Pattern Recognition 78 (2018) 358–370

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# A variance maximization criterion for active learning

Yazhou Yang[a,b,*], Marco Loog[a,c]

[a] Pattern Recognition Laboratory, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
[b] College of Information System and Management, National University of Defense Technology, Changsha, China
[c] Image Group, University of Copenhagen,Universitetsparken 5, DK-2100, Denmark

ARTICLE INFO

*Article history:*
Received 23 June 2017
Revised 9 January 2018
Accepted 22 January 2018
Available online 31 January 2018

*Keywords:*
Active learning
Retraining information matrix
Variance maximization

ABSTRACT

Active learning aims to train a classifier as fast as possible with as few labels as possible. The core element in virtually any active learning strategy is the criterion that measures the usefulness of the unlabeled data based on which new points to be labeled are picked. We propose a novel approach which we refer to as maximizing variance for active learning or MVAL for short. MVAL measures the value of unlabeled instances by evaluating the rate of change of output variables caused by changes in the next sample to be queried and its potential labelling. In a sense, this criterion measures how unstable the classifier's output is for the unlabeled data points under perturbations of the training data. MVAL maintains, what we refer to as, retraining information matrices to keep track of these output scores and exploits two kinds of variance to measure the informativeness and representativeness, respectively. By fusing these variances, MVAL is able to select the instances which are both informative and representative. We employ our technique both in combination with logistic regression and support vector machines and demonstrate that MVAL achieves state-of-the-art performance in experiments on a large number of standard benchmark datasets.

## 1. Introduction

In many real-world applications of classification problems, we face the problem that obtaining labels is more difficult than collecting input data: we can easily acquire a large amount of such input data, but labeling these instances is quite burdensome, time-consuming, or expensive [46]. For a large part, this is because of the heavy involvement of human supervision during the labeling process. For example, a hospital produces large amounts of digital images every day, but when categorizing these medical images one often needs to rely on medical doctors with a particular, and therefore expensive, expertise. Hence, it is essential to reduce the need for human annotation, bringing down cost by labeling fewer yet more informative samples. The problem studied in active learning is how to select the most valuable subset and how to measure the value of individual instances or collections of these.

In this work, we focus on, what we refer to as, retraining-based active learning in which one measures the usefulness of particular instances based on all the possible models that are obtained by adding the instances to the labeled dataset and retraining the classifier with the different labels possible [40,44,51]. This means that with $n$ unlabeled points and $k$ different classes to choose from, we train $nk$ different classifiers. The key idea behind this is that the value of an unlabeled instance can be estimated by the change it brings to the model when it is queried and used to retrain the model.

Here we propose a new retraining-based active learning method: maximizing variance for active learning (MVAL). Our method selects the instances with maximum retraining variance. This variance stems from the variation presented in the next sample to query and the possible labels those samples can have. The idea is that if the output of an instance changes dramatically, it means that this instance is very susceptible to the variations of input training data. On the other hand, if an instance's output does not vary much, this indicates that the current classifier is very certain about it. A sample with the largest changes in output value is most uncertain and this rate of change can be naturally measured by the variance. Thus, the larger the variance of the output of an unlabeled instance, the higher the uncertainty it has. We propose to keep track of the estimated probability (or decision output) of each unlabeled instance during the retraining procedure. The recorded information is utilized to produce so-called retraining information matrices (RIMs), which are used to calculate the variances for all unlabeled samples. More specifically, two

[*] Corresponding author at: Pattern Recognition Laboratory, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands.
*E-mail addresses:* y.yang-4@tudelft.nl (Y. Yang), m.loog@tudelft.nl (M. Loog).

https://doi.org/10.1016/j.patcog.2018.01.017
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

different kinds of variance are computed to measure the informativeness and representativeness. By selecting the instances with maximum variance, MVAL is able to query instances that are both informative and representative. Furthermore, MVAL can be incorporated with both probabilistic and non-probabilistic classifiers, such as logistic regression, Naive Bayes, support vector machines and least squares classifier. In this paper, we construct the experiments of MVAL with logistic regression and support vector machines.

The remainder is organized as follows. Section 2 reviews related work, focussing on retraining-based active learning algorithms. The proposed method is presented in detail in Section 3, followed by an extension of the proposed method to multiclass classification problems in Section 4. Sections 5 and6 report the experimental results on binary and multi-class classification problems, respectively. Finally, we conclude this paper in Section 7.

## 2. Related work

In the past decades, various active learning algorithms, based on many different selection criteria, have been proposed. These approaches rely on different heuristics. We can roughly divide these heuristics into two categories: informativeness and representativeness. Informativeness estimates the ability of an instance in decreasing the uncertainty of a statistical model, while representativeness indicates whether a sample is representative of the underlying distribution [46]. For example, query-by-committee [49], uncertainty sampling [27,51,55], error reduction [15,40], model change [2,13,20,48], expected variance reduction [44] belong to the informativeness category, but each of them has its own criterion of informativeness. Clustering-based approaches [36,43,58] and variance minimization methods [18,32,33,61] are included in the representativeness group. There are also methods that try to combine the two criteria, such as min-max view active learning [17], density or diversity weighted methods [1,30,47,60,64] and multi-criteria fusion [7,52,54,56].

