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/

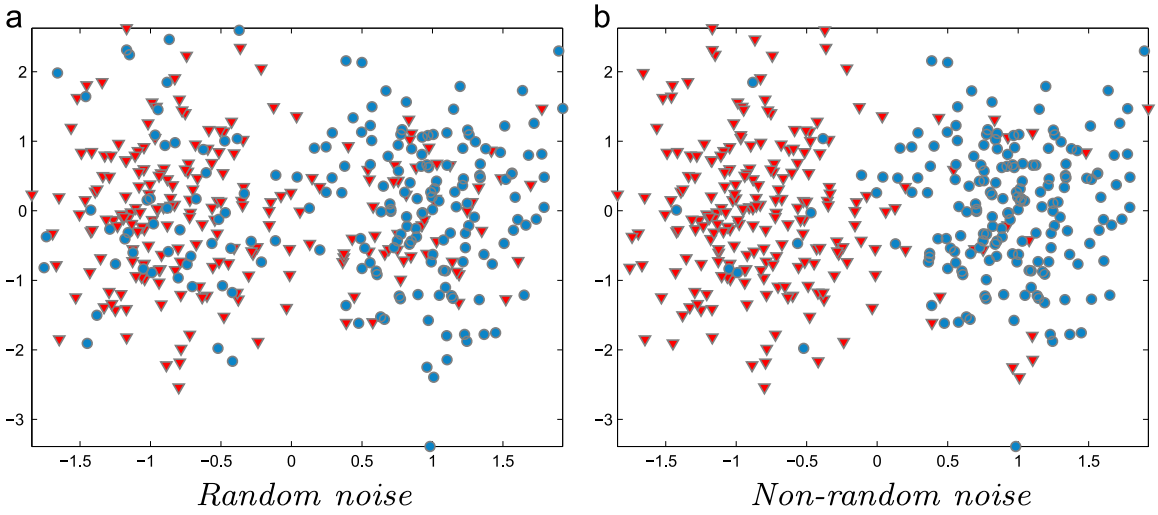

Fig. 1. Random noise occurs independently of the input features while non-random noise is influenced by the input features (in this case, there is less noise in the region further away from the decision boundary).

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

there will be a single noise function that works well in all cases. Nevertheless, with the proposed noise model, it boils down to choosing an appropriate noise function and plugging it into the model instead of developing a new model from scratch each time we see a new set of data.

For the sake of exposition and according to our initial assumption that points that 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:

$$p(\tilde{y}_n = 1 | y_n = 0, \mathbf{x}_n, \theta) = \frac{\exp(-Z_n^1/\gamma_0)}{\gamma_0} = \omega_n^{01} \tag{3}$$

$$p(\tilde{y}_n = 0 | y_n = 0, \mathbf{x}_n, \theta) = 1 - \omega_n^{01} = \omega_n^{00} \tag{4}$$

$$p(\tilde{y}_n = 0 | y_n = 1, \mathbf{x}_n, \theta) = \frac{\exp(-Z_n^0/\gamma_1)}{\gamma_1} = \omega_n^{10} \tag{5}$$

$$p(\tilde{y}_n = 1 | y_n = 1, \mathbf{x}_n, \theta) = 1 - \omega_n^{10} = \omega_n^{11} \tag{6}$$

Since Z_n is non-negative, $\exp(-Z_n^k/\gamma_j)/\gamma_j \in [0, 1]$ when $\gamma_j > 1$. We will employ a log barrier function, $\log(\gamma_j - 1)$ to impose this constraint.

2.1. Generalised robust logistic regression

The proposed label noise model could readily be incorporated into a margin-based probabilistic classifier to yield a robust classification algorithm. 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:

$$\mathcal{L} = \sum_{n=1}^{N} \left(\tilde{y}_n \log \left[\omega_n^{11} P_n^1 + \omega_n^{01} P_n^0 \right] + (1 - \tilde{y}_n) \log \left[\omega_n^{00} P_n^0 + \omega_n^{10} P_n^1 \right] \right) - \sum_{i=1}^{m} \alpha_i |w_i| + \sum_{j=0}^{1} \lambda_j \log(\gamma_j - 1) \tag{7}$$

