Neurocomputing 167 (2015) 122–131

Contents lists available at ScienceDirect

Neurocomputing

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

Gaussian process versus margin sampling active learning

Jin Zhou, Shiliang Sun *

Shanghai Key Laboratory of Multidimensional Information Processing, Department of Computer Science and Technology, East China Normal University, 500 Dongchuan Road, Shanghai 200241, PR China

ARTICLE INFO

Article history:
Received 3 February 2015
Received in revised form
23 March 2015
Accepted 28 April 2015
Communicated by Wei Wu
Available online 14 May 2015

Keywords:
Active learning
Gaussian process
Margin sampling
Support vector machine
Manifold-preserving graph reduction

ABSTRACT

There are a large number of unlabeled examples in real-world application, and if the labels of these unlabeled examples are given manually, then the cost will be very high. The problem about how to label these massive unlabeled instances with the minimal cost is paid more and more attention. Active learning efficiently solves this bottleneck by selecting the most informative examples from the unlabeled examples and establishing a classifier with a higher classifier accuracy to label unlabeled examples, which greatly improves work efficiency. In this paper, we compare two kinds of traditional active learning algorithms relying on a single classifier, namely Gaussian process and margin sampling active learning, in two aspects of classification error rates and computing time. Moreover, we compare their improved versions (GPMAL and IMS) which apply the manifold-preserving graph reduction (MPGR) algorithm. MPGR constructs a subset which well exploits the structural spatial connectivity and spatial diversity among examples. By using MPGR, an active learner selects the informative and representative candidates from the subset instead of the whole unlabeled data set. In addition, a comparison with a state-of-the-art active learning method, QUIRE, is provided. Experimental results on multiple data sets show that both GPMAL and IMS have their own advantages.

1. Introduction

For any supervised classifier to perform well, it not only needs to be trained on sufficient labeled examples but also requires the quality of the labeled data to be high. Therefore, the training set should be carefully selected to lead the classifier to obtain good performance. This constraint makes collecting training data a potentially important process. However, in many sophisticated supervised learning problems, a large number of unlabeled examples can be obtained easily, but labeling them are generally costly and time-consuming. In order to reduce the difficulties as much as possible, it is imperative for procedures to find an appropriate training set which really helps us to improve the performance of the classifier automatically, or semi-automatically.

In the machine learning area, this problem is known as active learning, which has been the subject of significant theoretical and experimental studies in machine learning [18]. As stated in [18], there are mainly three scenarios where the learner may be able to ask queries, which are respectively membership query synthesis, stream-based selective sampling and pool-based sampling. In this paper, we focus on the pool-based active learning, which selects an unlabeled example from a given pool for manual labeling. Starting with a small training set, active learning can be seen as a sampling process by following some criterion to actively select and label those examples which improve the performance of the classifiers during the iteration from the unlabeled example pool. Instead of randomly picking unlabeled examples, it selects the examples that are considered the most informative for human labeling and then updates the classification model by incorporating them into the existing labeled set [20]. Through this process, a predictor trained on a small number of well-chosen examples can perform comparably to a predictor trained on a large set of randomly chosen examples [3,11,25].

From the algorithm perspective, active learning methods may be grouped into three classes [22]. The first class of active learning methods is based on uncertainty sampling which relies on the estimation of the posterior probability distribution function of the classes [9,16]. It selects the examples which are the most uncertain according to the values of their posterior probabilities. For a binary problem, the selected examples are the ones which give the class membership probability closest to 0.5. In this paper, we use Gaussian processes (GP) to provide a probabilistic prediction estimate of uncertainty. In addition, active learning of Gaussian processes (GPAL) has been successfully used for object categorization [7,8]. The second class of active learning methods is based on large margin-based heuristics which utilize the geometrical features of SVM [1,13] and has been widely used in many practical applications, such as species recognition [10], text mining [19] and remote sensing image retrieval [21]. For example, the margin sampling (MS) strategy selects for

* Corresponding author. Tel.: +86 21 54345186; fax: +86 21 54345119.
E-mail address: slsun@cs.ecnu.edu.cn (S. Sun).

http://dx.doi.org/10.1016/j.neucom.2015.04.086
0925-2312/

labeling the candidates lying within the margin of the current SVM since these examples are the most likely to become new support vectors [1,17]. The efficiency and robustness of this method have been discussed and proved in [12,23]. The last class of active learning is committee-based heuristics [4,28]. The committee members are composed of several different classifiers, which are trained to label the unlabeled examples. The task is to select the candidates where the disagreement between the classifiers is maximal.

From the three classes of active learning methods, we can see that the first one and second one are based on a single classifier, while the third one is based on multiple classifiers. Although some active learning methods have been proposed for selecting unlabeled examples for tagging, comparison among the different types of active learning mechanism can rarely be found in the active learning literature. This paper aims to address this issue. In this paper, we compare the effectiveness of the first two active learning methods (take MS and GPML as the representative) since they both rely on the single classifier.

