REPRESENTATIVE SAMPLING WITH CERTAINTY PROPAGRATION FOR IMAGE RETRIEVAL

Jian Cheng, Biao Niu, Yikai Fang, Hanqing Lu

a National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, China
b Nokia Research Center, Beijing, China

ABSTRACT
Selective sampling has been widely used in relevance feedback of image retrieval to alleviate the burden of labeling by selecting the most informative instances for user to label. Traditional sample selection scheme often selects a batch of instances each time and label them simultaneously, which ignores the correlation among instances and results in redundant labeling. In this paper, we propose an improved representative sampling method with certainty propagation to improve the performance of sampling. In our method, two kinds of correlations among instances are explored to reduce the redundancy in sampling. One is the correlation between labeled instances and unlabeled instances. The other is the correlation among unlabeled instances. Extensive experiments show that the proposed method achieve encouraging results.

Index Terms—Selective sampling, image retrieval, SVM, Certainty propagation

1. INTRODUCTION

Relevance feedback is an effective approach for image retrieval to narrow the semantic gap between low-level feature and high-level semantic interpretation [1]. In relevance feedback process, users are taken into the loop of retrieval by iteratively labeling some unlabeled instances. However, labeling is tedious and time-consuming, and more labeling does not necessarily lead to better results. Therefore, the critical issue in relevance feedback is how to efficiently and effectively select the helpful unlabeled instances for user to label.

Selective sampling is one of the most popular active learning methods [2,3]. Typically, selective sampling chooses the most informative samples for user to label. The classic sampling strategies include query-by-committee [2], optimal experimental design [4], and uncertainty sampling [5]. One of the earliest works in image retrieval is SVM_{active} proposed by Tong and Chang [6]. SVM_{active} chooses the informative samples that can reduce the version space as fast as possible, i.e. eliminating the hypotheses. Thus, in each round of relevance feedback, the instances that are closest to the hyperplane of SVM classifier are regarded as most uncertain instances, and are returned to users for labeling. SVM_{active} retrains the classifier after a single instance is labeled in each iteration, which is named one-by-one mode in this paper. The one-by-one mode is rather time-consuming due to the complex SVM classifier learning.

In real relevance feedback process, the system usually allows users to label multiple instances simultaneously, which is called batch mode. As a consequence, the traditional selective sampling methods are possible to select multiple unlabeled instances that contain much redundancy. In order to facilitate users to feedback, most of selective sampling methods adopt the batch mode which updates classifier after labeling multiple instances [7,8]. Hoi, et al. [7] presented a semi-supervised SVM batch mode active learning algorithm. In [8], a Dynamic certainty propagation algorithm (DCP) is proposed to investigate the certainty propagation between the currently labeled instance and the remaining unlabeled data. However, there still exist amount of redundancy among the unlabeled instances which has been rarely discussed in literature. Recently, Huang, et al. designed a new selective sampling approach based on [7], called QUIRE [9]. QUIRE attempts to find the unlabeled samples that are both informative and representative. Although QUIRE reduces the redundancy among unlabeled instances, it does not yet consider the correlation between labeled instances and unlabeled instances.

In this paper, we propose an improved representative sampling method with certainty propagation to improve the performance of batch mode sampling in relevance feedback. Our method combines the merit of DCP and QUIRE to explore two kinds of correlations among instances. One is the correlation between labeled sample and unlabeled samples. The proposed method attempts to reduce the redundancy existing in the first correlation by certainty propagation. The other is the correlation among unlabeled samples, which is minimized by min-max view of active learning [7]. The extensive experiments show that the proposed method can improve the performance of selective sampling in image retrieval.
2. REPRESENTATIVE SAMPLING WITH CERTAINTY PROPAGATION

In batch mode selective sampling, two critical issues need to be addressed. One is that the correlation between the labeled instances and the unlabeled instances, i.e. the selected unlabeled instances to be labeled should be less redundant with the previous labeled instances. The other issue is the correlation among the unlabeled instances, i.e. the selected instance should be representative in the unlabeled sample set. In order to address the two issues, we propose a new batch mode selective sampling, named representative sampling with certainty propagation, for relevance feedback in image retrieval. We use a toy data in Fig.1 to illustrate the difference of our method. In Fig.1, the white circle represents unlabeled instances and the black disc denotes sampled instances.