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 term represents the log-likelihood, the second 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 [34] to take care of the discontinuity of the objective at the origin caused by the L1 regularisation. We used $\eta = 10^{-8}$ in the reported experiments. The gradient of the objective function w.r.t $\mathbf{w}$ is:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \sum_{n=1}^{N} \left[\left(\frac{\tilde{y}_n}{\tilde{P}_n^1} - \frac{1 - \tilde{y}_n}{\tilde{P}_n^0} \right) (1 - \omega_n^{10} - \omega_n^{01}) \right] P_n^1 P_n^0 \mathbf{x}_n - \sum_{i=1}^{m} \frac{\alpha_i w_i}{\sqrt{(w_i^2 + \eta)}} \tag{8}$$

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

$$\frac{\partial \mathcal{L}}{\partial \gamma_0} = \sum_{n=1}^{N} \left(\frac{\tilde{y}_n}{\tilde{P}_n^1} - \frac{1 - \tilde{y}_n}{\tilde{P}_n^0} \right) \left(\frac{2 Z_n^1}{\gamma_0^2} + \frac{2}{\gamma_0} \right) \omega_n^{01} P_n^0 + \frac{\lambda_0}{\gamma_0 - 1} \tag{9}$$

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

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:

$$\log p(\mathbf{w}|D) = \log p(D|\mathbf{w}) + \log p(\mathbf{w}|\boldsymbol{\alpha}) + \text{const.} \tag{11}$$

For sparsity inducing effect, the conditional prior $p(\mathbf{w}|\boldsymbol{\alpha})$ is modelled using a product of independent Laplace distributions,

$$p(\mathbf{w}|\boldsymbol{\alpha}) \approx \frac{\prod_{i=1}^{m} \alpha_i}{2^m} \exp\left(- \sum_{i=1}^{m} \alpha_i (w_i^2 + \eta)^{1/2} \right) \tag{12}$$

The parameter-free Jeffrey's prior is used to model each of the hyperparameter, $\boldsymbol{\alpha}$: $p(\alpha_i) \propto \frac{1}{\alpha_i}$. To eliminate the dependency on α_i and obtain the marginal prior $p(\mathbf{w})$, we complete an integration using the Gamma integral: $\int_0^\infty x^{\nu-1} e^{-\mu x} \, dx = \frac{\Gamma(\nu)}{\mu^\nu}$

$$\int_o^\infty p(\mathbf{w}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) \, d\boldsymbol{\alpha} = \frac{1}{2} \prod_{i=1}^{m} \frac{1}{(w_i^2 + \eta)^{1/2}}, \tag{13}$$

which implies that $-\log p(\mathbf{w}) = \sum_{i=1}^{m} \log (w_i^2 + \eta)^{1/2}$. Taking derivative of the resulting negative log of the marginal prior, we have:

$$-\frac{\partial \log p(\mathbf{w})}{\partial w_i} = \frac{1}{(w_i^2 + \eta)^{1/2}} \frac{\partial \sum_{i=1}^{m} \log ((w_i^2 + \eta)^{1/2})}{\partial w_i} \tag{14}$$

From the above we read off the estimate of the regularisation parameter:

$$\alpha_i = 1/(w_i^2 + \eta)^{1/2} \tag{15}$$

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 two folds: 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. The parameter λ_j, which is too small translates to large flipping probability and is undesirable because in such case the algorithm will be too pessimistic, i.e., believing that most of the examples are mislabelled. On the other hand, with large λ_j, the flipping probability will be near zero, thus making the algorithm think that there is no noise in the dataset. Our heuristic is then to discourage λ_j from taking such extreme values.

With everything in place, the optimisation is then to alternate between updating $\mathbf{w}$, γ_0, γ_1 and the regularisation parameters in turn. An algorithm to learn the generalised robust logistic regression is summarised in Algorithm 1.

Algorithm 1. Optimisation of generalised robust Logistic Regression (gLR).

Input: Set of training data $(\mathbf{x}_n, y_n)_{n=1}^N$
Initialise $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_{D \times D})$, $\gamma_{0,1} \leftarrow 5$
while Iteration < MaxIteration
Calculate $Z_n = \mathbf{x}_n^T \mathbf{w}/\|\mathbf{w}\|$
Calculate ω_n^{jk} using Eqs. (3)–(6)
Update $\mathbf{w}$ using Eq. (8) and the pre-calculated ω_n^{jk}
Update γ_i using Eqs. (9) and (10)
Update α_i using Eq. (15)

Update λ_j using $\lambda_j = 1/(\gamma_j - 1)$
end while
Output: Optimised weight vector, **w**. Optimised γ.

