Next, we need to decide whether $\mathbf{x}_o$ belongs to a novel class or not. Inspired by (Prewitt, Judith, and L. Mendelsohn 1966) which is prevalent in picture precessing, we use a iterative method to determine the threshold for distinguish known and novel classes. Given an instance $\mathbf{x}_o$ of the test set X_t, we firstly calculate the score $s_o = \arg\max_{k=1:K} C_o^k$, and then find the maximum score s_{max} and minimal score

s_{min} among all test data. After that the threshold is temporally set as $(s_{max} + s_{min}/2)$, and the test set is divided into two subsets according to the threshold. Then we update s_{max} and s_{min} by the mean scores of the two subsets, respectively. And the threshold updates as $(s_{max} + s_{min}/2)$. This process will be repeated until the threshold value converges to a stable value. Figure 2 shows an example of the score distribution over the test data on ExtendedYaleB. It demonstrate that the above method can find a accurate threshold for separating known and novel classes. At last, if the test instance $\mathbf{x}_o$ is identified as from known classes, then its class label will be further determined as $\arg\max_k C_o^k$.

Although some efficient solutions have been proposed for the problem in Eq. 6 (Zhang, Lin, and Zhang 2013; Liu and Yan 2011), it is still can not get the solution directly because we do not know the value of S_l; moreover, it is not scalable because we need to compute the affiliation matrix for each test example during the test phase. To overcome this problem, we adopt the incremental SVD (Berry, Dumais, and O'Brien 1995) to obtain a fast solution for calculating the affiliation matrix. Assume that we have the initial affiliation matrix of training data $Z_l = V_l S_l V_l^\top$ for X_l, and $\mathrm{SVD}(X_l) = U_l \Sigma_l V_l^\top$. We add test data matrix $X_t \in \mathbb{R}^{d \times m}$ to X_l to form the new data matrix $X_{new} = [X_l; X_t]$. We can make use of the SVD results over X_l instead of perform SVD from scratch on X_{new}. With the result from (Berry, Dumais, and O'Brien 1995), we have:

$$\mathrm{SVD}(X_{new}) = U_{new}\Sigma_{new}V_{new}^\top = U_l\Sigma_l[V_l^\top; V_t^\top],$$

where $V_t = X_t^\top U_l \Sigma_l^{-1}$. Thus for the case of adding a new test instance $\mathbf{x}_o$, we can have:

$$\hat{Z}^* = \hat{V} S_l \hat{V}^\top,$$

and $\hat{V}^\top = \left[V_l^\top; V_t^\top[:,o]\right]$. It can be further written as:

$$\hat{Z}^* = \hat{V}(V_l^\top Z_l V_l)\hat{V}^\top.$$

With this incremental solution to calculate the affiliation matrix, we can finally get efficient method to perform classification and novel class detection. The whole procedure of our method is summarized in Algorithm 2.

Algorithm 2 The ASOCIA Algorithm

```
Input: Unlabeled dataset $X \in \mathbb{R}^{d \times n}$; Initialize $A \in \mathbb{R}^{n \times n}$;
Perform Algorithm1 on $X$ to select the representative examples $X_l$;
Learn the low-rank representation $Z_l$ of $X_l$;
Clear $Z_l$ by setting non-block diagonal area to zero
Perform SVD over $X_l$: $\mathrm{SVD}(X_l) = U_l \Sigma_l V_l^\top$
Incremental SVD over test data $X_t$: $V_t = X_t^\top U_l \Sigma_l^{-1}$
Compute the affiliation matrix $Z_i^*$ for each example
Compute the score $s_i$ for each example
Compute the threshold: $\tau$
for each $\mathbf{x}_i \in X_t$ do
    if $s_i < \tau$ then
        $\mathbf{x}_i$ is detected as from novel class
    else
        $\mathbf{x}_i$ is classified as $y_i = \mathrm{argmax}_k C_i^k$
    end if
end for
```

Experiments

Datasets and algorithms

In our experiments, three datasets are used to examine the performance of the compared methods. ExtendedYaleB (Lee, Ho, and Kriegman 2005) has 28 classes, each of which has 576 images. Each image is resized to 48 × 42. Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017) consists of 70000 examples from 10 classes, where each example is a 28 × 28 grayscale image of clothes or shoes. We sample 500 instances for each class from those two datasets. Coil20 (S.A.Nene, Nayar, and H.Murase 1996) contains 20 classes, each of which has 72 examples.

