Neurocomputing 160 (2015) 93–107

Contents lists available at ScienceDirect

Neurocomputing

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

Making risk minimization tolerant to label noise

Aritra Ghosh [a], Naresh Manwani [b,*], P.S. Sastry [a]

[a] Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
[b] GE Global Research, John F. Welch Technology Centre, 122, EPIP Phase 2, Whitefield Road, Hoodi Village, Bangalore 560066, India

ARTICLE INFO

Article history:
Received 14 March 2014
Received in revised form
2 August 2014
Accepted 15 September 2014
Available online 11 February 2015

Keywords:
Classification
Label noise
Loss function
Risk minimization
Noise tolerance

ABSTRACT

In many applications, the training data, from which one needs to learn a classifier, is corrupted with label noise. Many standard algorithms such as SVM perform poorly in the presence of label noise. In this paper we investigate the robustness of risk minimization to label noise. We prove a sufficient condition on a loss function for the risk minimization under that loss to be tolerant to uniform label noise. We show that the 0–1 loss, sigmoid loss, ramp loss and probit loss satisfy this condition though none of the standard convex loss functions satisfy it. We also prove that, by choosing a sufficiently large value of a parameter in the loss function, the sigmoid loss, ramp loss and probit loss can be made tolerant to non-uniform label noise also if we can assume the classes to be separable under noise-free data distribution. Through extensive empirical studies, we show that risk minimization under the 0–1 loss, the sigmoid loss and the ramp loss has much better robustness to label noise when compared to the SVM algorithm.

1. Introduction

In a classifier learning problem we are given training data and when the class labels in the training data may be incorrect (or noise-corrupted), we refer to it as label noise. Learning classifier in the presence of label noise is a classical problem in machine learning [1]. This challenging problem has become more relevant in recent times due to the current applications of Machine Learning. In many of the web based applications, the labeled data is essentially obtained through user feedback or user labeling. This leads to data with label noise because of a lot of variability among different users while labeling and also due to the inevitable human errors. In traditional pattern recognition problems also, we need to tackle label noise. For example, overlapping class-conditional densities give rise to training data with label noise. This is because we can always view data generated from such densities as data that is originally classified according to, say, Bayes optimal classifier and then subjected to (non-uniform) label noise before being given to the learning algorithm. Feature measurement errors can also lead to label noise in the training data.

In this paper, we discuss methods for learning classifiers that are robust to label noise. Specifically we consider the risk minimization strategy which is a generic method for learning classifiers. We focus on the issue of making risk minimization robust to label noise.

Risk minimization is one of the popular strategies for learning classifiers from training data [2,3].[1] Many of the standard approaches for learning classifiers (such as Bayes classifier, neural network or SVM based classifier) can be viewed as (empirical) risk minimization under a suitable loss function. The Bayes classifier minimizes risk under the 0–1 loss function. One would like to minimize risk under 0–1 loss as it minimizes probability of mis-classification. However, in general, minimizing risk under 0–1 loss is computationally hard because it gives rise to a non-convex and non-smooth optimization problem. Hence many convex loss functions are proposed to make the risk minimization efficient. Square loss (used in feed-forward neural networks), Hinge loss (used in SVM), log-loss (used in logistic regression) and exponential loss (used in boosting) are some common examples of such convex loss functions. Many such convex loss functions are shown to be *classification calibrated*; that is, low risk under these losses implies low risk under 0–1 loss [4]. However, these results do not say anything about the robustness of such risk minimization algorithms to label noise. In this paper we present some interesting theoretical results on when risk minimization can be robust to label noise.

A learning algorithm can be said to be robust to label noise if the classifier learnt using noisy data and noise free data, both have same classification accuracy on noise-free test data [5]. In Manwani and Sastry [5], it is shown that risk minimization under 0–1 loss is tolerant to uniform noise (with noise rate less than 50%). It is also tolerant to non-uniform noise under some additional conditions. It is also shown in [5] through counter-examples that risk minimization under many

* Corresponding author.
E-mail addresses: aritraghosh.iem@gmail.com (A. Ghosh), nareshmanwani@gmail.com (N. Manwani), sastry@ee.iisc.ernet.in (P.S. Sastry).

[1] Risk minimization strategy is briefly discussed in Section 3.1.

http://dx.doi.org/10.1016/j.neucom.2014.09.081
0925-2312/

of the standard convex loss functions such as hinge loss, log loss or exponential loss is not noise-tolerant even under uniform noise.

In this paper, we extend the above theoretical analysis. We provide some sufficient conditions on a loss function so that risk minimization with that loss function becomes noise tolerant under uniform and non-uniform label noise. While 0–1 loss satisfies these, none of the standard convex loss functions satisfy the conditions. We also show that some of the non-convex loss functions such as sigmoid loss, ramp loss and probit loss satisfy the sufficiency conditions. Our results show that risk minimization under these loss functions is tolerant to uniform noise and that it is also tolerant to non-uniform noise if the Bayes risk (under noise-free data) is zero and if one parameter in the loss function is properly chosen. Hence we propose that risk minimization using sigmoid or ramp loss (which can be viewed as continuous but non-convex approximations to 0–1 loss) would result in learning methods that are robust to label noise. Through extensive empirical studies, we show that such risk minimization has good robustness to label noise.

