A Generalised Label Noise Model for Classification

Jakramate Bootkrajang*

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

Abstract. Learning from labelled data is becoming more and more challenging due to inherent imperfection of training labels. In this paper, we propose a new, generalised label noise model which is able to withstand the negative effect of both random noise and a wide range of non-random label noises. Empirical studies using three real-world datasets with inherent annotation errors demonstrate that the proposed generalised label noise model improves, in terms of classification accuracy, over 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 shown to effect the performance of a classifier in a wide range of classification problems [1, 2, 3, 4]. 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.

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. A non-random noise, on the other hand, is a noise which is influenced by the input features and hence is more general. Also, the non-random noise may be encountered more often than random noise in real-world classification problems. Interestingly, previous model-based approaches to learning from noisy labels have been focused on random noise due to simplicity [5, 2, 6]. The study of the latter type is still scarce [7, 8, 9].

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 [8] 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

*This work is supported by the Faculty of Science, Chiang Mai University.

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 need to estimate a huge number of noise parameters which unfortunately grows with the number of training instances. 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 [5], 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 [5, 2, 6]. 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.

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 [5, 4]. The approach represents the class posterior probability of the observed label with a weighted posterior probability of the true class labels. The probability of the observed label being class k for a point $\mathbf{x}_n$ under the latent variable model is then given by: $\tilde{P}_n^k = \sum_j p(\tilde{y}_n = k|y = j) \cdot p(y = j|\mathbf{x}_n, \theta)$. 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, which is independent of the input vector.

Arguably, such assumption is rather unrealistic for real-world problems as input features can have an 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.

$$\tilde{P}_n^k = \sum_j \mathcal{F}(\mathbf{x}_n, \tilde{y}_n, y_n = j, \theta) p(y_n = j|\mathbf{x}_n, \theta) =: \sum_j \mathcal{F}(\mathbf{x}_n, \tilde{y}_n, y_n = j, \theta) P_n^j \quad (1)$$

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 [5, 2] is a special case of the above noise model, where $\mathcal{F}$ is defined to be a constant polynomial.

According to our initial assumption that points lie close to the decision boundary have higher chance of being mislabelled than those that live further away, we find that a probability density function of the exponential distribution would suit our purpose. The noise function will take as input a distance of the point $(\mathbf{x}_n, \tilde{y}_n = k)$ from the decision boundary. Denoting the distance by Z_n^k, we define the label flipping probabilities to be:

$$p(\tilde{y}_n = 1 | y_n = 0, \mathbf{x}_n, \theta) = \frac{\exp\left(-Z_n^1/\gamma_0\right)}{\gamma_0} = \omega_n^{01} \tag{2}$$

$$p(\tilde{y}_n = 0 | y_n = 1, \mathbf{x}_n, \theta) = \frac{\exp\left(-Z_n^0/\gamma_1\right)}{\gamma_1} = \omega_n^{10} \tag{3}$$

Using $\sum_k p(\tilde{y} = k | y, \mathbf{x}, \theta) = 1$, one can easily derive ω_n^{00} and ω_n^{11}. Since Z_n is non-negative, $\exp\left(-Z_n^k/\gamma_j\right)/\gamma_j \in [0, 1]$ when $\gamma_j > 1$. We will employ a log barrier function, $\log(\gamma_j - 1)$ to impose this constraint.

For the sake of exposition, we will use a Logistic Regression parametrised by $\theta = \mathbf{w}$ as our base classifier. Here, the parameter $\mathbf{w}$ is the weight vector orthogonal to the decision boundary. The Euclidean distance of a point from the decision boundary of the classifier is given by $Z_n = \mathbf{x}_n^T \mathbf{w} / \|\mathbf{w}\|$. Putting everything together, the objective function of the *generalised robust Logistic Regression* (gLR) is a penalised log-likelihood:

$$\begin{aligned}\mathcal{L} = &\sum_{n=1}^{N} \tilde{y}_n \log\left[\omega_n^{11} P_n^1 + \omega_n^{01} P_n^0\right] + (1 - \tilde{y}_n) \log\left[\omega_n^{00} P_n^0 + \omega_n^{10} P_n^1\right] - \sum_{i=1}^{m} \alpha_i |w_i| \\ &+ \sum_{j=0}^{1} \lambda_j \log(\gamma_j - 1)\end{aligned} \tag{4}$$

where $\alpha_i > 0$ is a Lagrange multiplier and λ_j is a parameter expressing the sharpness of the barrier function at the boundary. The first two terms represent the log-likelihood, the third term is the L1 regulariser and the last term enforces the constraint that $\gamma_j > 1$. The class posterior probability is modelled by the sigmoid function: $P_n^1 = 1/(1 + e^{-\mathbf{w}^T \mathbf{x}})$.

