Active Learning to Select Unlabeled Examples with Effective Features for Document Classification

Minoru Sasaki

Ibaraki University
Department of Computer and Information Science
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511
minoru.sasaki.01@vc.ibaraki.ac.jp

Abstract. In this paper, for document classification task and text mining based on machine learning, I propose a new pool-based active learning method to select unlabeled data that have effective features not found in the training data. Given a small set of training data and a large set of unlabeled data, the active learner selects the most uncertain data that has effective features not found in the training data from the unlabeled data and asks to label it. After capturing these uncertain data from the unlabeled data repeatedly, I apply the existing pool-based active learning to select training data from the unlabeled data efficiently. Therefore, by adding data with effective features from unlabeled data to training data, I consider that it is effective to improve the performance of the pool-based active learning. To evaluate the efficiency of the proposed method, I conduct some experiments and show that our active learning method achieves consistently higher accuracy than the existing algorithm.

1 Introduction

In recent years, research to solve various automatic classification problems has been done such as spam mail filtering and documents classification. To solve these problems, machine learning methods are applied for classification problem. In the machine learning, many training data are needed to construct a classifier that can correctly predict the classes of new objects. For document classification and text mining based on supervised learning, learning algorithms require enough labeled training data to construct a classifier. However, it is hard to obtain a large amount of labeled data and it is time-consuming with a lot of cost to label a large number of data. In addition to the quantity of the training data, the quality of the training data used to obtain the classifier is critical for accurate and repeatable results. Even when a large number of labeled data are available, sometimes a good classifier cannot be obtained. Therefore, I need to improve the quality of the training data and reduce the amount of noise to achieve the better performance of the classifier.

To overcome the above labeling problems, active learning techniques have been successfully applied to classification problems to reduce the labeling effort. Active learning aims to automatically select the next document to label for training accurate classifiers. Though there are many variants of active learning in the literature, the focus of this article is the pool-based active learning, model which is the most widely used[1][2][3].

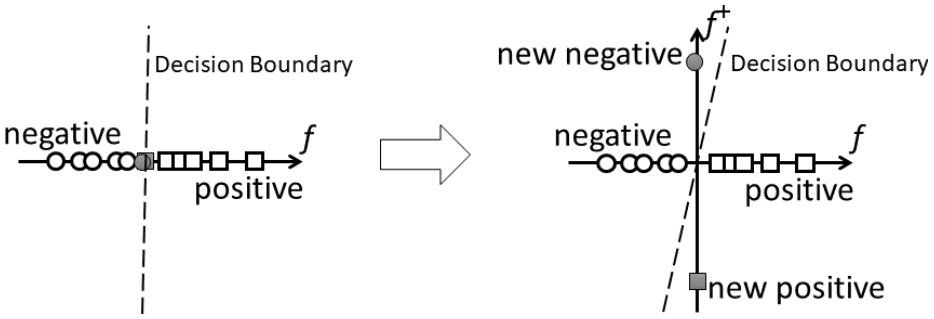

Fig. 1. Proposed active learning method to select unlabeled data that have effective features not found in the training data

Given a small set of training data L and a large set of unlabeled data U, a classifier is trained with L and the active learner selects the most uncertain data for a classifier from U and asks to label them. The L is augmented with this data and the process is repeated until a stopping criterion is met. However, in this method, uncertainty of unlabeled data is calculated with features in training data. In the previous paper [3], it was reported that active learning method for insufficient-evidence uncertainty performs worse than the existing uncertainty sampling. The insufficient-evidence uncertainty represents an uncertainty of a model due to insufficient evidence for each class. Insufficient-evidence uncertain instances do not have effective features in the training data for classifications. In this case, it is difficult to select informative samples from an unlabeled dataset pool. Therefore, in the early iteration of active learning, the existing methods tend not to improve predictive performance of the classifier (Fig. 1. left.). By solving this problem, I consider that a optimal classifier can be learned efficiently for initial small training data set.