The rest of the paper is organized as follows. In Section 2, we provide a brief review of methods for tackling label noise and then summarize the contributions of this paper. In Section 3 we define the notion of noise tolerance of a learning algorithm and formally state our problem. In this section we also provide a brief overview of the general risk minimization strategy. Section 4 contains all our theoretical results. We present simulation results on both synthetically generated data as well as on some benchmark datasets in Section 5. Some concluding remarks are presented in Section 6.

2. Prior work

Learning in the presence of noise is a long standing problem in machine learning. It has been approached from many different directions. A detailed survey of these approaches is given in Frénay and Verleysen [1].

In a recent study, Nettleton et al. present an extensive empirical investigation of robustness of many standard classifier learning methods to noise in training data [15]. They showed that the Naive Bayes classifier has the best noise tolerance properties. We comment more about this after presenting our theoretical results.

In general, when there is label noise, there are two broad approaches to the problem of learning a classifier. In the first set of approaches, data is preprocessed to clean the noisy points and then a classifier is learnt using standard algorithms. In the second set of approaches, the learning algorithm itself is designed in such a way that the label noise does not affect the algorithm. We call these approaches *inherently noise tolerant*. We briefly discuss these two broad approaches below.

2.1. Data cleaning based approaches

These approaches rely on guessing points which are corrupted by label noise. Once these points are identified, they can be either filtered out or their labels suitably altered. Several heuristics have been used to guess such noisy points.

For example, it is reasonable to assume that the class label of a point which is situated deep inside the class region of a class should match with the class labels of its nearest neighbors. Thus, mismatch of the class label of a point with most of its nearest neighbors can be used as a heuristic to decide whether a point is noisy or not [6]. This method of guessing noisy points may not work near the classification boundary. The performance of this heuristic also depends on the number of nearest neighbors used.

Another heuristic is that, in general, noisy points are tough to classify correctly. Thus, when we learn multiple classifiers using the noisy data, many of the classifiers may disagree on the class label of the noisy points. This heuristic has also been used to identify noisy points [7–9]. Decision tree pruning [10], distance of a point to the centroid of its own class [11], points achieving weights higher than a threshold in boosting algorithm [12], margin of the learnt classifier [13] are some other heuristics which have been used to identify the noisy examples.

As is easy to see, the performance of such heuristics depend on the nature of label noise. There is no single approach for identifying noisy points which can work for all problems. While each of the above heuristics has certain advantages, none of them are universally applicable. A non-noisy point can be detected as a noisy point and vice versa under any of these heuristics. This could eventually increase the overall noise level in the training data. Moreover, removal of the noisy points from the training data may lead to losing important information about the classification boundary [14].

2.2. Inherently noise tolerant approaches

These approaches do not do any preprocessing of the data; but the algorithm is designed in such a way that its output is not affected much by the label noise in the training data.

Perceptron algorithm, which is the simplest algorithm for learning linear classifiers, is modified in several ways to make it robust to the label noise [16]. Noisy points can frequently participate in updating the hyperplane parameters in the Perceptron algorithm, as noisy points are tough to be correctly classified. Thus, allowing a negative margin around the classification boundary can avoid frequent hyperplane updates caused due to the misclassifications with small margin. Putting an upper bound on the number of mistakes allowed for any example also controls the effect of label noise [16]. Similar techniques have been employed to improve Adaboost algorithm against noisy points. Overfitting problem in Adaboost, caused due to the label noise, can be controlled by introducing a prior on weights which can punish large weights [17]. In boosting algorithms, making the coefficients of each of the base classifiers input-dependent, also controls the exponential growth of weights due to noise [18]. SVM can be made robust to label noise by modifying the kernel matrix [19]. All these approaches are based on heuristics and work well in some cases. However, for most of these approaches, there are no provable guarantees of noise tolerance.

Noise tolerant learning has also been approached from the point of view of efficient probably approximately correct (PAC) learnability. By efficiency, we mean polynomial time learnability. Kearns [20] proposed a PAC learning algorithm for learning under label noise using statistical queries. However, the specific statistics that are calculated from the training data are problem-specific. PAC learning of the linear threshold functions is, in general, NP-hard [21]. However, linear threshold functions are efficient PAC learnable under uniform noise if the noise-free data is linearly separable with appropriate large margin [22]. For the same problem, Blum and Frieze [23] present a method to PAC-learn in the presence of uniform label noise without requiring the large margin condition. But the final classifier is a decision list of linear threshold functions. Cohen [24] proposed an ellipsoid algorithm which efficiently PAC learns linear classifiers under uniform label noise. This result is generalized further for class conditional label noise [25]. (Under class conditional noise model, the probability of a label being corrupted is the same for all examples of one class though different classes can have different noise rates.) All these results are given for linear classifiers and for uniform label noise. There are no efficient PAC learnability results under non-uniform label noise.

Recently Scott et al. [26] proposed a method of estimating Type 1 and Type 2 error rates of any specific classifier under the noise-free distribution given only the noisy training data. This is for the case of a 2-class problem where the training data is corrupted with class conditional label noise. They used the concept of mutually irreducible