Neurocomputing 71 (2008) 2489–2498

Contents lists available at ScienceDirect

Neurocomputing

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

# Semi-supervised and active learning with the probabilistic RBF classifier

Constantinos Constantinopoulos *,[1], Aristidis Likas

*Department of Computer Science, University of Ioannina, GR 45110 Ioannina, Greece*

ARTICLE INFO

Available online 7 May 2008

*Keywords:*
Probabilistic RBF network
Active learning
Semi-supervised learning
EM algorithm
Unlabeled data

ABSTRACT

The probabilistic RBF network (PRBF) is a special case of the RBF network and constitutes a generalization of the Gaussian mixture model. In this paper we propose a semi-supervised learning method for PRBF, using labeled and unlabeled observations concurrently, that is based on the expectation–maximization (EM) algorithm. Next we utilize this method in order to implement an incremental active learning method. At each iteration of active learning, we apply the semi-supervised method, and then we employ a criterion to select an unlabeled observation and query its label. This criterion identifies points near the decision boundary. In order to assess the effectiveness of our method, we propose an adaptation of the well-known Query by Committee (QBC) algorithm for the active learning of the PBFR, and present experimental comparisons on several data sets that indicate the effectiveness of the proposed method.



## 1. Introduction

The active learning of a classifier constitutes a special learning problem, where the training data are available as a stream of classified observations and are *actively collected* by the classifier during training. At each learning iteration, the learning system determines regions of interest in the data space, asks for labeled training data that lie in these regions, and exploits the acquired class labels to improve its classification performance.

The importance of active learning is well established, see [3] for a study on the increase of classifier's accuracy as the number of labeled data increases. Various active learning methods have been suggested for almost all types of classifiers; in [5] a learning method for Gaussian mixture models [15] has been proposed, that selects data minimizing the variance of the learner. The Query by Committee (QBC) algorithm has been proposed in [17,9] for the active learning of a committee of classifiers, which picks those data for which the committee members disagree. Based on this selection method, in [13] they proposed the exploitation of available unlabeled data by employing the EM algorithm [8] to form a better selection criterion, that is used to train a naive Bayes classifier. In [25] Gaussian random fields and harmonic functions are trained, while data selection is based on the estimated expected classification error. In [23] an active learning methodology for support vector machine (SVM) classifiers has been proposed with applications to text classification.

In this work we focus on *pool-based* active learning which is well-studied active learning problem [13,23,25]. In this case a set of labeled and unlabeled observations is available right from the start. At each training iteration we are allowed to query the labels of unlabeled points, and use the acquired labels to improve the classifier (see Fig. 1). In practice this learning scenario is important in two cases: (a) when querying a field expert is expensive, as in medical diagnosis, and (b) when there is a huge quantity of unlabeled data and is difficult to manually characterize all of them, as for example in document classification and e-mail filtering [23,11].

An intuition behind pool-based learning is that the unlabeled data can be exploited to construct a more detailed generative model for the data set. Thus this problem is closely related to *semi-supervised* learning. Algorithms for semi-supervised learning have been proposed for Gaussian mixtures in [10,18], as well as for the RBF network [16]. So it has been established that unlabeled data reveal useful information for the distribution of the labeled data (see [4] for an informative recent survey book on semi-supervised learning).

In order to implement active learning the following issues must be addressed:

- Define an effective *criterion* for selecting the unlabeled points to query their label.
- Use an *incremental* training algorithm so that learning does not start from scratch at each iteration.

* Corresponding author. Tel.: +30 26510 98810; fax: +30 26510 98882.
*E-mail addresses:* ccostas@cs.uoi.gr (C. Constantinopoulos), arly@cs.uoi.gr (A. Likas).

[1] This research was co-funded by the European Union in the framework of the program "Heraklitos" of the "Operational Program for Education and Initial Vocational Training" of the 3rd Community Support Framework of the Hellenic Ministry of Education, funded by 25% from national sources and by 75% from the European Social Fund (ESF).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2007.11.039

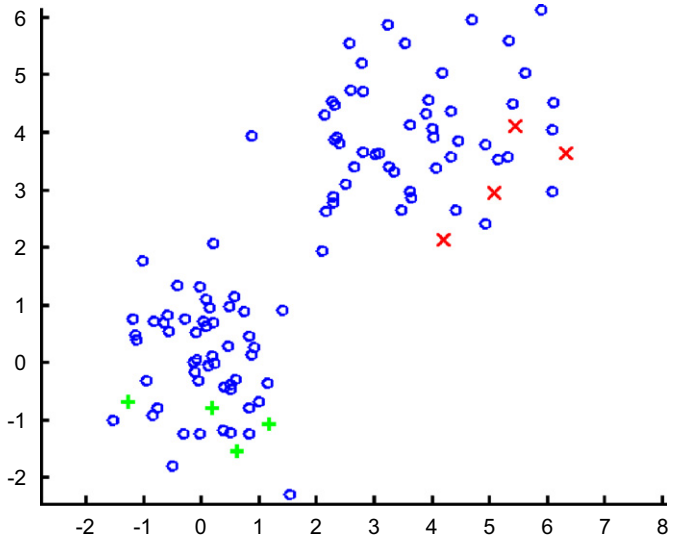

**Fig. 1.** Available data in a pool-based active learning scenario: "○" denotes unlabeled data, "+", "×" denote class labels. As learning proceeds "○" points are selected and their label ("+" or "×") is revealed.