To optimise the objective, we use the gradient-descent method to update $\mathbf{w}$, γ_0 and γ_1. We adopt an effective smooth approximation, $|w_i| \approx (w_i^2 + \eta)^{1/2}$, originally proposed by [10] to take care of the discontinuity of the objective at the origin caused by the regularisation. We used $\eta = 10^{-8}$ in the reported experiments. The gradient of the objective function w.r.t $\mathbf{w}$ is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \sum_{n=1}^{N} \left[\left(\frac{\tilde{y}_n}{\tilde{P}_n^1} - \frac{1 - \tilde{y}_n}{\tilde{P}_n^0} \right) (1 - \omega_n^{10} - \omega_n^{01}) \right] P_n^1 P_n^0 \mathbf{x}_n - \sum_{i=1}^{m} \frac{\alpha_i w_i}{\sqrt{(w_i^2 + \eta)}} \tag{5}$$

Next, the gradients of the objective w.r.t. γ_0 and γ_1 are found:

$$\frac{\partial \mathcal{L}}{\partial \gamma_0} = \sum_{n=1}^{N} \left(\frac{\tilde{y}_n}{\tilde{P}_n^1} - \frac{1-\tilde{y}_n}{\tilde{P}_n^0} \right) \left(\frac{2Z_n^1}{\gamma_0^2} + \frac{2}{\gamma_0} \right) \omega_n^{01} P_n^0 + \frac{\lambda_0}{\gamma_0 - 1} \tag{6}$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_1} = \sum_{n=1}^{N} \left(\frac{1-\tilde{y}_n}{\tilde{P}_n^0} - \frac{\tilde{y}_n}{\tilde{P}_n^1} \right) \left(\frac{2Z_n^0}{\gamma_1^2} - \frac{2}{\gamma_1} \right) \omega_n^{10} P_n^1 + \frac{\lambda_1}{\gamma_1 - 1} \tag{7}$$

Further, the value of the regularisation parameter α_i is determined using the Bayesian regularisation technique. In a similar spirit as in [6], a Bayesian interpretation of the objective w.r.t. γ and λ is given by: $\log p(\mathbf{w}|D) = \log p(D|\mathbf{w}) + \log p(\mathbf{w}|\boldsymbol{\alpha}) + \text{const.}$. The conditional prior $p(\mathbf{w}|\boldsymbol{\alpha})$ is the the product of independent Laplace distributions $p(\mathbf{w}|\boldsymbol{\alpha}) \approx \frac{\prod_{i=1}^{m} \alpha_i}{2^m} \exp(-\sum_{i=1}^{m} \alpha_i (w_i^2 + \eta)^{1/2})$. The parameter-free Jeffrey's priors is used to model each of the $\boldsymbol{\alpha}$: $p(\alpha_i) \propto \frac{1}{\alpha_i}$. The marginal prior $p(\mathbf{w})$ is then given by the integral $\int_o^\infty p(\mathbf{w}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) \mathtt{d}\boldsymbol{\alpha} = \frac{1}{2} \prod_{i=1}^{m} \frac{1}{(w_i^2+\eta)^{1/2}}$, which implies that $-\log(\mathbf{w}) = \sum_{i=1}^{m} \log(w_i^2 + \eta)^{1/2}$. Taking derivative of the resulting negative log of the marginal prior, we have:

$$-\frac{\partial \log p(\mathbf{w})}{\partial w_i} = \frac{1}{(w_i^2 + \eta)^{1/2}} \frac{\partial \sum_{i=1}^{m} \log((w_i^2 + \eta)^{1/2})}{\partial w_i} \tag{8}$$

From the above we read off the estimate of the regularisation parameter: $\alpha_i = 1/(w_i^2 + \eta)^{1/2}$. Next, we propose to use a simple heuristic, $\lambda_j = 1/(\gamma_j - 1)$, for setting the value of λ_j. The intuitions behind this heuristic are twofolds: first, to enforce increasingly larger penalty as γ_j approaches 1, in which case the penalty is amplified by $\lambda_j > 1$ and second, to prevent γ_j from being unreasonably large, in which case the corresponding λ_j will control the gain in likelihood.

With everything in place, the optimisation is then to alternate between updating $\mathbf{w}$, γ_0, γ_1 and the regularisation parameters in turn.

3 Experiments

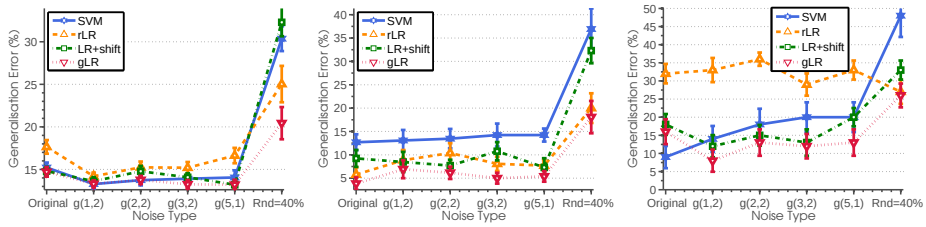
3.1 Experimental protocol and Datasets

We evaluated the proposed model on three real-world datasets which originally contain labelling errors according to literature. Two datasets are from biomedical domain namely *Colon* and *Breast* datasets (cf. [1]). The last one is an image classification dataset called the *Websearch* [4], constructed by querying a search engine for images matching a keyword and taking the keyword used to be the class label of the retrieved images (For more details see [4]). The characteristics of all datasets used in this study are summarised in Table 1.