Assume \( X = \{x_1, x_2, \ldots, x_n\} = U \cup L \) denote the entire dataset, where \( U \) and \( L \) denote the unlabeled and labeled dataset, respectively. \( U = U' \cup \{x_i\} \) represents the unlabeled data set, where \( x_i \) is the currently selected instance and \( U' \) is the rest of the unlabeled set. For each instance \( x_i \in U' (1 \leq i \leq n) \), \( y_i \in \{-1, 1\} \) is its corresponding class label. If \( x_i \in L \), the label is known; while if \( x_i \in U' \), the label is unknown and we use \( f(x_i) \) to predict \( y_i \), where \( f \) denotes the SVM classifier on \( L \).

For each unlabeled instance, we introduce the notion of certainty as a measure of the confidence in its predicted label. The term degree of certainty is used to denote the absolute value of certainty. The lower degree of certainty an unlabeled instance has, the more informative it is. The initial certainty of instance \( x_i \in U \) is:

\[
C^{(0)}(x_i) = f(x_i), x_i \in U
\]

First, we select the most representative instance from the unlabeled set by considering the cluster structure of unlabeled instances. The measure of the representativeness is the same as QUIRE \[9\]:

\[
\begin{align*}
    s &= \arg \min_{\alpha \in \ell} \hat{L}(L, U', x_i), \\
    l_{\text{ua}} &= \min_{\alpha \in \ell} \hat{L}(L, U', x_i).
\end{align*}
\]

Once the most representative instance is labeled, we change the certainty of the instance:

\[
C(x_i) = \begin{cases} 
    M & (y_i > 0) \\
    -M & (y_i < 0)
\end{cases}
\]

where \( M \) is a positive constant set beforehand to represent the labeled instance’s degree of certainty.

Then, the change of certainty of \( x_i \) is propagated to the remaining unlabeled instances, which is intuitive in practice. The propagated certainty will reduce with the similarity decrease:

\[
C^{(l+1)}(x) = C^{(l)}(x) + w_{ui} [C^{(l)}(x) - C^{(l)}(x_i)]
\]

where \( w_{ui} \) measures the pair-wise relationship between instance \( x_i \) and the current labeled instance \( x_j \) with the heat kernel:

\[
w_{ui} = \exp\left(-\frac{||x_i - x_j||}{t}\right)
\]

where two parameters need to be explained, \( t \) and \( M \). \( M \) is the degree of the labeled data. According to three sigma rule, we set it as follows:

\[
M = \mu_c + 3\sigma_c,
\]

\[
\mu_c = \frac{1}{|U|} \sum_{x \in U} |C^{(l)}(x)|,
\]

\[
\sigma_c^2 = \frac{1}{|U|} \sum_{x \in U} [C^{(l)}(x) - \mu_c]^2.
\]

Parameter \( t \) is a scale parameter controlling the range of certainty propagation. It is reasonable that \( t \) begin with a large scale so that the search does not fall into the local optimum. After several iterations the classifier is more confident with its performance. \( t \) can be gradually decreased to obtain a more accurate classification boundary. The minimum in eqn.2 can be used as the indication to adjust the clustering. The scale parameter is decreased as \( t = 2^{-\frac{l}{l_{\text{max}}}} \).

The detailed flowchart of the proposed method is depicted in Fig. 2:

**Input**: The classifier \( f \), the labeled training dataset \( L \), the unlabeled dataset \( U \), the weight matrix \( W \), and the number of samples to be labeled \( N_i \) in each iteration.