In this paper, I propose a new pool-based active learning method to select unlabeled data that have effective features not found in the training data for the document classification task and the text mining based on machine learning. Given a small set of training data L and a large set of unlabeled data U, the active learner selects the most uncertain data that has effective features not found in L from U and asks to label it (Fig. 1. right.). After capturing these uncertain data from U repeatedly, I apply the existing pool-based active learning to select training data from U efficiently. Therefore, by adding data with effective features from unlabeled data to training data, I consider that it is effective to improve the performance of the pool-based active learning.

The organization of residual of the paper is as follows. In Section 2, I introduce the active learning framework and uncertainty sampling used in this paper. In Section 3, I describe the proposed method. In Section 4, an outline of experiments and experimental results are presented. Finally, Section 5 concludes this paper.

2 Related Works

In this section, I provide a brief description of the active learning in the context of classification and review the relevant literature to explain the existing researches.

2.1 Active Learning

The main task of active learning is to automatically select the informative instances for efficiently reducing the sample complexity. By using active learning techniques, the number of labeled examples required by machine learning algorithms can be reduced.

Active learning can be divided into two main settings: stream-based active learning sampling and pool-based active learning[4]. In stream-based active learning, each instance is sampled from some distribution in a streaming manner and the learner has to decide whether to label this instance or discard it immediately[5]. In the pool-based setting, at each active selection step, the active learner chooses one or more instances that is (are) added to a training set from a large pool of unlabeled instances and the learner retrain the model. In this work, I focus on uncertainty sampling for the pool-based active learning, which is one of the most common setups in active learning [1].

2.2 Uncertainty Sampling

One intuitive approach in pool-based active learning is called uncertainty sampling which selects instance that the learner is the most uncertain about [1]. In previous uncertainty sampling methods, a popular uncertainty sampling strategy employs entropy to evaluate the uncertainty [6][7][8]. This entropy measure can be employed easily to probabilistic multi-label classifiers for complex structured instances [9][10]. In a recent paper, the entropy measure is used to evaluate the uncertainty of random variables via random walks on a graph [11].

In the uncertainty sampling topic, Sharma et al. distinguish between the two types of uncertainties, conflicting-evidence uncertainty and insufficient-evidence uncertainty to improve the performance of uncertainty sampling [3]. The experiments showed that the conflicting uncertain instances are effective for classification. However, the method that selects uncertain instance due to insufficient evidence performs worse than the baseline method that selects the t-th most uncertain instance for almost all datasets. In this paper, to solve the problem of this insufficient-evidence uncertainty, I propose a novel active learning framework to select unlabeled data that have effective features not found in the training data.

3 Proposed Method

In this section, I describe the algorithm of our proposed method to select unlabeled data that have effective features not found in the training data.

Figure 2 shows the proposed active learning framework. First, I randomly select five positive examples D_P and five negative examples D_N as the initial training set D_T. D_{UL} is a pool of unlabeled data samples. F_{UL} is a set of features that appear in

the unlabeled data D_{UL} but do not appear in the training data D_T. $D_{F_{UL}}$ is a set of examples with features in the F_{UL} in the D_{UL}.

First, the proposed method constructs a classification model M from the training data D_T using Naïve Bayes classifier. If the set F_{UL} is not empty, for each unlabeled data x in the set $D_{F_{UL}}$, conditional entropy $H(x)$ is calculated as an uncertainty measure as follows:

$$H(x) = -\sum_{i=1}^{Y} p(y_i|x) \log p(y_i|x).$$

The method selects the candidate unlabeled instance x_{max} that has the largest conditional entropy. Then, this candidate instance x_{max} is labeled with the correct answer and added to D_T. This process is repeated until $D_{F_{UL}}$ becomes empty.

If $D_{F_{UL}}$ becomes empty before the size of the training set n_{D_T} exceeds the maximum size n_{max}, the proposed method chooses representative instance for which the classifier M is uncertain due to conflicting evidence.

