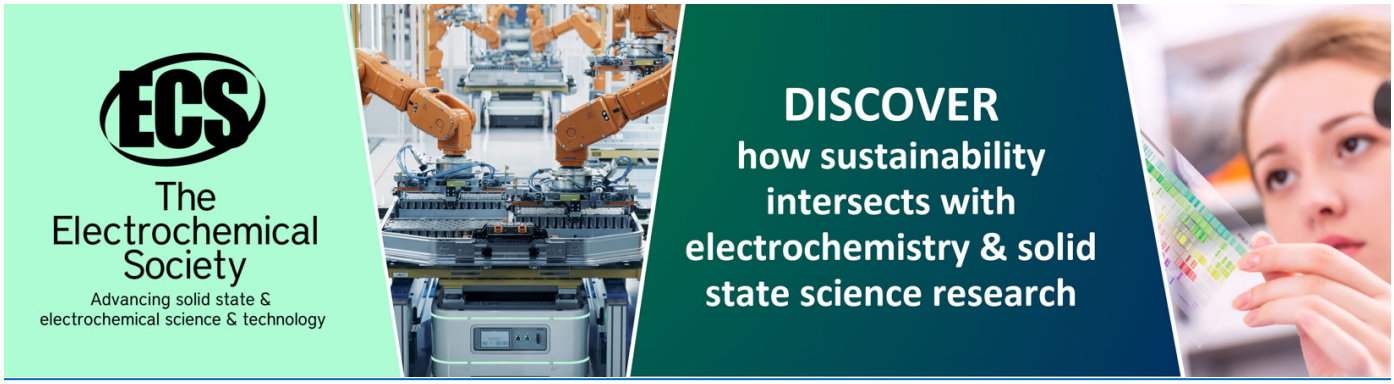
Multi-category Classification Problem Oriented Subsampling-Based Active Learning Method



Multi-category Classification Problem Oriented Subsampling-Based Active Learning Method

Wei Shi, Yanghe Feng, Guangquan Cheng, Shixuan Liu and Zhong Liu

College of Systems Engineering National University of Defense Technology
Changsha, China
Email: fengyanghe@yeah.net

Abstract. Traditional active learning methods have achieved gratifying results in the classification tasks of less categories such as binary classification, the application research of active learning in the field of big data problems still faces enormous challenges. Since many active learning query strategies need to perform matrix inversion, the amount of calculation increases exponentially with the increase of the scale of the problem, it is difficult to apply these active learning methods in large scale multi-category data classification task. In order to solve this problem, this paper designed a subsampling-based active learning model, and integrate unsupervised clustering algorithm with traditional active learning method, then conducted experiments on Binary Alphadigits and OMNIGLOT data sets. This paper compares the performance of five traditional active learning algorithms using this subsampling method, namely random sampling, uncertainty sampling, query-by-committee, density weighting and learning-based active learning. Through comparative experiments, the feasibility of active learning based on subsampling for solving the multi-category classification problem is verified, and it is found that the subsampling-based method can break the limitations of traditional active learning methods that cannot deal with large-scale data classification.
Keywords. Subsampling-based active learning method; multi-category classification; traditional active learning algorithm.

1. Introduction

Supervised learning is the process of utilizing labelled samples to continuously adjust model parameters to achieve the required performance. With the advancement of science and technology, it is no longer difficult to obtain large-scale sample data, which grants us the opportunity to improve the prediction performance of the model. However, labelling data is often a costly and time-consuming task even for pertinent experts. Ref. [1] employed many dermatologists to annotate the 129,450 skin-cancer clinical images used in the article. Ref. [2] directly points out that, assuming that 1,000 images are need to be labelled, it takes 3 to 4 days and 2,000-3,000 dollars for X-ray images, and 10 to 20 days and 5,000-7,000 dollars for CT images.