The framework of retraining-based active learning, which our method is also an instantiation of, was first proposed by Roy and Mccallum [40] to perform so-called expected error reduction (EER for short). Tong and Koller [51] used a retraining approach in combination with SVMs to find instances that, after labeling, approximately halve the version space. A series of active learning methods which propose a scheme similar to EER, but with somewhat different motivations, were put forward in [8,13,15,44]. All in all, retraining-based active learners can be roughly divided into four categories: error reduction [15,40], variance reduction [44], model change [2,13,21,47], and min-max view active learning [16,17]. The principal difference among the above methods lies in how they measure the usefulness of unlabeled samples after retraining the model. For example, error reduce methods like EER [40] attempt to estimate the future generalization error as an indicator of the value of an instance while variance reduction approach [44] turns to use the model variance as a measure of the informativeness. Similarly, model change algorithms seek various ways of defining such change, *e.g.* as gradient length [47], and choose the instance which leads to maximum change. The min-max view active learning directly measures the value of objective function during retraining procedure and selects the instance with minimum score in the worst case scenario. Recently, Yang and Loog [59] proposed to improve the retraining-based algorithms by integrating the uncertainty information in the selection criterion.

We finally note that there exist close relationships between the proposed method and various active learning techniques, such as query-by-committee (QBC) [49], and variance minimization [18,33,61]. Their connections will be particularly explained in Section 3.5.

## 3. Maximizing variance for active learning

We give a detailed description of the proposed method. We provide the full algorithm and introduce what is at the core of our method: so-called retraining information matrices (RIMs). Based on these RIMs, we introduce the two main types of variance and describe how these are fused into a single criterion for instance selection. In all of this, we focus on probabilistic classifiers. In Section 3.4, we show one way to adapt our method to a non-probabilistic classifier that does not directly provide a posterior probability estimate. We particularly focus on the SVM, which is the classifiers we are going to experiment with next to logistic regression. In Section 3.5, we analysis the connections of the proposed method and several existing active learning approaches. First however, we spend a few words on the specific active learning setting we consider.

### *3.1. Specific setting*

We study pool-based active learning in which the selection of individual instances to be labeled is sequential and myopic. This means that we assume we already have a large pool of unlabeled data with a small number of labeled data, and a single sample is selected for labeling at a time [46]. We start with the binary classification problem, then present how to extend the proposed method to multiclass tasks in the following section, Section 4. We take $\mathcal{U}$ to be the pool of $n$ unlabeled instances $\{x_i\}_{i=1}^{n}$ and $\mathcal{L}$ to be the already labeled training set, where $y_i = \{+1, -1\}$ is the class label of $x_i$. $P_{\mathcal{L}}(y|x)$ represents the conditional probability of $y$ given $x$ on the basis of a classifier trained on $\mathcal{L}$.

### *3.2. Retraining information matrices*

Fig. 1 gives a pictorial overview of the proposed method. The proposed method can be used with different types of classifiers. In addition, Algorithm 1 summarizes the overall training procedure of MVAL for probabilistic classifiers. The proposed method generates two matrices $\mathcal{P}$, $\mathcal{N}$, with the purpose of recording the probability of all unlabeled instances after each retraining procedure. We first assume that the next queried instance is labeled as $+1$, we then extend the current labeled set $\mathcal{L}^+ = \mathcal{L} \cup \{x_i, +1\}$, retrain the classifier on $\mathcal{L}^+$, and calculate the conditional probability $P_{\mathcal{L}^+}(+1|x_j)$ for all $x_j \in \mathcal{U}$. Each $x_i \in \mathcal{U}$ is used to retrain the model, resulting in a matrix $\mathcal{P}$ of size $n \times n$, where each element $(i, j)$ in $\mathcal{P}$ is assigned $P_{\mathcal{L}^+}(+1|x_j)$. For example, assuming that $\mathcal{U}$ consists of six unlabeled samples $x_i, i = 1, 2, \ldots 6$, we could get the matrix $\mathcal{P}$ in Fig. 1a. Equivalently, if we categorize all of the next queried instances as $-1$, we retrain the model with $\mathcal{L}^+ = \mathcal{L} \cup \{x_i, -1\}$ for all $x_i \in \mathcal{U}$, we can construct a matrix $\mathcal{N}$ that contains the elements $\mathcal{N}_{i,j} = P_{\mathcal{L}^+}(+1|x_j)$, of which an example is shown in Fig. 1b.

**Algorithm 1** Maximizing Variance for Active Learning

```
Input: Labeled data $\mathcal{L}$, unlabeled data $\mathcal{U}$
repeat
    Train on $\mathcal{L}$ and calculate entropy $e_j$ for all $x_j \in \mathcal{U}$;
    For each $x_i \in \mathcal{U}$, retrain on $\mathcal{L}^+ = \mathcal{L} \cup \{x_i, +1\}$, let $\mathcal{P}_{i,j} = P_{\mathcal{L}^+}(+1|x_j)$, $x_j \in \mathcal{U}$;
    For each $x_i \in \mathcal{U}$, retrain on $\mathcal{L}^+ = \mathcal{L} \cup \{x_i, -1\}$, let $\mathcal{N}_{i,j} = P_{\mathcal{L}^+}(+1|x_j)$, $x_j \in \mathcal{U}$;
    Obtain weighted $\hat{\mathcal{P}}$ and $\hat{\mathcal{N}}$ and compute the variance using Eq. 3;
    Query the instance $x^*$ with maximum variance and label it $y^*$, update $\mathcal{L} \leftarrow \mathcal{L} \cup \{x^*, y^*\}, \mathcal{U} \leftarrow \mathcal{U} \backslash \{x^*\}$;
until Stopping criterion is satisfied
```