3. Experiments

We shall now present empirical studies which validate the effectiveness of the proposed generalised label noise model against labelling errors in the training data. We organised the experiment into two parts. The first part is concerned with evaluating the proposed label noise model on simulated data with artificially created label noise. The latter part presents comparative performance of the proposed model to state-of-the-art classifiers on real-world datasets which genuinely contain label noise.

3.1. Synthetic datasets

We first present the results from a series of synthetic datasets. The purpose of this study is to gain better understanding of the performance of the newly proposed generalised label noise model compared to the existing latent variable model in controlled environment. Here we will compare the generalised logistic regression (gLR) to the existing robust logistic regression (rLR) [4] and the traditional logistic regression (LR). To ensure fair comparisons, all of the algorithms are equipped with the L1-regularisation.

There are two types of synthetic data employed in this study: discriminatively generated data and generatively generated data. For the discriminative dataset, we sampled N points from an m-dimensional spherical Gaussian and assigned the points from one half of the hypersphere to the positive class and the points from the other half to the negative class. For the generative dataset, we sampled points from two spherical Gaussians which are c-separated. The class separation [11] is used to quantify the degree of classes overlapping, and is defined as

$$c = \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\| / \sqrt{m \max(\lambda_{max}(\boldsymbol{\Sigma}_0), \lambda_{max}(\boldsymbol{\Sigma}_1))}, \quad (16)$$

where $\lambda_{max}(\boldsymbol{\Sigma})$ represents the largest eigenvalue of the covariance $\boldsymbol{\Sigma}$, μ_i is the mean vector of class i and m is the dimensionality of the data. A 0.5-separated data is highly overlapped dataset while a 5-separated data is rather well-separated dataset.

We then artificially injected five types of label noises, four of which are non-random label noises and the remaining is random label noise, into the dataset. To generate the non-random label noise we first train an SVM on untainted version of the dataset. The resulting optimal weight vector, together with a label noise function, are used to calculate the mislabelling probabilities of the input instances. Here, we used the PDF of the gamma distribution with the (shape,scale) parameters pair being (1,2), (2,2), (3,2) and (5,1) to simulate label flipping. The label flipping probabilities of the four non-random noise configurations are illustrated in Fig. 2. We used 30% random misclassification rate to test the classifiers on random label noise. In addition to varying the label flipping rate we also consider noise of symmetric and asymmetric types. A symmetric label noise is when both classes have equal probability of getting its label flipped while asymmetric class is where only instances from one class are flipped into the other class but not vice-versa.

We studied the effects of noise types, number of training examples N, data dimensionality m and additionally, class separation c in the case of generative dataset on the performance of the classifiers. We performed 20 experiment repetitions for each data setup and report the averaged classification errors and standard errors.

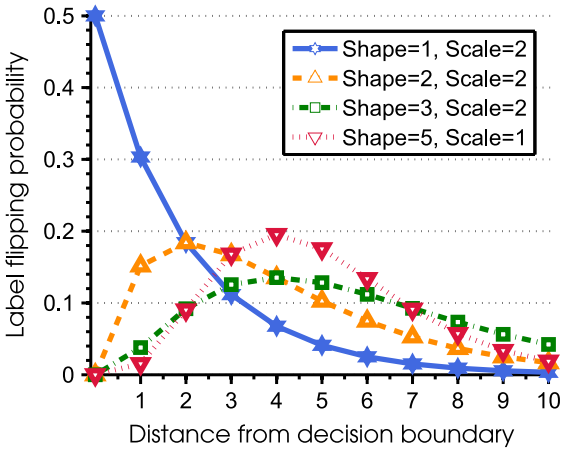

Fig. 2. Label flipping probabilities as a function of distance from the decision boundary.