For each unlabeled data y in the set D_{UL}, conditional entropy $H(y)$ is calculated and this method extracts the top t largest uncertain unlabeled examples. Among the top t uncertain instances T, the proposed method chooses representative instance z_{max} for which the model is uncertain due to con icting evidence as follows:

$$z_{max} = \arg\max_{z \in T} \log E_{+1}(z) + \log E_{-1}(z),$$

where the scores for the example z to belong to the positive and negative class is

$$\log E_{+1}(x) = \sum_{x_j \in Pos_x} \log \frac{p(x_j|+1)}{p(x_j|-1)},$$

$$\log E_{-1}(x) = \sum_{x_j \in Neg_x} \log \frac{p(x_j|-1)}{p(x_j|+1)}.$$

The Pos_x and Neg_x contain the attribute values of the example x_j that provide evidence for the positive class and the negative class respectively. Then, this example z_{max} is labeled with the correct answer and added to D_T.

4 Experiments

To evaluate the effectiveness of the proposed active learning method, I perform some experiments and compare the results of the previous method.

4.1 Data Set

In this paper, I experimented with two publicly available datasets that can be used for text classification and text mining such as Spambase [1] and Internet Advertisements [2],

[1] https://archive.ics.uci.edu/ml/datasets/spambase

[2] https://archive.ics.uci.edu/ml/datasets/internet+advertisements

function Uncertainty-Sampling (D_T, D_{UL});
Input : Training data $D_T = \{D_P, D_N\}$ and unlabeled data D_{UL}
Output : Labeled training data set D_L
Classifier M is trained with D_T;
while $n_{D_T} \leq n_{max}$ **do**
    $EN_{max} \leftarrow -\infty$;
    $F_{UL} \leftarrow$ features that are not found in D_T;
    $D_{F_{UL}} \leftarrow$ examples of D_{UL} with features in F_{UL};
    if $D_{F_{UL}} \neq \phi$ **then**
        for $x \in D_{F_{UL}}$ **do**
            if $H(x) > EN_{max}$ **then**
                $EN_{max} \leftarrow H(x)$;
                $x_{max} \leftarrow x$;
            end
        end
        $D_T \leftarrow D_T \cup x_{max}$;
        $D_{F_{UL}} \leftarrow D_{F_{UL}} \setminus x_{max}$;
        $D_{UL} \leftarrow D_{UL} \setminus x_{max}$;
        Re-train classifier M with D_T;
    end
    else
        $E_{max} \leftarrow -\infty$;
        $Top \leftarrow$ The top t most uncertain examples in D_{UL};
        for z *in* Top **do**
            if $\log E_{+1}(z) + \log E_{-1}(z) > E_{max}$ **then**
                $E_{max} \leftarrow \log E_{+1}(z) + \log E_{-1}(z)$;
                $z_{max} \leftarrow z$;
            end
        end
        $D_T \leftarrow D_T \cup z_{max}$;
        $D_{UL} \leftarrow D_{UL} \setminus z_{max}$;
        Re-train classifier M with D_T;
    end
end

Fig. 2. Proposed Uncertainty Sampling

from UCI machine learning repository. The Spambase dataset consists of 4601 email messages in which 1813 are spam and 2788 are non-spam emails. Each email has 57 numeric features that indicate the frequency of spam related term occurrences and lengths of uninterrupted sequences of capital letters. The Internet Advertisements is a popular dataset for predicting if a given image is an advertisement or not[12]. It contains 3279 examples and 1558 features which include phrases occurring in the URL, the anchor text, words occurring near the anchor text and the geometry of the image and so on. Since there are missing values in about 28% of the examples, I conduct experiments using 2359 examples (excluding missing values from the dataset).