Active learning method can effectively solve aforementioned problems. Its algorithm selects the most typical examples and asks human experts for the label information. In this sense, only through a few labelled training examples, the annotation of the data set can be achieved. However, the current mainstream active learning methods all require strong computational power to calculate the effective information and information density of the examples. Being shackle by its characteristics, active learning algorithm cannot solve practical problems with excessively large data volume. Therefore, an

effective method to build competitive classifiers through only a small number of subsets of the original data set is devastatingly needed.

This paper proposes an active learning method based on subsampling that combines unsupervised clustering method with traditional active learning method. To start with, the algorithm adopts unsupervised clustering method to roughly cluster the data set and extract the sub-datasets of partial categories, which is called sub-sampling, so as to reduce the dimension of the data set. Then the active learning method is applied to subsampling to realize the classification of unlabelled sets. After several iterations, the overall classification annotation of the original data set is completed. The proposed algorithm is tested on Binary Alphadigits and OMNIGLOT data sets, and the experimental results show that the proposed method can overcome the limitations of traditional active learning methods that cannot handle large-scale data tagging.

2. Related Work

Literature on active learning methods shows that some mainstream active learning algorithms, e.g. uncertainty sampling, focus on selecting one instance with the maximum uncertainty, i.e., containing the maximum information, for labelling. The strategy in Refs. [3-4] tends to select those samples, whose category is least determined by the current classifier, for annotation. The selection algorithm chooses the most valuable samples from the unannotated samples, passes them to the experts, and after annotation, adds them to the training set, so as to obtain the highest classification performance with the fewest samples as possible. This approach has achieved good results in many applications, but the algorithm is not stable enough. The principle of the query by committee algorithm, being a method of filtering information to query from random input stream, mentioned in Refs. [5-6] is to train a classifier committee and select the instance with which the committee members most disagree to analyse. However, the performance of the bagging query method adopted is sensitive to the base classifier, and the estimation error in this method also entails biases and variances, which requires a large amount of calculation.

The active learning strategies above are within the scope of myopic active learning, which gets its name because of their mere exploitation over annotated instances and their ignorance over the distribution information of the unannotated instances. Affected by the sample distribution of data set, outliers and invalid values often pop into the query results.

In view of the defects of the above algorithms, another class of algorithms tries to obtain information from a large number of unmarked instances and establish a classifier with good generalization performance for instances not visible in the problem domain. For example, in Ref. [7], a method is proposed to minimize the generalization error directly by reducing the expected error of unlabelled data relative to the estimated probability of a posterior label. A similar approach is to indirectly minimize generalization errors by reducing model variance, as in Refs. [8-9]. But these two methods are often computationally expensive.

Different from the method of querying single instances, active learning method in batch mode will select a batch of samples, update the selection criteria with the received authentic label information, and conduct the next round of training and selection. Depending on the varying size of the batch, a variety of one-shot active learning methods are proposed [10-11]. These methods advocate that smaller batch size brings faster selection criteria update, thus promoting the efficient.

There is also a class of active learning methods that do not require real labels to label samples [12-13], which try to minimize the expected variance of statistical models and omit label information when calculating such variances. However, as this method does not really make full use of the sample label information and only focuses on representative instances, the annotated instances will not be utilized by the algorithm even if they have valid information. Ref. [14] has improved the performance of this kind of one-shot active learning method to some extent by introducing multiple pseudo-annotators, but the disadvantages aforementioned still exist.

There is also a class of active learning methods that use heuristic approach to extract valid information from unlabelled data. [15] used the prior probability p(x) of unlabelled instances as the

weight of uncertainty. Methods in Refs. [16-17] explicitly combine intelligent algorithms with active learning to take advantage of both labelled and unlabelled instances. A similar framework is adopted in Ref. [18], which uses cosines to measure information density.

3. Proposed Approach

In view of the difficulty of applying traditional active learning in multi-category datasets, this paper presents a subsampling-based active learning method, and discusses the effectiveness of several classical active learning algorithms based on subsampling in multi-category datasets through experiments.