3.2. Results: Synthetic data

We start off with the effect of training set size on classification performance in discriminative data setting. Fig. 3 summarises the classification error on $m=10$ dimensional data as dataset size varies, grouped by types of label noise. We first note that the improvement of gLR over LR is clear. The newly proposed classifier outperformed the traditional logistic regression in all cases tested. The performance gap becomes more apparent as more training examples are available, in which case the estimation of noise is more accurate resulting in low classification error. We also see the degree of negative effect each type of noises has on the classification performance. Here, *gamma*(1,2) and 30% random label noise seem to perturb the learning most while *gamma*(3,2) and *gamma*(5,1) have relatively mild effect.

In comparison to rLR, the proposed gLR tends to perform better when training data is scarce, e.g., at $N < 100$. However, in abundance of training data, the performance gaps between the two are marginal. Interestingly, we observed that gLR performed as well as rLR, a model tailored to tackle random label noise, on data with random label errors. This somehow demonstrates the generality of gLR in dealing with both types of noises. Note that rLR is also surprisingly robust to non-random noises in this data setting where data dimensionality is low. In the sequel, we shall investigate further to see if rLR will remain robust as the dimensionality of the data increases.

Continuing from the preceding study, we now fix $N=1000$ and study the effect of data dimensionality. Fig. 4 presents our findings. We first observe that high dimensional data turns out to be more challenging such that the negative effect of label noise is more pronounced. We are starting to see the deterioration of rLR's robustness, which was rather impressive at low dimensional settings. The proposed gLR is, in general, more robust compared to rLR throughout the entire test range, and especially so at high dimensional data with non-random mislabelling. Interestingly, rLR is now lagging behind gLR in the case where random labelling errors are present. This signifies the advantage of the proposed generalised label noise model over the existing latent variable model.

Figs. 5 and 6 report the results from varying dataset size and data dimensionality in the case of generative data where class separation is fixed to $c=1.5$, respectively. The outcomes are inline with what we have seen on the discriminative data settings. The