4.2 Experiments on Active Learning

I experimented with the four uncertainty sampling methods such as the proposed method and the three existing methods(conflicting-evidence uncertainty, insufficient-evidence uncertainty and uncertainty sampling)[3], I evaluated each method using a multinomial Naïve Bayes classifier for class probability estimation.

In the experiments, I use five-fold cross validation to evaluate the proposed method. I divide the whole dataset into five equal-size subsets. For the 80% of the data, five positive examples D_P and five negative examples D_N are selected as the initial training set D_T and the rest of the data is used for a pool of unlabeled data D_{UL}. The remaining 20% of the data is used for test data. Uncertainty sampling methods for the conflict-evidence uncertainty and the insufficient-evidence uncertainty operate within top 10 uncertain instances as the initial training set. The maximum training data size n_{max} in Algorithm 2 was set to 500 instances. For each fold, the experiment was repeated five times and the final score is the average precision over the five results.

4.3 Experimental Results

In this section, I present the experimental results for the proposed method and the three existing methods.

Figure 3 shows the average learning curves for all methods on the Spambase dataset. The results of the experiments show that the proposed method achieves consistently higher accuracy than the existing methods in an early stage of the iteration process, leading to higher final accuracy overall. Since the Spambase dataset has only a small number of features, all the features appear in the training data in the early iteration of active learning (about 10 examples). By extracting features that are not included in the training data, the proposed method can perform uncertainty sampling more effectively than the existing methods. However, when the number of iteration increases, average precision tends to vary between 80% and 90% by using any method. In future, I would like to find the optimum number of iteration to improve the performance of the proposed method.

Figure 4 shows the average learning curves for all methods on the Internet Advertisements dataset. The learning curve of the proposed method fluctuates, while the proposed method selects the most uncertain example that has features not found in training data. However, the results of the experiments show that the proposed method achieves consistently higher accuracy from the middle of the learning iteration.

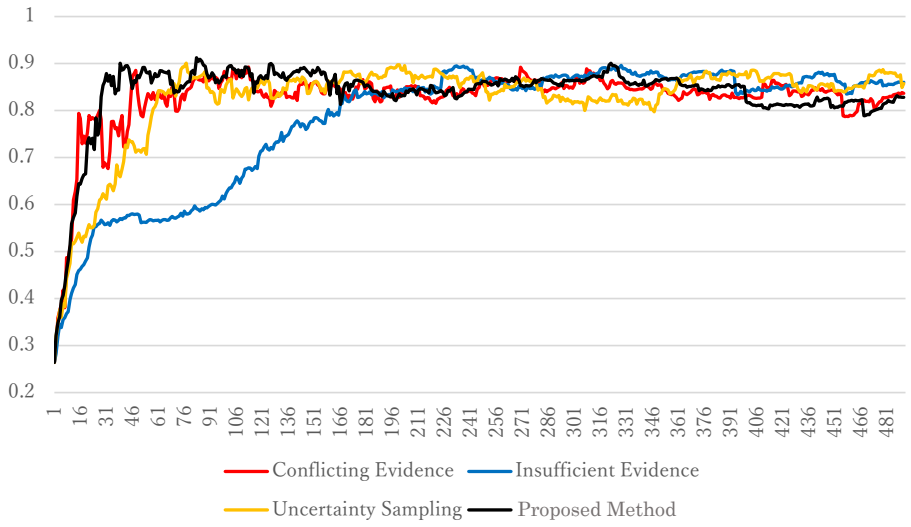

Fig. 3. Experimental Result on the Spambase

Even when a model is uncertain because it does not have sufficient evidence for either class, uncertainty sampling for insufficient-evidence performs significantly worse than the other methods. However, the proposed method can solve the sampling problem of insufficient-evidence uncertainty and improve the accuracy of active learning in the early iteration.

5 Conclusion

In this paper, I proposed pool-based active learning method to select the unlabeled data that has effective features not found in the training data. In traditional uncertainty sampling, uncertainty of unlabeled data is calculated in the feature space which is generated by the training data. By adding data with effective features using the proposed method, I consider that it is effective to improve the performance of the pool-based active learning.