Moreover, informativeness and representativeness are two main criteria which are widely used for active query selection [6]. Informativeness measures if an examples can reduce the uncertainty of a statical model, while representativeness measures the ability of an example in representing the overall input patterns of all the unlabeled examples [18]. However, most active learning methods only consider one of the two criteria when selecting unlabeled data, which greatly limit the availability of active learning. For example, both of GPAL and MS only deploy the informativeness criterion. In the process of active example selection, we usually select the most informative examples from all unlabeled data without considering the structural information and spatial diversity among them. This will lead to a result that in the same area there are more than one point to be selected, and thus it is possible to produce redundancy which can decrease the classification accuracy. To overcome this shortcoming, recently two improved versions of MS and GPAL were introduced, which are called GPMAL [26] and IMS [27], respectively. GPMAL and IMS combine the informativeness and representativeness criteria by applying an algorithm called manifold-preserving graph reduction (MPGR). By using MPGR, one can construct a subset which represents the global structure of the original distribution of samples. Such a modification of GPAL and MS can avoid oversampling on dense regions to a large extent.

The main contribution of this paper is that we compare the two classes of classical active learning methods (GPML and MS) and their improved versions (GPMAL and IMS) which are based on a single classifier. Besides, we introduce a state-of-the-art active learning approach which also combines the informativeness and representativeness of an examples, termed QUIRE [6]. We compare them in the aspects of classification performance and time cost. In this paper, the SVM classifier has been used to provide the comparison for all active learning methods.

This paper is organized as follows: In Section 2, we briefly review some background about GPAL and MS. In Section 3, we describe the two improved versions which apply MPGR to the original active learning methods. In Section 4, we show the experimental results on multiple datasets to present the performance and time comparison of all methods. Finally, we provide concluding remarks in Section 5.

2. Background

In this section, we briefly review some basic knowledge related to active learning of Gaussian process and margin sampling.

2.1. Active learning algorithms

Suppose we have a training set X composed by m labeled examples, i.e., $X=\{(x_1,y_1),\cdots(x_m,y_m)\}$ $(x_i\in R^d)$ with the corresponding labels $\mathbf{y}=[y_1,y_2,\ldots y_m]^\top$ $(y_i\in\{\pm 1\})$. We wish to add to the training set a series of examples from a set of n unlabeled points $U=\{u_1,u_2,\cdots u_n\}$, with $n\gg m$. X and U have the same features. Instead of choosing unlabeled examples randomly, active learning selects valuable examples according to a problem-oriented heuristic which aims at maximizing the performances of the classifiers. Generally speaking, active learning is a process of guiding the sampling process by actively selecting and labeling the most informative or representative candidates from a large pool of unlabeled examples, which can effectively reduce the size of the training set and simultaneously improve the performance of the model. For simplicity, in this paper we mainly consider the two-class problem. As to multi-class classification, we use one-vs-rest to convert the multi-class problem to multiple two-class problems.

2.2. Active learning of Gaussian processes

2.2.1. Gaussian processes

Here we first briefly summarize Gaussian processes to facilitate the subsequent introduction of GPAL.

As mentioned above, Gaussian processes provide probabilistic prediction estimates and thus are well-suited for active learning. A Gaussian process is a stochastic process specified by its mean and covariance function [15]. Given a data set with m examples $X=\{x_1,x_2,\ldots x_m\}$, the corresponding class labels are $\mathbf{y}=[y_1,y_2,\ldots y_m]^\top$ and latent variables are $\mathbf{t}=\{t_1,t_2,\ldots t_m\}$.

The prior distribution defines the probabilistic relationship between the examples X and the latent variables $\mathbf{t}$. After putting a Gaussian processes prior, the prior distribution for the latent variables is assumed to be Gaussian:

$$p(\mathbf{t}|X,\theta)=N(\mathbf{t}|\mathbf{0},K) \tag{1}$$

with a zero mean and a covariance matrix K. K is a kernel matrix which is parameterized by the hyperparameter θ.

The likelihood models the probabilistic relationship between the label $\mathbf{y}$ and the latent variable $\mathbf{t}$. In this work we assume that $\mathbf{y}$ and $\mathbf{t}$ are related via a Gaussian noise model. The Gaussian observation likelihood is

$$p(\mathbf{y}|\mathbf{t})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{-\frac{(\mathbf{y}-\mathbf{t})^2}{2\sigma^2}} \tag{2}$$

where σ^2 is the noise model variance. Although the Gaussian noise model is originally developed for regression, it has also been proved effective for classification, and its performance typically is comparable to the more complex probit and logit likelihood models used in classification problems [7]. For its simplicity and a closed form solution for iterations, we use the Gaussian noise model in our experiments. The joint likelihood can be written as

$$p(\mathbf{y},\mathbf{t}|X,\theta)=p(\mathbf{t}|X,\theta)p(\mathbf{y}|\mathbf{t}) \tag{3}$$

After integrating out the latent variables, the marginal likelihood will be

$$P(\mathbf{y}|X)=N(\mathbf{y}|\mathbf{0},K+\sigma^2 I). \tag{4}$$

The prediction distribution for the label y_u at a new point x_u is also a Gaussian:

$$P(y_u|X,\mathbf{y},x_u)\sim N(Y_u,\Sigma_u), \tag{5}$$

where

$$Y_u=k_*^\top(\sigma^2 I+K)^{-1}\mathbf{y}, \tag{6}$$

$$\Sigma_u=\tilde{k}(x_u,x_u)-k_*^\top(\sigma^2 I+K)^{-1}k_*^\top, \tag{7}$$

Here, $\tilde{k}$ is the covariance function and k_* is the vector of covariances between x_u and the training data, which is given by $k_*=[k(x_u,x_1),\ldots,k(x_u,x_m)]^\top$. Moreover, since y_u and t_u are linked by the Gaussian noise model, the predictive distribution over the