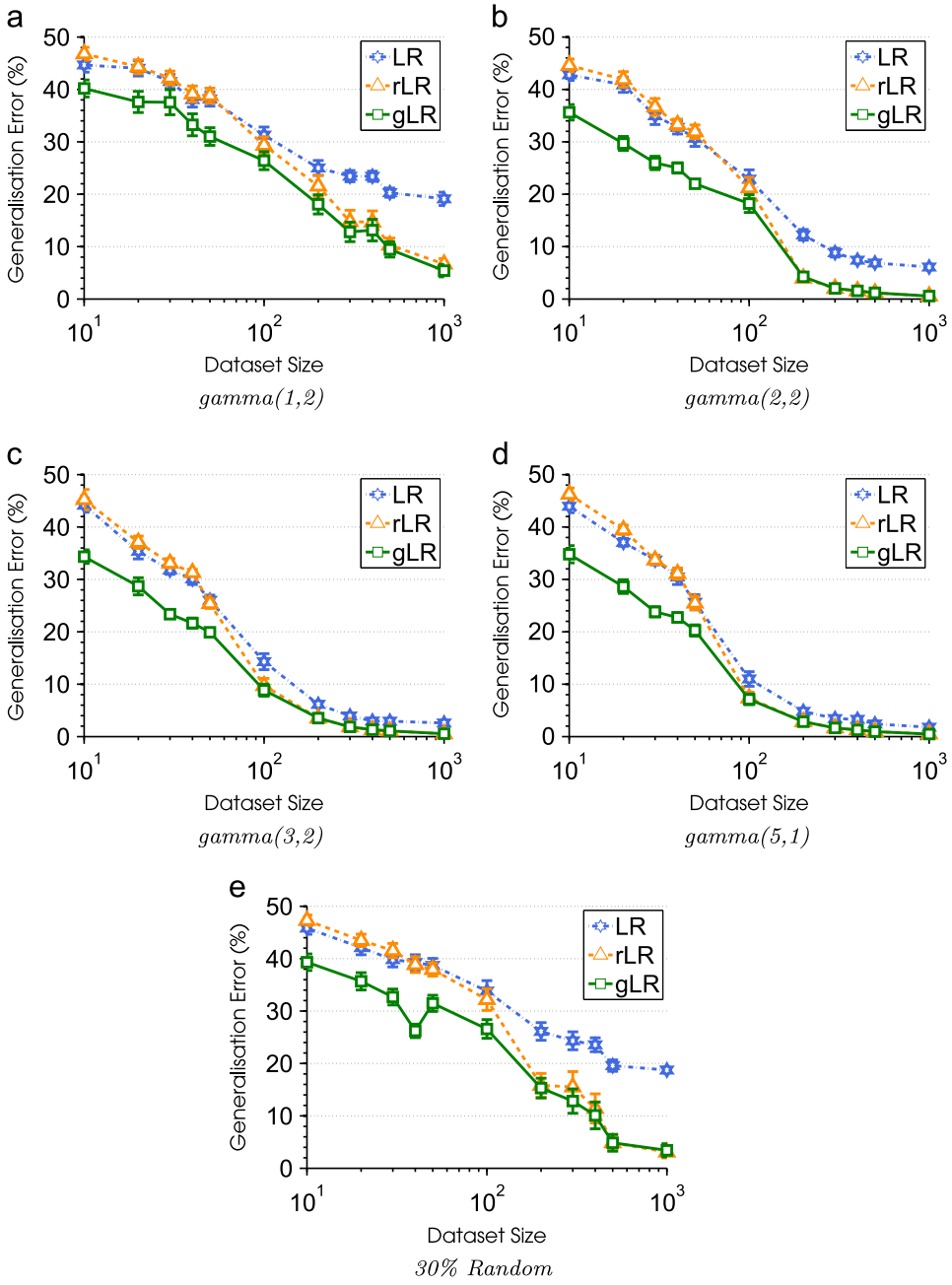

Fig. 3. Mean classification errors and their standard errors on 10-dimensional discriminatively generated data on five noise types with varying dataset size. The improvement of gLR over LR and rLR is well apparent especially when training examples are scarce.

proposed gLR clearly outperforms traditional LR and is superior to rLR in almost all data configurations, except in the cases of low dimensional data or in the situations where there are enough data points. In such cases, rLR is able to withstand the effect of mislabelling.

Lastly, on the battery of synthetic datasets, we study the comparative performance of the three classifiers in the dynamic of class separation. We can see from the results summarised in Fig. 7 that gLR is superior to the other two competing classifiers in almost all cases. The only scenario where the benefit of the proposed model is less significant is when the classes are highly overlapped. This is expected, though, since in such case it would be very difficult to differentiate noise from natural overlapping. The situation eases up as classes become more separated such that the performance gaps between gLR and the competing algorithms become apparent. Recall that mislabelling probability is, in some way, proportional to the distance between class means and for that reason we are starting to see the performance gaps between the competing algorithms diminish as c becomes larger in the non-random noise cases. However, the random label noise is invariant to class separation, so gLR continues to dominate even in the situation where c is large (Fig. 7e).

Based on the above empirical results, it is expected that gLR will perform better than rLR in general noisy cases and especially so in the case where data are high dimensional with limited training examples.

3.3. Real-world datasets

So far, we have only witnessed the robustness of the proposed method on artificially created testbed which is only part of the story. We will now move to real-world datasets which tend to be