The following methods are compared with our proposed method. OneClass SVM (Schölkopf et al. 2001) is a baseline to learn a model for the known classes, and then can be applied to detect novel classes. MOC-SVM incorporates the one-vs-rest strategy with OneClass SVM to perform open-set classification. SENC (Mu et al. 2017) uses matrix sketch techniques to store data information and distinguish new data and classify known class data by the sketch matrix. ASG (Yu et al. 2017) adopts adversarial learning to find a boundary around seen class data, and achieved state-of-the-art performance for open-set classification. ASOCIA-0 is a baseline of our method that simply ignore the diversity during active selection, i.e., it sets $\lambda = 0$ in Eq. 3.

Result on the open-set classification

For each dataset, we set 20% of the class labels as known, and others as novel classes. For ExtendedYaleB, half of the 5 known classes are randomly selected as the training set, from which the ASOCIA algorithm will actively select 150 examples from each class for annotation. At test stage, the other half of training examples from known classes together with all examples from unknown classes are used as the test set. Similar data partition is applied to Fashion-MNIST and Coil20, which have 2 and 4 known classes, respectively. The data partition is repeated for 10 times, and the average results are reported.

Because of the imbalanced size between known data and novel data, the measurements used here are accuracies of known class and novel class respectively: $Accuracy - known = \frac{M_{known}}{N_{known}}$, where M_{known} is the number of known data that are classified correctly to the true class, and N_{known} is the number of true known data in test data. Similarly, for novel class data, $Accuracy - novel = \frac{M_{novel}}{N_{novel}}$. Besides, we also adopt F1-measure to evaluate the overall performance on all test data (Yu et al. 2017; Mu et al. 2017). The results are showed in Table 1, 2 and 3, respectively. OneClass SVM can only distinguish between known and novel classes; and thus no *accuracy-known* and *F1-total* available for this method.

From the results in the tables, we can observe that in most cases the proposed ASOCIA method can achieve the best performance for both classification accuracy on known

Table 1: Open-set classification result of ExtendedYaleB: Best results are bold, OC could not classify known data with detail label, so some results are NA.

Method	OC	MOC	SENC	ASG	ASOCIA-0	ASOCIA
Accuracy-known	NA	0.515 ± 0.034	0.240 ± 0.029	0.876 ± 0.023	0.935 ± 0.018	**0.939** ± **0.016**
Accuracy-novel	0.404 ± 0.027	0.663 ± 0.064	0.528 ± 0.008	0.898 ± 0.036	**0.999** ± **0.002**	**0.999** ± **0.001**
F1-total	NA	0.340 ± 0.052	0.140 ± 0.032	0.838 ± 0.003	0.964 ± 0.008	**0.966** ± **0.009**

Table 2: Open-set classification result of Coil20: Best results are bold, OC could not classify known data with detail label, so some results are NA.

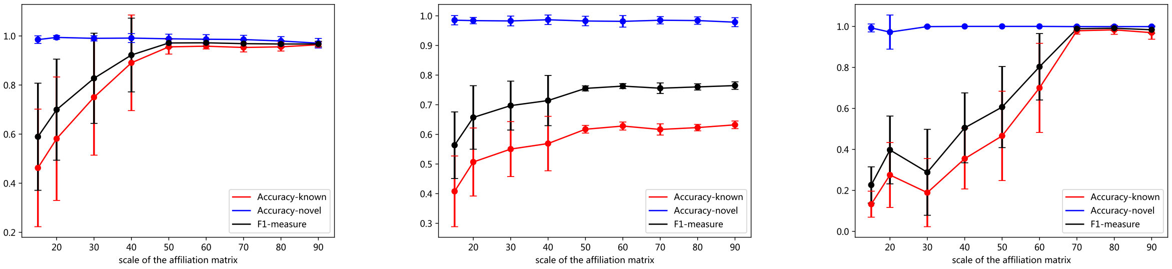
Method	OC	MOC	SENC	ASG	ASOCIA-0	ASOCIA
Accuracy-known	NA	0.751 ± 0.049	0.443 ± 0.013	0.896 ± 0.085	0.958 ± 0.026	**0.980** ± **0.018**
Accuracy-novel	0.591 ± 0.012	0.950 ± 0.014	0.377 ± 0.015	0.904 ± 0.034	0.970 ± 0.038	**0.996** ± **0.003**
F1-total	NA	0.770 ± 0.036	0.225 ± 0.007	0.788 ± 0.069	0.927 ± 0.059	**0.982** ± **0.009**

Table 3: Open-set classification result of Fashion-MNIST: Best results are bold, OC could not classify known data with detail label, so some results are NA.