In addition, since the ground truths labels are available for all the datasets, we further evaluate the model by artificially injecting label noise of various types into the *cleansed* version of the datasets. To generate the non-random label noise we first train an SVM on an untainted version of the datasets. The resulting "optimal" weight vector, together with a label noise function, are used to calculate

Dataset	# of samples (pos./neg.)	# wrong labels (pos./neg.)	# features
Colon	40(T)/22(N)	5/4	2000
Breast	25(ER+)/24(ER-)	4/5	7129
Websearch	515(bike)/515(not bike)	100/83	1318

Table 1: The characteristics of the datasets employed in this study



(a) *Websearch* (b) *Colon* (c) *Breast*

Fig. 1: Generalisation error (%) of the four classifiers on various types of label noises

the mislabelling probabilities of the input instances. Here we use the PDF of the gamma distribution with the shape/scale parameters pair being (1,2),(2,2),(3,2) and (5,1) to sample label flippings. We also test our model on a random noise scenario where 40% of the training labels are contaminated. We performed 20 experiment repetitions for each level of contaminations using 80/20 train/test split except on *Breast* dataset where we use 90% of the samples for training due to the small sample size and high dimensional nature of the dataset. We study comparative performances of the proposed gLR versus rLR [4], LR+shift [8] and the gold standard SVM with RBF kernel.

3.2 Results and Discussion

We first report the experimental results from the *Websearch* dataset. The noise in this dataset is less likely to appear at random. This is because the search engine might have used textual information around the image during the search process. The results in Fig.1a, demonstrate that gLR is capable of dealing with all types of noises, from the noise originally inherent in the dataset to other simulated noises including random label noise. The rLR, which relies on the random noise assumption, performs reasonably well only when random noise is presented (ranked second) but lags behind in non-random noise cases. The LR+shift is more robust than rLR on all type of noises studied except on extreme random noise. The SVM ranked second overall in this dataset but its performance drops significantly when random noise is presented.

Fig.1b and Fig.1c summarise the results from *colon* and *breast* datasets. Again, in these datasets the nature of the inherent noise would be far from being random. It is expected that noise would appear more in the region of maximum

confusion. As can be seen from the results, the gLR employing the proposed generalised label noise model performs better than the other robust classifiers and the SVM in almost all types of noise except the original *breast* dataset where SVM performs better. We speculate firstly that SVM can cope with a mild noise and secondly that the RBF kernel used might have helped. However the power of non-linearity leads to serious overfitting in some other cases. Overall, we can conclude based on these empirical evidences that the proposed generalised label noise model helps counteracting the negative effect of various types of noises including random noise, and that the existing random noise model is less suitable for real-world applications where non-random is present.

4 Conclusion

We presented a novel label noise model for classification where the training labels are inaccurate. The proposed model seeks to explain the label noise using any customised label noise function deemed appropriate for the task. We paired the proposed model with the Logistic Regression classifier and evaluated the robust classifier on a non-random noise scenario where noise appears more in the region near the optimal decision boundary. The experimental results revealed that the proposed model was able to counter the negative effect of such label noise, and outperformed both the gold standard SVM and the existing robust classifiers. The future work will be to investigate theoretical aspects of the proposed model.

References

[1] Andrea Malossini, Enrico Blanzieri, and Raymond T. Ng. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics*, 22(17):2114–2121, 2006.

[2] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[3] Benoît Frénay, Gael de Lannoy, and Michel Verleysen. Label noise-tolerant hidden markov models for segmentation: Application to ECGs. In *ECML-PKDD'11*, pages 455–470, 2011.

[4] Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In *ECML-PKDD'12*, pages 143–158, 2012.

[5] Neil D. Lawrence and Bernhard Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In *ICML'01*, pages 306–313. Morgan Kaufmann, 2001.

[6] Jakramate Bootkrajang and Ata Kabán. Learning kernel logistic regression in the presence of class label noise. *Pattern Recognition*, 47(11):3641–3655, 2014.

[7] Peter A. Lachenbruch. Discriminant analysis when the initial samples are misclassified ii: Non-random misclassification models. *Technometrics*, 16(3):pp. 419–424, 1974.

[8] Julie Tibshirani and Christopher D. Manning. Robust logistic regression using shift parameters. *CoRR*, abs/1305.4987, 2013.

[9] Benoît Frénay and Michel Verleysen. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[10] Liu Zhenqiu, Jiang Feng, Tian Guoliang, Wang Suna, Sato Fumiaki, Meltzer Stephen J., and Tan Ming. Sparse logistic regression with lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, 6(1):1–22, 2007.