3.1. Unsupervised Clustering

In order to solve the problem that active learning is not competent for large-scale data annotation, we introduce an unsupervised clustering link before active learning. This method uses the k-means algorithm, which is a clustering analysis algorithm with iterative solution. It randomly selects K objects as the initial clustering centre, then calculates the distance between each object and each clustering centre, and assigns each object to the nearest clustering centre. The cluster centre and the objects assigned to it represent a cluster. After each iteration, the algorithm recalculates the centre of each cluster based on the existing objects in the cluster. This process is repeated until a certain convergence condition is met.

There are four steps in the k-means algorithm, which are as follows:

Step 1: Selection of K-value

The number of centres is given by the user, denoted as K. The value of K is generally selected according to the actual needs. Each sample only belongs to one cluster, and the initial cluster is empty.

Step 2: Distance calculation

The nearest-neighbour metric is used to classify object points into the nearest cluster. Euclidean distance is used in Euclidean space, cosine similarity function is used in processing document objects, and Manhattan distance is also used in some occasions. The selection of metrics needs to be done according to different situations. Let the ith instance be x_i, the center of the jth cluster class be $Center_j$, and the distance from the data point to the center of cluster be dist $(x_i, Center_j)$.

Step 3: Calculate the new cluster center

When the second step is finished, K new clusters are obtained, and each sample is categorized into one of the K clusters. For the K clusters generated after classification, the point with the minimum mean distance to other points in the cluster is chosen as the centre. Assuming that the jth class cluster contains data points $x_{j1}, x_{j2}, \cdots x_{jm}$, then the coordinates of the new center of cluster are

$$Center_j = \frac{1}{m}\sum_{r=1}^{m} x_{jr} \tag{1}$$

Step 4: Determine whether k-means shall terminates

The algorithm stops when the difference between two iterations ΔJ is less than the iteration termination threshold δ, or we can set a maximum number of iterations iter_{max}. Otherwise, loop through steps 2 to 4.

Albeit the effectiveness of K-means, it is easily affected by the initialization of cluster centre. In general, when the distribution of sample data is unknown, the setting of the initial centre is random, making algorithm is prone to local optimization. Usually, we do a lot of repeated experiments to find the best initial Settings.

3.2. Sub-sampling

After unsupervised clustering of the original data set, $\text{L} = \{L_1, L_2, \cdots, L_K\}$ is obtained. Set L is the set of all subclasses, and L_h represents the hth cluster, $\text{h} \in [1, \text{K}]$ and $\text{h} \in \text{Z}$.

According to the actual needs, p elements are randomly selected from L to form a new set. The subset is set as $L_{sub,p} = \{L_{s1}, L_{s2}, \cdots, L_{sp}\}$, where, p represents the number of subclasses selected, $\text{p} \in$

$[1, \mathrm{K}]$ and $\mathrm{p} \in \mathrm{Z}$. $L_{sub,p}$ is used as a new unlabelled data set for active learning algorithm. Since the performance of active learning is greatly affected by the data set, the value of p should be selected according to the actual situation, and the appropriate value of p will bring about the improvement of performance.

3.3. Active Learning Method

Although all active learning methods can be theoretically applied, we conducted experiments on the current mainstream active learning algorithms following:

Random sampling (RS), where a specified number of samples are uniformly randomly selected from each subset as training data, is often used as the baseline for all active learning methods.

Uncertainty sampling (Unc), the uncertainty sampling based on entropy is adopted as the standard of query samples:

$$x_H^* = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(\widehat{y_i}|x) \log P_\theta(\widehat{y_i}|x) \tag{2}$$

Here, x is the sample in the subset, and x_H^* is the query sample selected by the model θ of information entropy as the uncertainty sampling criterion.Because of the subsampling method, the length of the label y_i here is p in the sampling subset $L_{sub,p}$, and the union of y_i on all subsets covers all possible labels on the entire dataset.

