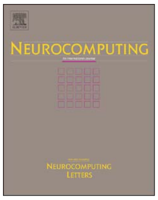
Neurocomputing 222 (2017) 183–190

Contents lists available at ScienceDirect

Neurocomputing

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



# Active learning for penalized logistic regression via sequential experimental design

Jing Wang, Eunsik Park*

*Department of Statistics, Chonnam National University, Republic of Korea*

ARTICLE INFO

Communicated by D. Wang

*Keywords:*
Active learning
Logistic regression
Experimental design

ABSTRACT

Penalized logistic regression is useful for classification that not only provides class probability estimates but also can overcome overfitting problem. Traditionally, supervised classifier learning has required a lot of labeled data. Due to technical innovation, it is easy to collect large amounts of unlabeled data, while labeling is usually expensive and difficult. Active learning aims to select the most informative subjects for labeling to decrease the amount of labeling requests. Recently, active learning using experimental design techniques have attracted considerable attention. The typical criteria attempt to reduce the generalization error of a model by minimizing either its estimation variance or estimation bias. However, they fail to take into account both components simultaneously. In this article, we introduce a new algorithm of active learning using penalized logistic regression. The most informative subjects are selected as those with the smallest mean squared estimation error. This criterion, integrated with the idea of sequential design, is exploited in our algorithms to guide a procedure for a new subject selection. Experiments on extensive real-world data sets demonstrate the effectiveness and efficiency of the proposed method compared to several state-of-the-art active-learning alternatives.

## 1. Introduction

Supervised classifier learning is a common and important task in many fields, including biomedical research, engineering and many others. In a standard setup, subjects with class labels are provided for training. Nowadays, due to technical innovation, one often easily encounters data sets that have large amounts of unlabeled data. Since labeling each of them is inefficient and usually costly, it is crucial to find a way to decide which subjects are labeled and included into the training set so as to achieve a good classifier within a reasonable number of labeling requests. An effective approach is active learning [1–3] among machine learning applications [4–6], whereby the learner actively selects the most informative subjects and requests their labels.

One category of active-learning methods is based on support vector machines (SVMs), which have performed excellently in many real-world classification tasks [7,8]. A standard query strategy for this type of learning process is to select the data points that lie closest to the SVM's dividing hyperplane [9,10]. The classification decision then is made according to the sign of the decision function's output. One weakness of the SVM classifier is that it only addresses the class membership estimation for a given subject. Many applications, however, require more than classification [11,12]. In particular, accurate class-probability estimation plays much more important role. For instance [13], in targeted marketing, the estimated probability that a customer will respond to an offer is combined with the estimated profit to evaluate various offer propositions. Logistic regression [14] is a simple and efficient probability-based classifier. However, as noted in [15], maximum likelihood (ML) estimation for logistic regression can exhibit overfitting when the number of unknown parameters is large compared to the number of observations. In addition, for linearly separable data sets, some ML estimates will be infinite, thus assigning an unreasonable class probability, either zero or one, to every subject [16]. Using enalized logistic regression (PLR) [17] may overcome these problems. Zhu and Hastie [17] have shown that PLR performs comparably to SVM.

Besides, many other active-learning approaches have been developed on the basis of experimental design techniques [18]. Cohn [19] proposed an active-learning approach for neural networks, selecting queries that reduced the error by minimizing its estimated variance while ignoring its bias or assuming that the bias was approximately zero. Results in [19] demonstrated that minimizing variance alone was not enough. Cohn [20] introduced a bias minimization criterion based on locally weighted regression and showed empirically that this algorithm outperformed the common variance-minimization approach.

* Corresponding author.
*E-mail address:* espark02@jnu.ac.kr (E. Park).

http://dx.doi.org/10.1016/j.neucom.2016.10.013
Received 26 March 2016; Received in revised form 31 August 2016; Accepted 9 October 2016
Available online 24 October 2016

**Table 1**
Description of UCI benchmarks.

| Data sets | Features | Size | Class |
|---|---|---|---|
| spect | 22 | 267 | 2 |
| wdbc | 30 | 569 | 2 |
| diabetes | 8 | 768 | 2 |
| biodegradation | 41 | 1055 | 2 |
| optdigits3vs5 | 64 | 1130 | 2 |
| spambase | 57 | 4601 | 2 |

In these works, queries were generated while we were interested more in identifying informative examples from an existing pool of available unlabeled examples. Le Ly and Lipson [21] presented an active-learning method, based on Shannon information criterion, and used it for model disambiguation. Similarly, Li et al. [22] introduced manifold optimal experimental design via dependence maximization, selecting subjects that minimized the variance of the model parameters. Instead of focusing on variance, Pauwels et al. [23] proposed a criterion that attempted to control the contribution of both variance and bias jointly to the error of parameter estimation. These optimization criteria on parameters, however, did not directly characterize prediction quality. Using linear regression models, Yu et al. [24] developed transductive experimental design, minimizing the average predictive variance over a pre-given set, and used SVM as the base classifier. Schein and Ungar [25] re-derived the variance-reduction principle for logistic regression.