To evaluate the efficiency of the proposed method, I conduct some experiments to compare with the result of the baseline method. The results of the experiments on the Spambase dataset show that our active learning algorithm achieves consistently higher accuracy than the existing algorithm in an early stage of the iteration process, leading to higher final accuracy overall. Therefore, it is shown that the proposed method is effective for active learning.

In another dataset, Internet advertisements, I got a result that classification accuracy stabilizes near the highest point after switching to conventional method. By these results, I could indicate that selecting unlabeled data which have features not found in the training data before switching to conventional method is effective.

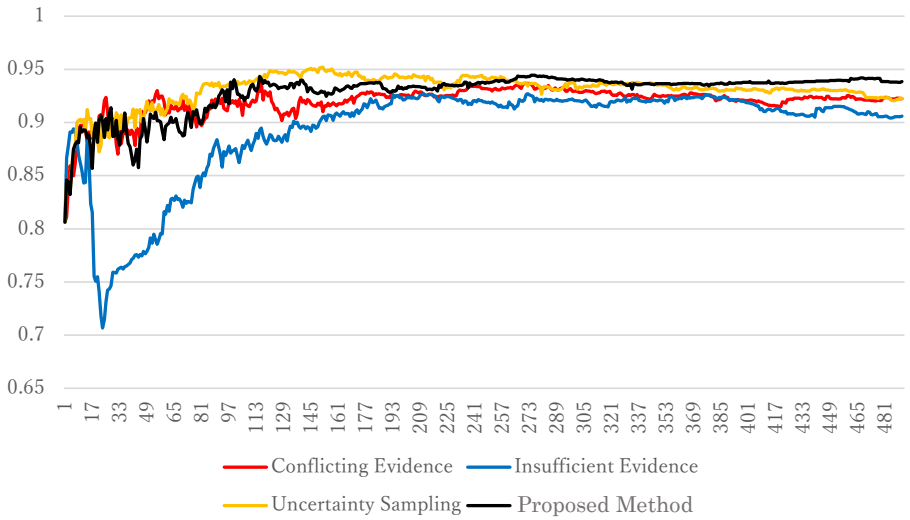

Fig. 4. Experimental Result on the Internet Advertisements

References

1. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '94, New York, NY, USA, Springer-Verlag New York, Inc. (1994) 3–12
2. McCallum, A., Nigam, K.: Employing em and pool-based active learning for text classification. In: Proceedings of the Fifteenth International Conference on Machine Learning. ICML '98, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1998) 350–358
3. Sharma, M., Bilgic, M.: Evidence-based uncertainty sampling for active learning. Data Mining and Knowledge Discovery **31** (2017) 164–202
4. Settles, B.: Active learning literature survey. Technical report (2010)
5. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning **15** (1994) 201–221
6. Tang, M., Luo, X., Roukos, S.: Active learning for statistical natural language parsing. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 120–127
7. Chen, J., Schein, A., Ungar, L., Palmer, M.: An empirical study of the behavior of active learning for word sense disambiguation. In: Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. HLT-NAACL '06, Stroudsburg, PA, USA, Association for Computational Linguistics (2006) 120–127
8. Zhu, J., Hovy, E.: Active learning for word sense disambiguation with methods for addressing the class imbalance problem. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007)

9. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP '08, Stroudsburg, PA, USA, Association for Computational Linguistics (2008) 1070–1079
10. Hwa, R.: Sample selection for statistical parsing. Comput. Linguist. **30** (2004) 253–276
11. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. International Journal of Computer Vision **113** (2015) 113–127
12. Kushmerick, N.: Learning to remove internet advertisements. In: Proceedings of the Third Annual Conference on Autonomous Agents. AGENTS '99, New York, NY, USA, ACM (1999) 175–181