- Exploit during training not only the labeled but also the unlabeled data (use *semi-supervised* learning training algorithms).

In this work we treat the problems of semi-supervised learning and pool-based active learning for the probabilistic RBF (PRBF) classifier [20,22]. This is a special case of the RBF network [2] that computes at each output unit the density function of a class. It adopts a cluster interpretation of the basis functions, where each cluster can generate observations of any class. This is a generalization of a Gaussian mixture model [15,2], where each cluster generates observations of only one class. In [6] an incremental learning method based on expectation–maximization (EM) for supervised learning is proposed that provides classification performance comparable to SVM classifiers.

The core of our work is the proposition of a *semi-supervised* learning method for the PRBF network, which is based on the EM algorithm for maximization of the *joint likelihood* of both the labeled and unlabeled data. Thus we obtain closed form update equations for the network parameters. We are facilitated by the fact that each node of the PRBF describes the local density of potentially all the classes. In order to handle the unlabeled data, it is possible to marginalize the class labels from the update equations of EM, thus using both labeled and unlabeled data in parameter estimation. Building on this method, we further present an incremental method for semi-supervised training. Exploiting the later method, we develop an *active* learning framework, by defining a suitable criterion for selecting the unlabeled observations, and asking for their label. Our contribution is concluded with the application of the QBC algorithm for the active learning of the PRBF, and its experimental comparison with the proposed algorithm.

In the following section we describe the PRBF network, and in Section 3 we present the semi-supervised training method based on the EM algorithm. In Section 4 we first propose an incremental semi-supervised training method, and in the following we propose an active learning algorithm. In the same section we also present the application of the QBC algorithm for the active learning of the PRBF. Section 5 provides the results from our experimental study, while the discussion in Section 6 concludes this work.

## 2. The PRBF classification network

Consider a classification problem with $K$ classes, where $K$ is known and each pattern belongs to only one class. We are given a training set $X=\{(x^n,y^n), n=1,\ldots,N\}$ where $x^{(n)}$ is a $d$-dimensional pattern, and $y^n$ is a label $k\in\{1,\ldots,K\}$ indicating the class of pattern $x^n$. The original set $X$ can be partitioned into $K$ independent subsets $X_k$, so that each subset contains only the data of the corresponding class. Let $N_k$ denote the number of patterns of class $k$, i.e. $N_k=|X_k|$.

The PRBF network has the same architecture as the typical RBF network, i.e. an input layer with $d$ units for the input vector $x=(x_1,\ldots,x_d)$, an output layer with $K$ units (one for each class) and a single hidden layer with an arbitrary number $M$ of component functions (hidden units), which are probability density functions. In the PRBF network all component density functions $p(x|j)$, $(j=1,\ldots,M)$ are utilized for estimating the conditional densities of all classes by considering the components as a common pool [19,20]. The $k$-th PRBF output models the class conditional density function $p(x|k)$ of class $k$ as a mixture model of the form

$$p(x|k)=\sum_{j=1}^{M}p(j|k)p(x|j),\quad k=1,\ldots,K, \tag{1}$$

where $p(x|j)$ denotes the component density $j$, while the mixing coefficient $p(j|k)$ represents the prior probability that a pattern has been generated from the density function of component $j$, given that it belongs to class $k$. The priors take positive values and satisfy the following constraint:

$$\sum_{j=1}^{M}p(j|k)=1,\quad k=1,\ldots,K. \tag{2}$$

Once the outputs $p(x|k)$ have been computed, the class of data point $x$ is determined using the Bayes rule, i.e. $x$ is assigned to the class with maximum posterior $p(k|x)$ computed by

$$p(k|x)=\frac{p(x|k)p(k)}{\sum_{\ell=1}^{K}p(x|l)p(\ell)}. \tag{3}$$

The class priors $p(k)$ are computed as $p(k)=N_k/N$, according to the maximum likelihood solution.

Also using Bayes theorem, the posterior probabilities $p(j|x,k)$ that component $j$ generated a pattern $x$ given its class $k$ can be easily computed:

$$p(j|x,k)=\frac{p(j|k)p(x|j)}{\sum_{i=1}^{M}p(i|k)p(x|i)}. \tag{4}$$

In the following, we assume Gaussian component densities of the general form:

$$p(x|j)=\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(x-\mu_j)\right\}, \tag{5}$$

where $\mu_j\in\Re^d$ represents the mean of component $j$, while $\Sigma_j$ represents the corresponding $d\times d$ covariance matrix. The whole adjustable parameter vector of the model consists of the mixing coefficients $p(j|k)$ and the component parameters (means $\mu_j$ and covariance matrices $\Sigma_j$) and we denote it by $\Theta$.

It is apparent that the PRBF model is a special case of the RBF network, where the outputs correspond to probability density functions and the second layer weights are constrained to represent the prior probabilities $p(j|k)$. Furthermore, the separate mixtures model [14] can be derived as a special case of PRBF, if each component $j$ generates only patterns of the class $\ell$, and we set $p(j|k)=0$ for all classes $k\neq\ell$.