Query by committee (QBC), the QBC method is designed to find the minimum version space consistent with the labelled training data on the current subset. Here, we use the method of bagging query. Bagging in the context means to resample the input sample and obtain a fixed distribution. Thereby, the final hypothesis is obtained by averaging the output. In this method, the prediction error is composed of bias and variance, where bias is the estimation error necessary for the size of input data, and variance is the statistical difference existing in specific data. Bagging isolates these two factors and minimizes the variance of the error. It is assumed that a total of T queries are performed on each subset. Therefore, the samples being queried at time t are:

$$x_{QBC}^* = \underset{x}{\operatorname{argmax}} \left| |\{t \le T | h_t(x) = 1\}| - |\{t \le T | h_t(x) = 0\}| \right| \tag{3}$$

where, $h_t(x)$ represents the version space at the time of the tth query on the subset.After T queries, the final hypothesis is:

$$h_{fin}(x) = \underset{y}{\operatorname{argmax}} \ |\{t \le T | h_t(x) = y\} \tag{4}$$

Density weighting. We adopted subsampling method to improve the graph-based density weighting method. The graph-based density weighting method describes the relationship between samples by constructing a k-nearest neighbour graph structure and using adjacency matrix. The structure of the graph constructed on each subset is symmetric, and the weight between the two samples is expressed as:

$$W_{ij} = P_{ij} \exp\left(\frac{-\mathrm{dist}(x_i, Center_j)}{2\sigma^2}\right) \tag{5}$$

$\mathrm{Dist}(x_i, Center_j)$ represents the Manhattan distance between the two samples.In order to distinguish data points with multiple fields with small weights, we normalized these weights by the number of edges, and the calculation formula of the query samples was as follows:

$$x_{GD}^* = \underset{x}{\operatorname{argmax}}\, Gra(x) = \underset{x}{\operatorname{argmax}} \frac{\sum_i W_{ij}}{\sum_i P_{ij}} \tag{6}$$

Learning-based active learning (LAL). Similarly, this paper uses the method of subsampling to improve the learning-based active learning (LAL) proposed by Konyushkova and Sznitman. For each subset, by considering the query selection process as a regression problem, the model trains a regressor to predict the expected error reduction of candidate samples on each subset.

On each subset, LAL selects samples through the following formula for tag query:

$$x_{LAL}^{*} = \underset{x \epsilon u_t}{\operatorname{argmax}}\, g(\Phi_t, \Psi_x) \tag{7}$$

where, t represents the number of queries in the current iteration; u_t represents the unlabelled data set on the current subset; $g(\cdot)$ is a regression function that can predict the potential error reduction of annotating a particular sample in a given classifier state; Φ_t marks the parametric classifier in current subset, Ψ_x is expressed as the characteristic parameters of unannotated sample in the current subset.

3.4. Standard Framework

The proposed method integrates several sub-algorithms and improves active learning algorithm in solving classification problem. The methods of unsupervised clustering, subsampling and active learning are given above. Our goal is to establish a comprehensive framework to integrate the advantages of the three. The main idea is as follows: as active learning algorithm with excellent performance has a huge amount of computation, when confronted with the practical problems of large-scale data sets, it is difficult to directly use the existing active learning algorithm to train the classifier. Therefore, to solve this problem, we use k-means algorithm at first to achieve unsupervised clustering for the original data set. Although the classification effect is not as good as the active learning algorithm, the rough classification of the data set can be preliminarily realized. In the second step, we need to randomly select some subclasses as subsamples of the original data set. This step is actually the compression of the original data set, which reduces the size of experimental data. Afterwards, in the third step, the sub-sampling data set is taken as the unlabelled data set and plugged into the active learning algorithm to complete the labelling. Repeat the second and third steps until the complete annotation of the clustering is generated. The whole process needs to select the appropriate number of iterations according to the size of the data set. After repeating experiments for several times, the optimal results can be screened out.

The flow chart of the method is shown in figure 1.

4. Experiment

We apply five classical active learning methods based on subsampling to the problem of multi-category data labeling and conduct case studies on the performance of those methods on multi-category data.