**Step1.** Classify \( U \) with \( f \), and initialize the certainty of unlabeled data:

\[
C^{(0)}(x_i) = f(x), x_i \in U.
\]

\[ S = \emptyset \]

**Step2.** For \( r = 1 \) to \( N_i \), repeat the following procedure:

1) Select the most informative and representative instance \( x_i, x_j \in U \) according to equation (2);
2) Add $x_i$ to $S$ and remove it from $U$:
$$S = S \cup \{x_i\}; U = U - \{x_i\}.$$ 
3) Change $x_i$’s certainty according to equation (3); 
4) Propagate the change in $x_i$’s certainty to the rest unlabeled data according to equation (4); 

**Step3.** Train a new classifier $f^*$ using $L = L \cup S$.

![Fig. 2. The flowchart of the proposed method.](image)

### 3. EXPERIMENTS

To evaluate performance of the proposed algorithm, we conduct experiments by comparing it with several state-of-the-art selective sampling algorithms in image retrieval. The experiments are performed on a subset select from the Corel image CDs. In particular, we collect a 50-category dataset that contains 5,000 images from 50 different categories. Each category represents a different semantic topic, such as tiger, car, flag, etc. Color and texture features are extracted to represent images. The color features consist of 125-dimensional color histogram and 6-dimensional color moment in each color channel (R, G, and B), respectively. The texture features are extracted using 3-level discrete wavelet transformation, and the mean and variance averaging on each of 10 sub-bands form a 20-dimensional vector.

#### 3.1. Experimental Setup

All SVM classifiers in our experiments use the same RBF kernel [10]. In each round of relevance feedback, we label the same amount of images selected by different algorithms. In experiments, the performance is evaluated by average accuracy. The first 10 images of each category, 500 in total, are chosen as query to test performance of the compared methods.

In experiments, we compare the proposed algorithm with the following algorithms for selective sampling:
- $SVM_{Active}$ (denoted as Active SVM in figure 3-6): the baseline which just chooses the instances closest to the classification boundary for labeling [6].
- Dynamic Certainty Propagation (DCP): a batch model selective sampling algorithm which explores the correlation between the labeled instance and unlabeled instances [8].
- QUIRE: A selective sampling algorithm which exploits the correlation among the unlabeled instances to look for the representative and informative instance [9].

#### 3.2. Comparison of Selective Sampling Strategies

![Fig. 3. The accuracy of image retrieval with 3 feedbacks](image)

![Fig. 4. The accuracy of image retrieval with 4 feedbacks](image)

Fig. 3 and Fig. 4 show the accuracy vs. scope curves after 3 and 4 rounds of relevance feedback, respectively. Here scope = x means the accuracy is calculated within the top x returned images. The comparison results indicate that DCP, QUIRE and our method are all superior to the baseline, $SVM_{Active}$ (Active SVM), which suggest that considering the correlation among instances can improve the effectiveness of selective sampling. QUIRE attempts to find the representative and informative instance so that it is slightly better than DCP which only look for the informative instance. Due to consider both labeled-unlabeled correlation and unlabeled-unlabeled correlation, our method also has the advantage of DCP and QUIRE. Fig.5 and Fig.6 show that the relevance feedback can significantly improve the performance of image retrieval.
4. CONCLUSION

Relevance feedback has been widely adopted to narrow the semantic gap in image retrieval. The selective sampling is one of the popular active learning methods and is proved to be effective for relevance feedback. In this paper, we propose an improved representative sampling method with certainty propagation to find the informative and representative instances for user to label. In our method, two kinds of correlations among instances are explored to reduce the redundancy in sampling. One is the correlation between labeled instances and unlabeled instances. The other is the correlation among unlabeled instances. Extensive experiments show that the proposed method can achieve encouraging results.

Fig. 5. The top 30 accuracy of image retrieval

Fig. 6. The top 50 accuracy of image retrieval

ACKNOWLEDGEMENTS

This work is partially supported by the National Natural Science Foundation of China (Grant No.60975010, 61070104, and 60873185), and Young Talents Project of CASIA.

5. REFERENCES


