Pattern Recognition 83 (2018) 401–415

Contents lists available at ScienceDirect

Pattern Recognition

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

A benchmark and comparison of active learning for logistic regression

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

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

ARTICLE INFO

Article history:
Received 1 October 2017
Revised 19 April 2018
Accepted 6 June 2018

Keywords:
Active learning
Logistic regression
Experimental design
Benchmark
Preference maps

ABSTRACT

Logistic regression is by far the most widely used classifier in real-world applications. In this paper, we benchmark the state-of-the-art active learning methods for logistic regression and discuss and illustrate their underlying characteristics. Experiments are carried out on three synthetic datasets and 44 real-world datasets, providing insight into the behaviors of these active learning methods with respect to the area of the learning curve (which plots classification accuracy as a function of the number of queried examples) and their computational costs. Surprisingly, one of the earliest and simplest suggested active learning methods, i.e., uncertainty sampling, performs exceptionally well overall. Another remarkable finding is that random sampling, which is the rudimentary baseline to improve upon, is not overwhelmed by individual active learning techniques in many cases.

1. Introduction

In practice, it is easy to acquire a large amount of data, yet difficult, time-consuming, and expensive to label data since human experts are usually involved [65]. For instance, collecting millions of images from Google is not that difficult, while categorizing these images may need a lot of manpower and other resources. Active learning addresses this challenge by selecting the most valuable subset from the whole data set for human annotation. Many research studies have demonstrated that active learning is effective in maintaining good performance while reducing the overall labeling effort over a diverse range of applications, such as text categorization [5,73], medical image classification [34,62], remote sensing [18,63,74], image retrieval [13,53,80] and natural language processing [71].

To choose the most informative subset, it is of vital importance to choose an appropriate criterion which measures the usefulness of unlabeled instances. Most commonly used criteria in active learning include query-by-committee [69], uncertainty sampling [73], expected error minimization [30,37,61], and variance reduction [64,79,81], variance maximization [77], maximum model change [6,24,42,68] and the "min-max" view active learning [35,38]. They are derived from diverse heuristics and classifier dependent. Some of them are specifically designed for one particular classifier, *e.g.* the simple margin criterion for support vector machines [73], while others can be adapted to different types of classifiers, *e.g.* expected error reduction for logistic regression and naive Bayes [61].

In this work, we benchmark the state-of-the-art active learning algorithms built on logistic regression. Logistic regression is chosen because it is the most widely applied classifier in general and especially outside of machine learning in the applied sciences.[1] In addition, it is also used by most active learners (see, for instance, [17,28,30,31,34,36,43,44,52,64]). In part, the latter is because logistic regression readily provides an estimate of the posterior class probability, which is often exploited in active learning. In the binary classification setting, logistic regression models a posterior probability $P(y_i|x_i) = 1/(1+\exp^{-y_i w^T x_i})$, where $x_i \in \mathbb{R}^d$ is a training feature vector labeled with $y_i \in \{+1, -1\}$ and w is the d-dimensional parameter vector that is determined at training time. During training, we minimize the log-likelihood of the training data $\mathcal{L}$ to learn the model parameter w as follows:

$$\min_w \frac{\lambda}{2}\|w\|^2 + \sum_{x_i \in \mathcal{L}} \log(1+\exp^{-y_i w^T x_i}) \tag{1}$$

where $\|w\|^2$ is a regularization term for which λ controls its influence.

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

[1] An advanced search on www.nature.com on October 1, 2017, gives us, for example, 1,126 hits for "support vector machine", 6,182 for "nearest neighbor" (containing more hits than just to the classifier), 1,231 for "LDA", and 14,715 for "logistic regression". Other classifiers are retrieved even less often.

https://doi.org/10.1016/j.patcog.2018.06.004
0031-3203/

All in all, we study six different categories of active learning algorithms in which nine active learners are compared in an extensive benchmark study. Our work differs from two relevant earlier surveys on active learning [25,65] in two important respects: (1) our work constructs extensive experiments to investigate the empirical behaviors of these active learning algorithms while these two surveys do not compare the performance of different methods; (2) our paper presents a detailed summary of the active learning algorithms on the basis of logistic regression classifier because of its popularity while these two surveys offer an overview of existing active learning algorithms without specifying a type of classifiers. We believe that an empirical comparison can lead to a better understanding of the characteristics of active learning algorithms and provide guidance to the practitioner to choose a proper active learning algorithm. We should also mention the work by Schein and Ungar [64] here, that already provided an evaluation of active learning methods using logistic regression. In this paper, however, we compare some new methods, which appeared only recently [6,38,50], and we generally provide a fair and comprehensive comparison with much more extensively conducted experiments. We also investigate how active learning algorithms generally perform in comparison to random sampling, and point out the underlying relationships among the compared methods. The computational cost of each method is also evaluated.