4.1. Datasets

4.1.1. Binary Alphadigits. The Binary Alphadigits data set is a data set comprised of handwritten character images. There are 36 types of handwritten characters in the data set, consisting of 26 uppercase letters " A " to " Z " and 10 numeric characters " 0 " to " 9 ". Each type of character in the data set is composed of 39 binary images. Through zeroing, we expanded each sample image to 20 * 20 pixels, so each image can be represented by a 400-dimensional vector. The dataset image is shown in figure 2.

4.1.2. Omniglot. the omniglot dataset is also a handwritten character image dataset. The dataset consists of 32, 460 (1623 classes) different handwritten characters composed of 50 different letters. Each type of character was drawn online by different 20 people via amazon's Mechanical Turk. The size of each image is 105 * 105 pixels. In our experiment, we compressed each sample image to 28 * 28 pixels. Therefore, each image can be represented by a 784 - dimensional vector. The OMNIGLOT dataset is harder to classify than the Binary Alphadigits dataset because of the large number of categories in the dataset and the small number of samples per category. Therefore, OMNIGLOT dataset is a standard dataset for small-sample learning. Some of the images in the dataset are shown in figure 3.

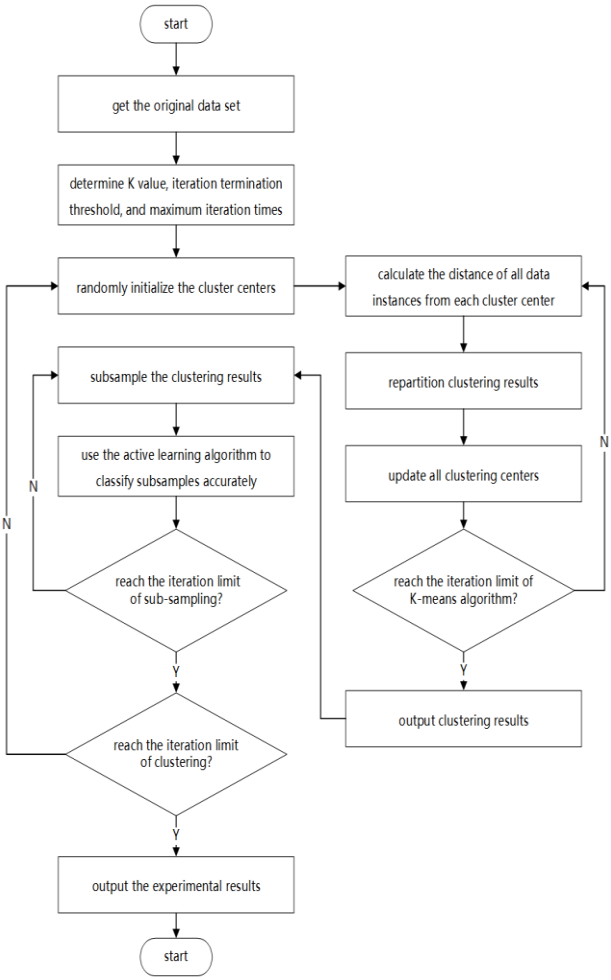


Figure 1. Standard framework flowchart.

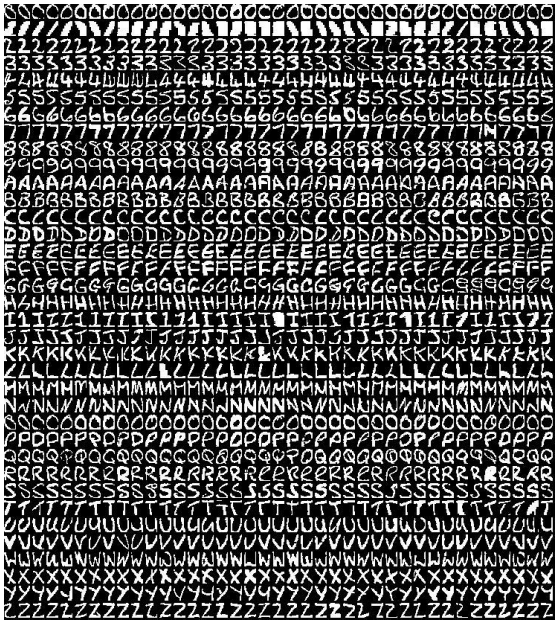


Figure 2. Binary Alphadigits dataset image [19].

