Neurocomputing 192 (2016) 61–71

Contents lists available at ScienceDirect

Neurocomputing

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

A generalised label noise model for classification in the presence of annotation errors ☆

Jakramate Bootkrajang

Department of Computer Science, Chiang Mai University, Muang, Chiang Mai, 50200, Thailand

ARTICLE INFO

Article history:
Received 8 July 2015
Received in revised form
2 November 2015
Accepted 7 December 2015
Available online 27 February 2016

Keywords:
Non-random label noise
Classification
Logistic regression

ABSTRACT

Supervised learning from annotated data is becoming more challenging due to inherent imperfection of training labels. Previous studies of learning in the presence of label noise have been focused on label noise which occurs randomly, while the study of label noise that is influenced by input features, which is intuitively more realistic, is still lacking. In this paper, we propose a new, generalised label noise model which is able to withstand the negative effect of random label noise and a wide range of non-random label noises. Empirical studies using a battery of synthetic data and four real-world datasets with inherent annotation errors demonstrate that the proposed generalised label noise model improves, in terms of classification accuracy, upon existing label noise modelling approaches.

1. Introduction

A classification problem is a task where one wants to infer a {0,1}-valued function $\hat{h} : \mathcal{X} \rightarrow \mathcal{Y}$ using a finite sample $D = (\mathbf{x}_n, y_n)_{n=1}^{N}$: $\mathbf{x}_n \in \mathcal{X}, y_n \in \mathcal{Y} = \{0, 1\}$ drawn from some joint distribution on $\mathcal{X} \times \mathcal{Y}$. One can then use the estimated $\hat{h}$ to predict y for any new data $\mathbf{x}$ drawn from the same distribution. Here $\mathbf{x}$ is an m-dimensional feature vector and y is its label assignment. In an idealised scenario, y_n are assumed to be perfect. However, in reality, there is a possibility that the true label, y_n, is corrupted by some unknown factor so that we observe a flipped noisy $\tilde{y}_n$ instead of the true y_n. The quality of training labels has been theoretically [25,9,13,28,18] and empirically [26,15] shown to effect the performance of a classifier in a wide range of classification problems. Ensuring a close to perfect labelling turns out to be too costly in practice, especially with the scale and complexity of today's classification tasks. For example, the recent crowdsourcing practice for obtaining labelled training data cheaply and quickly could introduce label noise into the data set [30,29]. Label errors can also be found in complex classification tasks such as the classification of biomedical data [17,27,5,33] and the classification of textual data [21,20,2].

Class label noise can be loosely categorised into two types: random and non-random noise. The random label noise occurs independently of the input features [22]. The probability of label flipping is assumed to be class-conditional and is shared among all members in the same class. A non-random noise, on the other hand, is a noise which is influenced by the input features and hence is more general [23]. In the non-random noise case, the label flipping rates of all the members in the class are not necessarily equal and can vary within the class. Also, the non-random noise may be encountered more often than random noise in real-world problems. Pictorial illustrations of the two types on label noise on two dimensional data are given in Fig. 1a and b.

Interestingly, existing approaches to learning from noisy labels have been focused on random noise due to simplicity. Notable model-based robust classifiers include robust Kernel Fisher Discriminant [24], robust Normal Discriminant Analysis [3] and robust Logistic Regression [4,29], all of which are based on a weighed surrogate loss function. Relating to the above are the methods which utilise the so-called 'soft label' to quantify the degree of uncertainty of the training labels [12,14,10]. However, the study of the latter type of label noise is still scarce [23,8,31]. The reader is referred to [16] for an extensive survey on label noise problems.

Label noise modelling can be done at several levels of granularity. At the finest level, a noise model is associated with each data point. For example, a robust Logistic Regression proposed in [31] treats label noise of each training instance individually by incorporating a shift parameter into the sigmoid function. The parameter's role is to control the cutting point of the posterior probabilities of the two classes. This kind of *local* approximation is seemingly an ideal approach for the problem as it provides all the flexibility needed for capturing variations of noises. However, the method needs to estimate a huge number of noise parameters which unfortunately grows with the number of training instances.

☆Extended version of the work presented at ESANN 2015.