Method	OC	MOC	SENC	ASG	ASOCIA-0	ASOCIA
Accuracy-known	NA	0.760 ± 0.057	0.345 ± 0.093	**0.824** ± **0.043**	0.818 ± 0.128	0.810 ± 0.031
Accuracy-novel	0.196 ± 0.010	0.574 ± 0.042	0.574 ± 0.026	0.178 ± 0.060	0.674 ± 0.162	**0.720** ± **0.160**
F1-total	NA	0.439 ± 0.030	0.225 ± 0.055	0.322 ± 0.017	0.539 ± 0.077	**0.571** ± **0.092**

Table 4: Performance on multiple test stages of ExtendedYaleB: Best results are bold, OC could not classify known data with detail label, so some results are NA.

Method	Stage one			Stage two			Stage three		
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
OC	NA	0.599 ± 0.046	NA	NA	0.661 ± 0.061	NA	NA	0.406 ± 0.046	NA
MOC	0.605 ± 0.047	0.938 ± 0.042	0.725 ± 0.048	0.640 ± 0.037	0.899 ± 0.079	0.736 ± 0.041	0.354 ± 0.058	0.692 ± 0.111	0.426 ± 0.071
SENC	0.398 ± 0.035	0.677 ± 0.016	0.461 ± 0.031	0.437 ± 0.045	0.557 ± 0.011	0.464 ± 0.038	0.267 ± 0.096	0.612 ± 0.032	0.317 ± 0.099
ActMiner	0.578 ± 0.037	0.958 ± 0.015	0.720 ± 0.0.026	0.383 ± 0.043	0.628 ± 0.108	0.476 ± 0.063	0.529 ± 0.037	0.831 ± 0.061	0.646 ± 0.041
ASG	0.956 ± 0.016	0.948 ± 0.019	0.952 ± 0.011	0.981 ± 0.019	0.926 ± 0.026	0.945 ± 0.013	**0.996** ± **0.006**	0.865 ± 0.030	0.926 ± 0.016
ASOCIA-0	0.937 ± 0.101	**0.969** ± **0.023**	0.949 ± 0.060	0.975 ± 0.025	**0.997** ± **0.010**	**0.985** ± **0.012**	**0.996** ± **0.003**	0.942 ± 0.019	**0.968** ± **0.010**
ASOCIA	**0.973** ± **0.024**	0.956 ± 0.015	**0.964** ± **0.015**	**0.999** ± **0.001**	0.968 ± 0.017	0.984 ± 0.009	0.986 ± 0.016	**0.944** ± **0.017**	0.964 ± 0.014



(a) ExtendedYaleB

(b) Fashion-MNIST

(c) Coil20

Figure 3: Performance changes with the scale of affiliation matrix

classes and detection accuracy on novel classes. Especially, the superiority of ASOCIA over ASOCIA-0 validates that considering the diversity is helpful for active sampling of representative examples. Among the compared methods, ASOCIA-0 tends to be effective in most cases, probably benefits from the low-rank representation learning. ASG also achieves descent performance on some datasets, but is less effective on the others. One possible reason is that ASG need to identify the boundary for generating examples. While on some small datasets with limited labeled data, e.g., Coil20, it may lose its edge due to the less accurate boundary. In addition, ASG generates data around each known class based on the distance. So if data from different classes are close to each other, ASG may confuse to recognize the labels of data. In contrast, our model based on low-rank representation learning can distinguish data based on the subspace where data lies in.

SENC gets relatively less effective results. In (Mu et al. 2017), the matrix sketch technique aims at large data matrix, and requires a large number of labeled data. While in our experiments, only a few examples are used to train the model. Besides, SENC computes inner products of the test data with every row of the sketch matrix. The datasets used here are image files. For example, for grayscale images, inner product of two undertint images is smaller than one undertint image with a dark image, so inner products of data to sketch matrix rows may not appropriate.