Figure 3. OMNIGLOT dataset image [20].

4.2. Experimental Setting

In the experiment, for each kind of active learning algorithm, we adopted the method of subsampling, and made it select p classes of samples from the data set to form the subset $L_{sub,p}$, where $p = \{3,5,8\}$. Notice that the samples in this subset are likely to be unbalanced, with many samples in one category and few or no samples in others, which adds difficulty to our experiment. Since active learning algorithm requires labelled samples as the initial samples, we provided p labelled samples as the initial samples in each subset. These initial samples were from different categories. For the data set after unsupervised clustering, 10,000 subsets were selected, i.e. 10,000 randomized experiments were conducted. In each subset, we divided the data set into the training set by 80% and the test set 20%.

4.3. Experimental Results and Analysis

4.3.1. Experiment I: Binary Alphadigits. We plot experimental results of various active learning algorithms, as shown in figures 4a-4c, where, the horizontal axis is the number of queries (here we take the number of initial given tag samples as the initial value of the horizontal axis), and the vertical axis is the accuracy of performance indicators. In each set of experiments, we set the maximum number of queries to 80% of the total number of samples, that is, the whole training set was eventually queried. As can be seen from the results in the figure, the accuracy of each algorithm improved when the number of queries increased. Since the initial labelled instances of each algorithm is the same in each round (each subset), the classification accuracy of each algorithm remains the same on the test set

at the outset. Similarly, at the end of the round, as our maximum number of queries equals to the number of samples in the whole test set, all algorithms have the same accuracy.

The average accuracy of each algorithm is recorded in table 1. According to the results in the table, on the Binary Alphadigits data set, the active learning model based on uncertainty sampling (Unc) had the best average performance and achieved the highest average accuracy. When the value of p is 3, the average accuracy of all the active learning algorithms is above 83%. With figures 4a-4c also taken into account, the density-weighting-based active learning model can achieve a high accuracy rate with a small number of queries. However, with the increase of queries, the classification performance of density improves poorly, bring accuracy rate nearly similar to random sampling method results. As it can been seen from the table, the average accuracy of each active learning algorithm on the Binary Alphadigits dataset decreased to different degrees with the increase of the number of categories to be distinguished on each subset.

4.3.2. Experiment II: OMNIGLOT. The OMNIGLOT dataset has more categories and fewer samples per category than the Binary Alphadigits dataset, making it a greater challenge for traditional active learning methods. The performance curves of each active learning algorithm on the OMNIGLOT dataset are shown in figures 4d-4f. Similarly, we treated the initial labelled samples as query samples, so the query samples per turn started at p. The maximum number of queries in each set of experiments was 80% of the total number of samples, that is, each algorithm shall eventually query the samples of the entire training set. As can be seen from the results in the figure, the accuracy of each algorithm improves with the increase of the number of queries.

The experimental results of each algorithm on the OMNIGLOT dataset are shown in table 2. According to table 2, when adopted on the OMNIGLOT dataset, the active learning model based on query-by-committee (QBC) has the best average performance and the highest average accuracy. The Density-based active learning model was the worst performing, lower than the accuracy baseline of the random sampling active learning method. When the value of p is 3, the average accuracy of all the active learning algorithms is above 70%. As the number of categories to be distinguished on each subset increases, the average accuracy of each active learning algorithm on OMNIGLOT dataset decreases significantly.

Table 1. Experimental results of the Binary Alphadigits dataset.

Algorithm	3 classes		5 classes		8 classes	
	Average Accuracy	Number of queries	Average Accuracy	Number of queries	Average Accuracy	Number of queries
Random	83.4%	24	75.7%	40	68.5%	64
Density	83.5%	24	76.1%	40	68.9%	64
LAL	84.5%	24	75.8%	40	68.2%	64
QBC	86.1%	24	78.5%	40	71.3%	64
Unc	86.6%	24	78.9%	40	71.5%	64

Table 2. Experimental results of the OMNIGLOT dataset.