E-mail address: jakramate.b@cmu.ac.th

http://dx.doi.org/10.1016/j.neucom.2015.12.106
0925-2312/

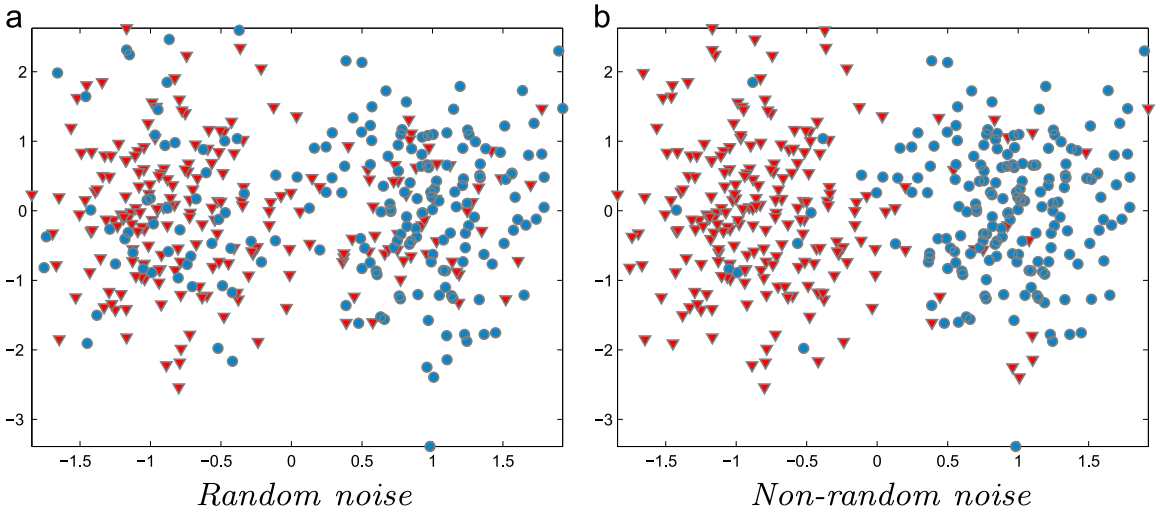

Fig. 1. Random noise occurs independently of the input features while non-random noise is influenced by the input features (in this case, there is less noise in the region further away from the decision boundary).

At the other end of the spectrum, a *global* statistic can be used for summarising the label flipping probabilities of all instances in the same class. For example, the work in [24], which targets random label noise, assumes that the instances in the class share the same label flipping probability. This significantly reduces the number of free parameters from $\mathcal{O}(N)$ to $\mathcal{O}(K)$, where N is the number of training instances and K is the number of classes. For this reason the global approach is widely adopted for solving random label noise problems [24,29,4]. Nonetheless, while the approach alleviates the curse of dimensionality, it is inevitably too restricted.

In this paper, we attempt to combine the advantages of the two approaches by proposing a more general label noise model which is flexible enough for dealing with both random and non-random label noises and is also simple such that the number of parameters is still merely of the order of the number of classes. We do this by expressing label flipping probabilities by a parametric function. We employ the probability density function of the exponential distribution to model the likelihood of label flipping. This function is chosen in order to capture noises in a scenario where points that live closer to the decision boundary have *relatively* higher chance of being mislabelled than those that live further away. Experiments show that the proposed method is able to counter the negative effect of the label noise while maintaining the computational feasibility of learning the model. We note that a similar assumption namely, points lies close to class mean have lower chance of being mislabelled has been investigated in the case of the normal discriminant analysis [8]. However, the study focuses on the theoretical aspect of the classifier while algorithmic solution for learning the model was not sufficiently described. In contrast, in this work we formulate the mislabelling probability as a function of distance from the decision boundary and propose a robust logistic regression employing the new label noise model together with an efficient algorithm to learn the robust model.

To sum up, the contributions of our work are the followings.

- We proposed a new label noise model which can deal with both random label noise and example-dependent label noise.
- We developed a new robust Logistic Regression employing the newly proposed noise model and devised an efficient algorithm to learn the classifier.
- We extensively evaluated the usefulness of the proposed method on a battery of synthetic datasets and real datasets which genuinely contain annotation errors.

The rest of the paper is organised as follows. Section 2 introduces the generalised label noise model, a new robust logistic regression employing the new noise model and an efficient algorithm to learn the classifier. Section 3 presents empirical evaluations and discussions of the results while Section 4 concludes the study.

2. The generalised label noise model

One of the principled ways for dealing with *random* label noise problem is the use of a latent variable model [24,4]. The approach represents the class posterior probability of the observed label with a weighted posterior probability of the true class labels. Under the latent variable model, the probability that the observed label of a point $\mathbf{x}_n$ is k is given by:

$$\tilde{P}_n^k = \sum_j p(\tilde{y}_n = k|y = j) \cdot p(y = j|\mathbf{x}_n, \theta) \tag{1}$$

Here $p(\tilde{y} = k|y = j)$ denotes a *label flipping probability* that the true class label j was flipped into the observed class label k. Clearly, the label flipping probability is class-conditional and is independent of the input vector.

Arguably, such assumption is rather unrealistic for real-world problems as input features can have some influence on the occurrence of mislabelling, so the random latent variable model may not be appropriate. To generalise the above noise model to accommodate label noise which may depend on the input vector, we redefine the label flipping probability to be a function of the input vector, its class label and the parameters of the classification model.

$$\tilde{P}_n^k = \sum_j \mathcal{F}(\mathbf{x}_n, \tilde{y}_n, y_n = j, \theta) p(y_n = j|\mathbf{x}_n, \theta) =: \sum_j \mathcal{F}(\mathbf{x}_n, \tilde{y}_n, y_n = j, \theta) P_n^j \tag{2}$$

where $\mathcal{F}(\mathbf{x}_n, \tilde{y}_n, y_n = j, \theta) \stackrel{def}{=} p(\tilde{y}_n = k|y_n = j, \mathbf{x}_n, \theta) = \omega_n^{jk}$. The function $\mathcal{F}$ can be any function which best describes the nature of the label flipping and has to satisfy the probabilistic constraint, i.e., outputting a value between zero and one. The proposed model will be referred to as the *generalised label noise model*. Note that the random label noise model used in [24,29] is a special case of the above noise model, where $\mathcal{F}$ is defined to be a constant function. It is worth mentioning that the selection of the noise function depends highly on the knowledge of noise. It is very unlikely that