In this paper, we focus on the pool-based setting, where few labeled samples and a large pool of unlabeled samples are available [65]. We consider the myopic active learning which assumes that a single unlabeled instance is queried at a time. Batch mode active learning, which selects a batch of examples simultaneously, is not considered in this work and we refer to [8–10,12,31,34] for further background of typical approaches.

The main contributions of this work can be summarized as follows:

1. A review of the state-of-the-art active learning algorithms built on logistic regression is presented, in which links and relationships between methods are explicated;
2. A preference map is proposed to reveal characteristic similarities and differences of the selection locations in 2D problems;
3. Extensive experiments on 44 real-world datasets and three artificial sets are carried out;
4. Insight is provided for the behaviors of classification performance and computational cost.

1.1. Outline

The remainder of the paper is organized as follows. Section 2 describes the general procedure of active learning and reviews the various approaches to active learning built on logistic regression. At the same time it sketches the relationships among different methods. Extensive experimental results on synthetic and real-world datasets are given in Section 3. The experimental setup is described and the outcomes are reported. More importantly, it provides an extensive discussion of the findings and aims to critically evaluate these compared methods. Section 4 concludes our work.

2. Active learning strategies and methods

For myopic active learning in the pool-based scenario, we assume that a small set of labeled instances with a large pool of unlabeled samples are available. Let $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{l}$ represent the training data set that consists of l labeled instances and let $\mathcal{U}$ be the pool of unlabeled instances $\{x_i\}_{i=l+1}^{n}$. Each $x_i \in \mathbb{R}^d$ is a d-dimensional feature vector and $y_i \in C$ is the class label of x_i. In this work we restrict ourselves to binary classification, which does not pose any essential limitation. For this reason, C is simply taken to be the set $\{+1, -1\}$. The active learner will select an instance x^* from the unlabeled pool based on its measure of utility, obtain the corresponding label y^* by manual annotation and extend the training set with the new labeled sample $\mathcal{L} = \mathcal{L} \cup (x^*, y^*)$. The whole procedure is repeated until some stopping criteria are satisfied.

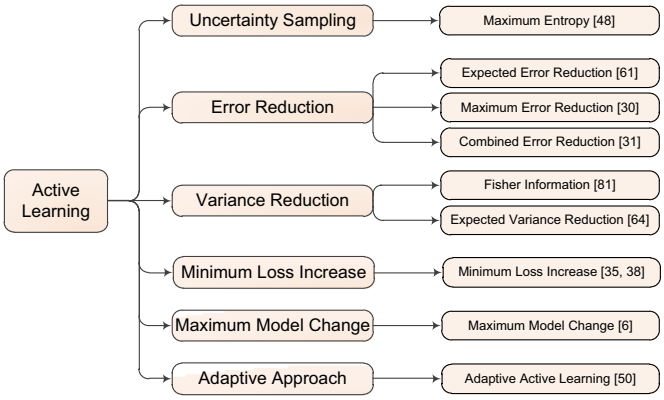

Fig. 1. Nine active learners from six different categories are used in our comparison.

The remaining part of this section presents six different categories of active learning algorithms built on logistic regression, i.e., uncertainty sampling, error reduction, variance reduction, minimum loss increase, maximum model change and an adaptive approach, one per subsection. As also shown in Fig. 1, nine different active learners which relate to the above six categories are used in our benchmark and comparison.

2.1. Uncertainty sampling

Uncertainty sampling, which selects the instances for which the current classifier is least certain, is a widely used active learning method [48,65]. Querying these least certain instances can help the model refine the decision boundary. Intuitively, the distances between unlabeled instances and the decision boundary can be measures of the uncertainty. Tong and Koller [73] proposed a simple margin approach which queries the instance closest to the decision boundary.

Entropy is a different and more widely used general measure of uncertainty [70]. Entropy-based approaches query the instances with *maximum entropy*:

$$x^* = \arg\max_{x \in \mathcal{U}} -\sum_{y \in C} P_{\mathcal{L}}(y|x) \log P_{\mathcal{L}}(y|x) \tag{2}$$

where $P_{\mathcal{L}}(y|x)$ is the conditional probability of y given x according to a logistic classifier trained on $\mathcal{L}$. This method is called ENTROPY for short. It calculates the entropy of each $x \in \mathcal{U}$ and selects the instance x^* which has maximum entropy. It can be used with any classifier that produces probabilistic outputs. For binary classification, ENTROPY is equivalent to the simple margin approach [73].

One of the main risks of such uncertainty sampling based approaches lies in the fact that, due to a lack of exploration, they can get stuck at suboptimal solutions, continuously selecting instances which do not improve the current classifier at all [38].

2.2. Error reduction

Error reduction approaches are another type of popular active learning methods [30,31,37,61]. These approaches attempt to measure how much the generalization error is likely to be reduced when adding one new instance into the labeled dataset. Though one does not have direct access to the future test data, Roy and