Algorithm	3 classes		5 classes		8 classes	
	Average Accuracy	Number of queries	Average Accuracy	Number of queries	Average Accuracy	Number of queries
Random	72.5%	24	63.2%	40	55.0%	64
Density	70.2%	24	61.5%	40	53.7%	64
LAL	73.3%	24	63.4%	40	54.9%	64
QBC	74.2%	24	64.9%	40	56.5%	64
Unc	73.9%	24	64.2%	40	55.3%	64

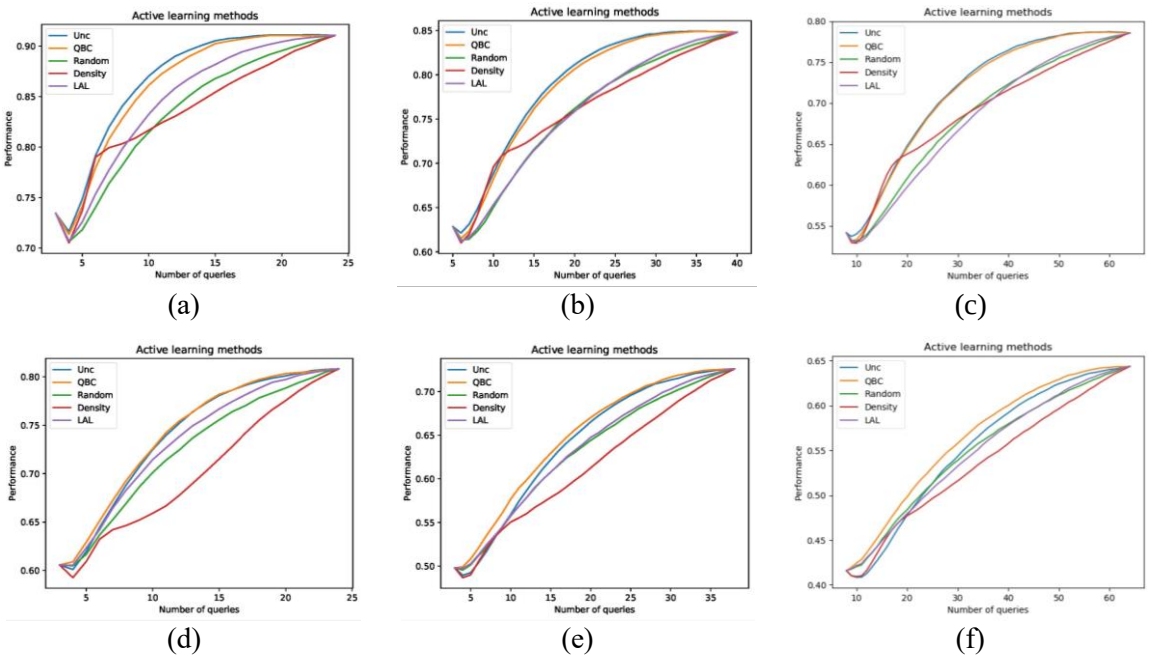


Figure 4. Algorithm performance comparison diagram: (a)-(c) is the experimental results of the Binary Alphadigits dataset. The number of sample categories contained in the subsample is 3,5,8, respectively; (d)-(f) is the experimental results of the OMNIGLOT dataset, and the number of sample categories contained in the subsample is 3,5,8, respectively.

In 8-fold classification problems, the average accuracy of all the algorithms came out less than 60%. In this sense, with the number of classes increase, the above active learning methods remains incompetent. If subsampling-based active learning method is not adopted, traditional active learning methods could not handle multi-classification tasks of small samples but with immense categories and scarce samples per category, such as the OMNIGLOT dataset.

5. Conclusion

In view of the limitations of traditional active learning methods in multi-category data sets, this paper proposes a subsampling-based active learning method for multi-category data sets. This method adopts a standard framework integrated by unsupervised clustering and active learning methods. Through subsampling process of clustering results, active learning method trains the model through the annotated sub-dataset instead of the original large-scale dataset. The proposed method was tested on Binary Alphadigits and OMNIGLOT datasets.