Result with different scales of affiliation matrix

In previous experiments, the average budget number is 30 for each class in the affiliation matrix. Here we examine how the performance changes with the increase of the affiliation matrix size. For each dataset, 6 classes are used and 3 of them are known classes and the other 3 classes are novel. The number of examples in each class is the same as previous experiments. We select data from 3 known classes and the budget number is from 15 to 90. The result is showed in Figure 3.

Figure 3 shows that the method achieves a high accuracy of novel class detection with various affiliation matrix sizes. Because in this experiment the score statistical distribution shows that the score range of novel data is smaller and lies in a narrow interval of unimodal distribution, so after the threshold separated the score list, novel data can always be classified correctly with high accuracy. But the score of known data is widely distributed. In the beginning, the accuracy of known data and the F1-measure are at low level and the standard deviation measurements are larger, After the training data grows to a certain scale, the performance stays at high level with a more steady state.

Update the model on multiple test stages

In this subsection, we examine the performance of the proposed method in a more challenging case with multiple test stages. In the beginning, no labeled data are available and we select 90 data from unlabeled dataset which contains 3 classes, then in every test stage, test dataset contains 6 classes, i.e., 3 novel classes are added. In each stage, we label the selected data and merge them with labeled data selected previously for the model training. We set three stages until 9 classes are included in the model. Due to the equal number of known data and novel data in test dataset, we use Precision, Recall and F1-measure as measurements where known data classified correctly are seen as true positive data, novel data classified correctly are seen as true negative data. The result is showed in Table 4. ASOCIA and ASOCIA-0 have the best classification results and new data detection effects. ActMiner (Masud et al. 2010) uses clustering to control the boundary which is a hypersphere determined by the furthest point to the corresponding cluster center, so it is more sensitive to outliers and may not filter new class data accurately.

Conclusion

We study a challenging case of open-set classification, where the training data is collected from known class but fully unlabeled, while the test data is from a open-set of both known and novel classes. We propose an active learning approach to perform both classification of known classes and detection of novel classes. On one hand, representative sampling is incorporated with diversity to actively select the most important examples for label annotation; on the other hand, low-rank representation along with a online solution is learned to achieve discriminative features. The effectiveness of the proposed ASOCIA approach is validated on multiple datasets with regard to different performance measures. In the future, we plan to incorporate other classification techniques with representation learning to further improve the performance.

References

Berry, M. W.; Dumais, S. T.; and O'Brien, G. W. 1995. Using linear algebra for intelligent information retrieval. *Siam Review* 37(4):573–595.

Bouguelia, M. R.; Belaid, Y.; and Belaid, A. 2014. Efficient active novel class detection for data stream classification. In *International Conference on Pattern Recognition*, 2826–2831.

Cai, J.; Candes, E. J.; and Shen, Z. 2010. A singular value thresholding algorithm for matrix completion. *Siam Journal on Optimization* 20(4):1956–1982.

Da, Q.; Yu, Y.; and Zhou, Z. 2014. Learning with augmented class by exploiting unlabeled data. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 1760–1766.

Donoho, D., and Grimes, C. 2003. Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Technical report.

Guru, D. S.; Suhil, M.; Gowda, H. S.; and Raju, L. N. 2016. Detection of a new class in a huge corpus of text documents through semi-supervised learning. In *International Conference on Advances in Computing, Communications and Informatics*.

Hu, Y.; Zhang, D.; Jin, Z.; Cai, D.; and He, X. 2013. Active learning via neighborhood reconstruction. In *Proceedings of* *the 23rd International Joint Conference on Artificial Intelligence*.

Huang, S.-J., and Zhou, Z.-H. 2013. Active query driven by uncertainty and diversity for incremental multi-label learning. In *Proceedings of the 13th IEEE International Conference on Data Mining*, 1079–1084.

Huang, S.-J.; Jin, R.; and Zhou, Z.-H. 2010. Active learning by querying informative and representative examples. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 892–900.

