Incremental One-Class Learning with Bounded Computational Complexity

Rowland R. Sillito and Robert B. Fisher
School of Informatics, University of Edinburgh, UK

Abstract. An incremental one-class learning algorithm is proposed for the purpose of outlier detection. Outliers are identified by estimating - and thresholding - the probability distribution of the training data. In the early stages of training a non-parametric estimate of the training data distribution is obtained using kernel density estimation. Once the number of training examples reaches the maximum computationally feasible limit for kernel density estimation, we treat the kernel density estimate as a maximally-complex Gaussian mixture model, and keep the model complexity constant by merging a pair of components for each new kernel added. This method is shown to outperform a current state-of-the-art incremental one-class learning algorithm (Incremental SVDD [5]) on a variety of datasets, while requiring only an upper limit on model complexity to be specified.

1 Introduction

The problem of one-class learning (also known interchangeably as “outlier / novelty / anomaly detection”) arises in a wide variety of different application domains. The fundamental goal of one-class learning is to generate a rule that distinguishes between examples of a known class of items and examples from previously-unseen novel classes, on the exclusive basis of training examples from the known class.

This problem presents itself in cases where one wishes to distinguish between members of a class for which examples are abundantly available, and members of another rarely observed class. This often arises when attempting to detect abnormal activity, eg. jet engine failure, computer network intrusions, disease symptoms, etc. In each of these domains, anomalous examples may be scarce or entirely absent during training, but their subsequent identification is of crucial importance. A wide variety of different methods have been proposed to address this problem (see [3] for a review). However, almost all existing one-class classification algorithms require all training examples to be available at once, for a single “batch” learning step: if a new example is presented, the classifier must be retrained from scratch.

Since outliers might only be identifiable by their deviation from a normal model, a key problem in one class learning is the choice of model complexity. In some cases training data may be well described by the parameters of a single Gaussian distribution, while in other cases - eg. where the data has multiple
modes or lies on a non-linear manifold - a more complex model is required. It is important to select the correct level of model complexity: if it is too low, the learned normal model may also include the anomalies that we wish to detect; if it is too high, the model may not include the majority of normal examples.

In many cases it would be useful to be able to incrementally train a classifier as data became available, without needing to pre-specify the level of model complexity. In this paper we propose a new technique for performing incremental one-class learning, where only an upper limit on model complexity needs to be specified. While computationally feasible, our algorithm attempts to estimate the underlying p.d.f. (probability density function) of the training data using non-parametric kernel density estimation (with Gaussian kernels), thereby generating a maximally complex one-component-per-example Gaussian mixture model. Once a maximum number of mixture components has been reached, it is kept constant by merging a pair of components for every new component added. We choose pairs of components for merging based on an information theoretic merging-cost function originally proposed by Goldberger and Roweis in [2].

Currently, at least two related techniques exist in the literature. In [9] Yamanishi et al. propose an unsupervised outlier detection procedure “SmartSifter” in which a Gaussian mixture model is trained using an on-line adaptation of the EM (Expectation Maximization) algorithm. A key aspect of their adaptation is the inclusion of “discounting” parameters which ensure that the effect of older training examples on the model parameters is rapidly displaced by new examples. Each new training example is given a score based on the extent to which it changes the model parameters: a comparatively high score, indicating a large change in model parameters, indicates the possibility of an outlier. This algorithm seems inappropriate for comparison with the proposed algorithm as it models a finite window of training data preceding each new example, rather than attempting to incrementally build a complete normal class description.

A more closely related algorithm has been proposed in [5], where Tax and Laskov present an incremental training procedure for the SVDD (Support Vector Data Description) algorithm originally proposed by Tax and Duin in [7]. The SVDD algorithm attempts to find the smallest hypersphere that encloses the training data, and allows more complex hyper-volumes (which may fit the data better) to be obtained by introducing kernel functions which map the training data to a higher dimensional space [7]. In both batch and incremental forms, the SVDD algorithm relies crucially on the correct choice of model complexity parameters. Various methods have been proposed to address this issue in the absence of example outliers, including: procedures for generating synthetic outliers (Tax and Duin [14]) and, more recently, a consistency based approach which takes the simplest possible classifier and increases its complexity parameter until the proportion of correctly recognized training data starts to fall (Tax and Muller [13]). The incremental variant of SVDD does not include any on-line model complexity selection, but it is conceivable that the batch optimization methods could be applied to a pre-existing dataset to optimally parametrize the classifier before using it for on-line training.