Experimental results show that the subsampling-based approach proposed in this paper can handle the problem that it is difficult for active learning algorithm. This method possesses considerable applicability and prospects in a variety of practical application backgrounds. It can effectively resolve the problem of missing annotations in deep learning model.

Acknowledgments

Thanks to Honglan Huang, Shixuan Liu and Yanghe Feng for their help and guidance.

References

[1] Esteva A, Kuprel B, Novoa R A, Ko J, Swetter S M, Blau H M and Thrun S 2017 Dermatologist-level classification of skin cancer with deep neural networks *Nature* **542** (7639) 115-118.

[2] Zhou Z, Shin J, Zhang L, Gurudu S, Gotway M and Liang J 2017 Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally *The IEEE Conf. on Computer Vision and Pattern Recognition* (Honolulu, HI) pp 4761-4772.
[3] Lewis D D and Gale W A 1994 A sequential algorithm for training text classifiers *17th Annual International ACM SIGIR Conf. on Research & Development in Information Retrieval* (Dublin, Ireland) pp 3-12.
[4] Settles B and Craven M 2008 An analysis of active learning strategies for sequence labeling tasks *Conf. on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii) pp 1070-1079.
[5] Freund Y, Seung H S, Shamir E and Tishby N 1997 Selective sampling using the query by committee algorithm *Machine Learning* **28** (2-3) 133-168.
[6] McCallum A and Nigam K 1998 Employing EM in pool-based active learning for text classification *Proc. of the 15th International Conf. on Machine Learning* (Madison, Wisconsin) pp 350-358.
[7] Nicholas R and Andrew M 2001 Toward optimal active learning through sampling estimation of error reduction *Proc. of the 18th International Conf. on Machine Learning* pp 441-448.
[8] Cohn D, Ghahramani Z and Michael J 1996 Active learning with statistical models *Journal of Artificial Intelligence Research* **4** 129-145.
[9] Zhang T and Oles F J 2000 A probability analysis on the value of unlabeled data for classification problems *Proc. of the 17th International Conf. on Machine Learning* (Stanford University) pp 1191-1198.
[10] Huang H L, Huang J C, Feng Y H and Zhang J R 2019 On the improvement of reinforcement active learning with the involvement of cross entropy to address one-shot learning problem *PLoS ONE* **14** (6) 1-17.
[11] Huang H L, Feng Y H, Huang J C and Zhang J R 2019 A reinforcement one-shot active learning approach for aircraft type recognition *IEEE Access* **7** 147204-147214.
[12] Brinker K 2003 Incorporating diversity in active learning with support vector machines *20th International Conf. on Machine Learning* pp 59-66.
[13] Kai Y, Jinbo B and Volker T 2006 Active learning via transductive experimental design *Proc. of the 23rd International Conf. on Machine Learning* (Pittsburgh, Pennsylvania) pp 1081-1088.
[14] Yang Y Z and Marco L 2019 Single shot active learning using pseudo annotators *Pattern Recognition* **89** 22-31.
[15] Zhang C and Chen T 2002 An active learning framework for content-based information retrieval *IEEE Trans. on Multimedia* **4** 260-258.
[16] Pinar D, Jaime C and Paul B 2007 Dual strategy active learning *European Conf. on Machine Learning* (Warsaw, Poland) pp 116-127.
[17] Hieu T N and Arnold S 2004 Active learning using preclustering *Conf. on Machine Learning* (Banff, Canada) pp 623-630.
[18] Burr S and Mark C 2008 An analysis of active learning strategies for sequence labeling tasks *Conf. on Empirical Methods in Natural Language Processing* (Honolulu, Hawaii) pp 1070-1079.
[19] Sam R *Vision, Learning and Graphics Group* www.cs.nyu.edu/~roweis.
[20] Simon *A Omniglot-Writing Systems and Languages of the World* www.omniglot.com.