Júnior, P. R. M.; de Souza, R. M.; de O. Werneck, R.; Stein, B. V.; Pazinato, D. V.; de Almeida, W. R.; Penatti, O. A. B.; da S. Torres, R.; and Rocha, A. 2016. Nearest neighbors distance ratio open-set classifier. *Machine Learning* 106(3):1–28.

Kuzborskij, I.; Orabona, F.; and Caputo, B. 2013. From n to n+1: Multiclass transfer incremental learning. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3358–3365.

Lee, K. C.; Ho, J.; and Kriegman, D. J. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 27(5):684–698.

Liu, G., and Yan, S. 2011. Latent low-rank representation for subspace segmentation and feature extraction. In *Proceedings of the 13th International Conference on Computer Vision*.

Liu, G.; Lin, Z.; and Yu, Y. 2010. Robust subspace segmentation by low-rank representation. In *Proceedings of the 31th International Conference on Machine Learning*.

Liu, F. T.; Ting, K. M.; and Zhou, Z. 2008. Isolation forest. In *International Conference on Data Mining*.

Masud, M. M.; Gao, J.; Khan, L.; Han, J.; and Thuraisingham, B. 2010. Classification and novel class detection in data streams with active mining. In *Proceedings of the 14th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining-Volume Part II*, 311–324.

Mu, X.; Zhu, F.; Du, J.; Lim, E.-P.; and Zhi-Hua. 2017. Streaming classification with emerging new class by class matrix sketching. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*.

Mu, X.; Ting, K. M.; and Zhou, Z.-H. 2017. Classification under streaming emerging new classes: A solution using completely-random trees. *IEEE Transactions on Knowledge and Data Engineering* 29.

Narayanan, H., and Mitter, S. 2010. Sample complexity of testing the manifold hypothesis. In *Advances in Neural Information Processing Systems 23*, 1786–1794.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint $l_{2,1}$ -norms minimization. In *Advances in Neural Information Processing Systems 23*.

Nie, F.; Wang, H.; Huang, H.; and Ding, C. 2013. Early active learning via robust representation and structured sparsity. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*.

Prewitt, M. S.; Judith; and L. Mendelsohn, M. 1966. The analysis of cell images. 128:1035–53.

S.A.Nene; Nayar, S.; and H.Murase. 1996. Columbia object image library (coil-20), technical report cucs-005-96. Technical report, Columbia University.

Scheirer, W.; Rocha, A.; Michaels, R.; and Boult, T. E. 2011. Meta-recognition: The theory and practice of recognition score analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 33.

Scheirer, W. J.; Rocha, A.; Sapkota, A.; and Boult, T. E. 2013. Towards open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 35.

Scheirer, W. J.; Jain, L. P.; and Boult, T. E. 2014. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 36.

Schölkopf, B.; Platt, J. C.; Shawe-Taylor, J.; Smola, A. J.; and Williamson, R. C. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation* 13(7):1443–1471.

Shi, L., and Shen, Y. 2016. Diversifying convex transductive experimental design for active learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

Vapnik, V.; Vashist, A.; and Pavlovitch, N. 2009. Learning using hidden information (learning with teacher). In *Proceedings of International Joint Conference on Neural Networks*, 3188–3195.

Xian, Y.; Akata, Z.; Sharma, G.; Nguyen, Q.; Hein, M.; and Schiele, B. 2016. Latent embeddings for zero-shot classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 69–77.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

Yu, K.; Zhu, S.; Xu, W.; and Gong, Y. 2008. Non-greedy active learning for text categorization using convex transductive experimental design. In *Proceedings of the 31st Annual International ACM SIGIR Conference*.

Yu, Y.; Qu, W.; Li, N.; and Guo, Z. 2017. Open category classification by adversarial sample generation. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3357–3363.

Yu, K.; Bi, J.; and Tresp, V. 2006. Active learning via transductive experimental design. In *Proceedings of the 23rd International Conference on Machine Learning*.

Zhang, H., and Patel, V. M. 2017. Sparse representation-based open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)* 39.

Zhang, H.; Lin, Z.; and Zhang, C. 2013. A counterexample for the validity of using nuclear norm as a convex surrogate of rank. *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013* 226–241.

Zhou, P.; Lin, Z.; and Zhang, C. 2016. Integrated low-rank-based discriminative feature learning for recognition. *IEEE Transactions on Neural Networks Learning Systems* 27(5):1080–1093.