Another category is heuristic active learning. A simple strategy is uncertainty sampling, choosing subjects for which the learner has the lowest certainty [26]. Query by committee (QBC) is to select subjects that maximize disagreement among an ensemble of hypotheses [27]. MacKay [28] and Roy and McCallum [29] have studied active-learning algorithms for probabilistic models, called classifier certainty methods, which minimize the entropy of their predictions. Based on logistic regression with ML estimation, Schein and Ungar [25] evaluated the variance-based approach compared to five variations of the above mentioned heuristic techniques. They concluded that among the competitors, the proposed experimental design method is most likely to match or beat random sampling, though time consumingly, whereas the heuristic alternatives are computationally efficient but fail to beat random sampling on some portions of the evaluation.

In this paper, we propose a new active-learning algorithm built on penalized logistic regression. Our algorithm is fundamentally based on experimental design. Beyond the common variance-only or bias-only minimization criteria, we identify particularly informative subjects as the ones that minimize the generalization error of a model, which takes into account both the bias and the variance. Since designs for generalized linear models depend on the unknown parameters of the models, a sequential strategy commonly used in sequential designs in statistics is exploited in the subject-selection stage during the learning process. We show empirically that the proposed method is effective and computationally efficient compared to several state-of-the-art active-learning

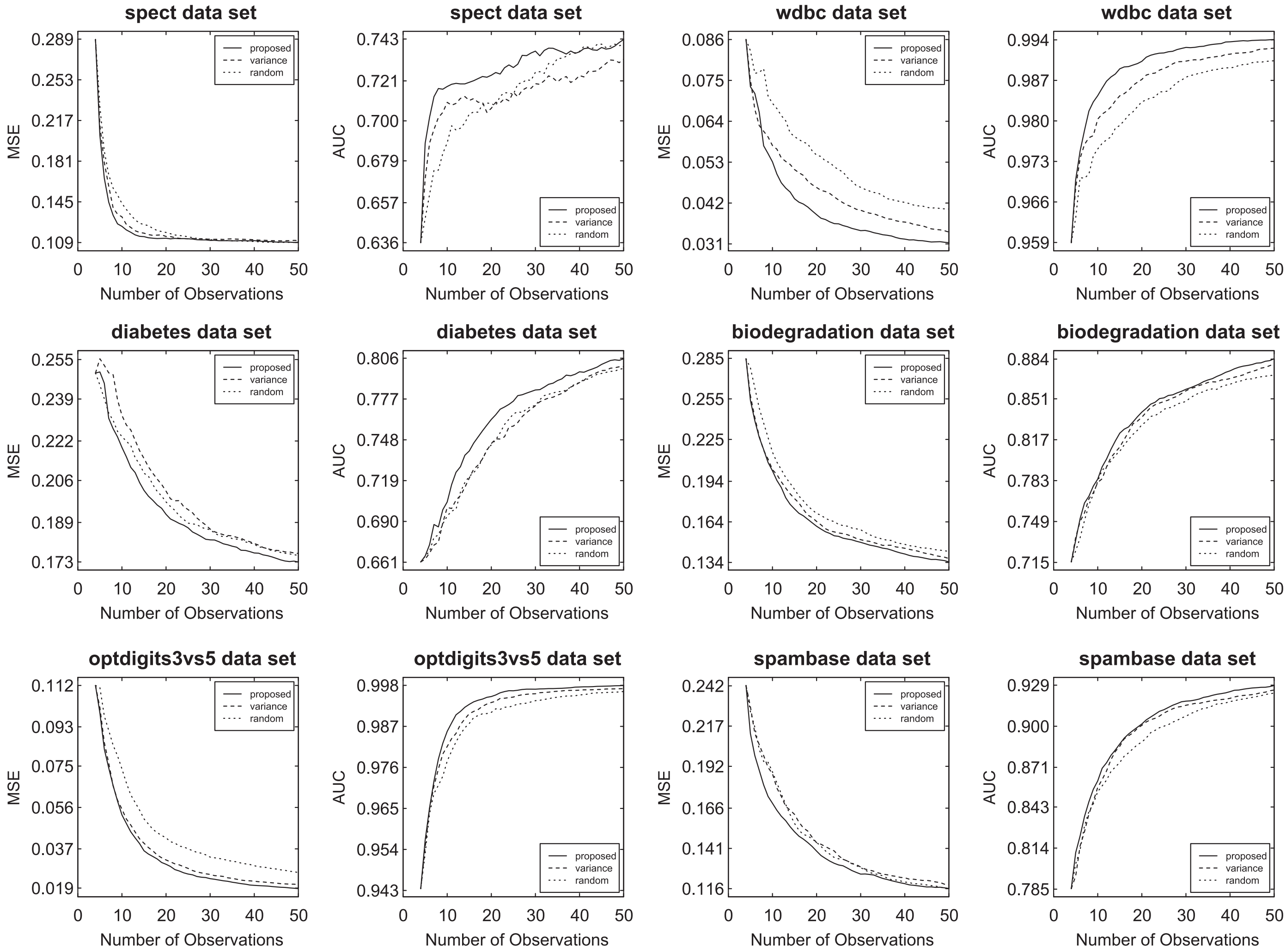


**Fig. 1.** Learning curves of the proposed method compared to variance-based active learning and random sampling in terms of mean MSE and AUC on UCI benchmarks.