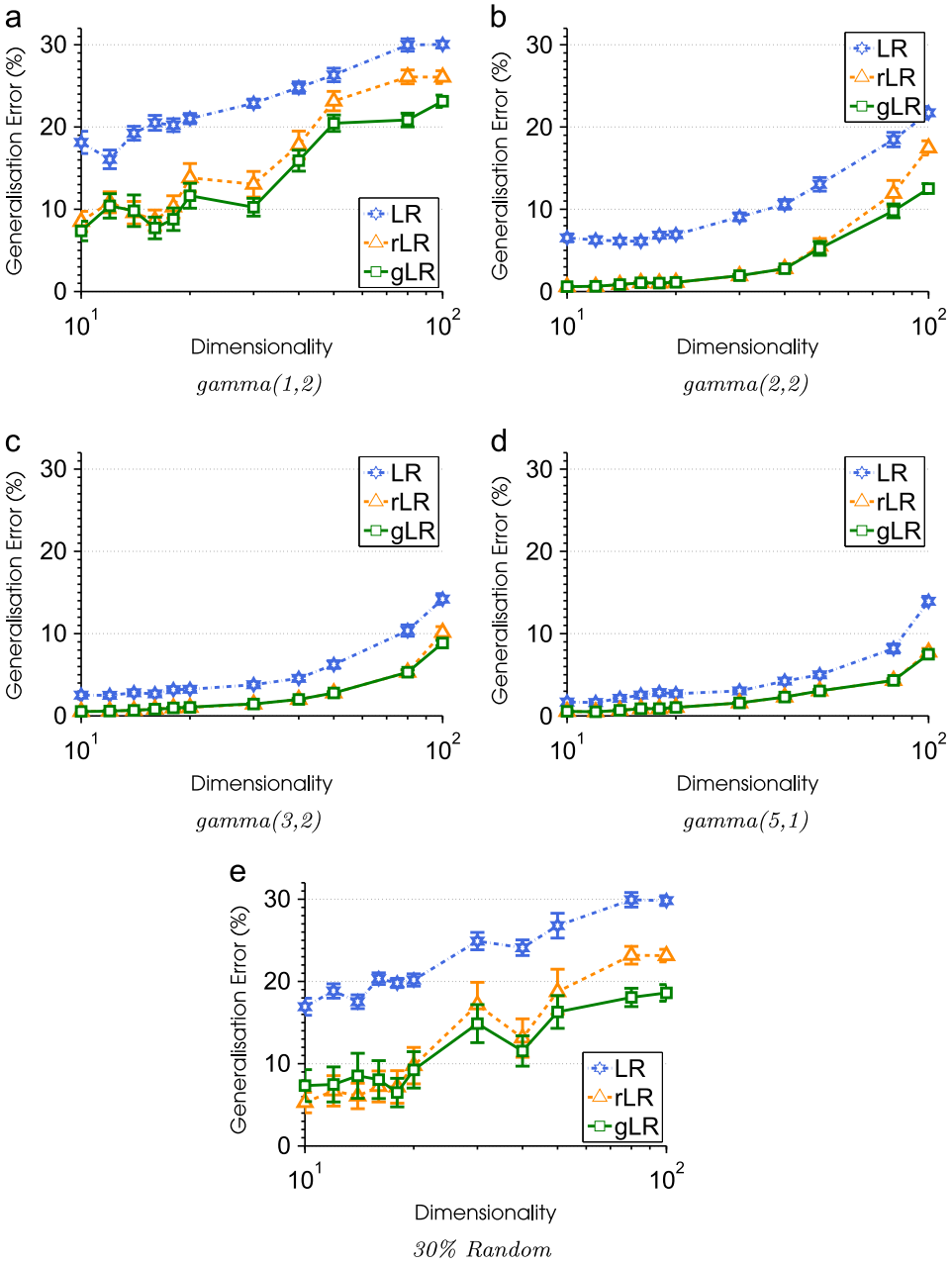

Fig. 4. Mean classification errors and their standard errors on discriminatively generated data with 1000 datapoints on five noise with varying data dimensionality. The proposed gLR is, in general, more robust compared to rLR throughout the entire test range, and especially so at high dimensional data.

more complex and may not fit our model assumption. We will evaluate the proposed model on four real-world datasets which are reported to originally contain labelling errors according to literature. Three datasets are from bio-medical domain namely *Colon* [1], *Breast* [32] and *Leukaemia* [19] datasets. 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 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.

We note that the mislabelled data points in the mentioned datasets were identified with biological/supporting evidence [1,32,4] except for *Leukaemia* dataset where the only mislabelling was identified by a consensus of label noise detecting algorithm in the literature. 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. The protocol for generating label noise is identical to the case of synthetic datasets presented above. We performed 20 experiment repetitions for each level of contaminations using 80/20 train/test split except on *Breast* dataset where we used 90% of the samples for training due to the small sample size and high dimensional nature of the dataset. Evaluation is made by considering original noise 'Original' and other artificially added noises namely 'g(1,2)', 'g(2,2)', 'g(3,2)', 'g(5,1)' and 'Random'. We studied comparative performances of the proposed gLR versus rLR [4], LR+shift [31], robust Normal Discriminant Analysis (rNDA) with full covariance matrix [24] and the gold standard SVM with RBF kernel.[1] A 10-fold cross validation technique was adopted for selecting optimal kernel parameter and optimal C parameter of the

[1] We used LIBSVM [7] in this study.

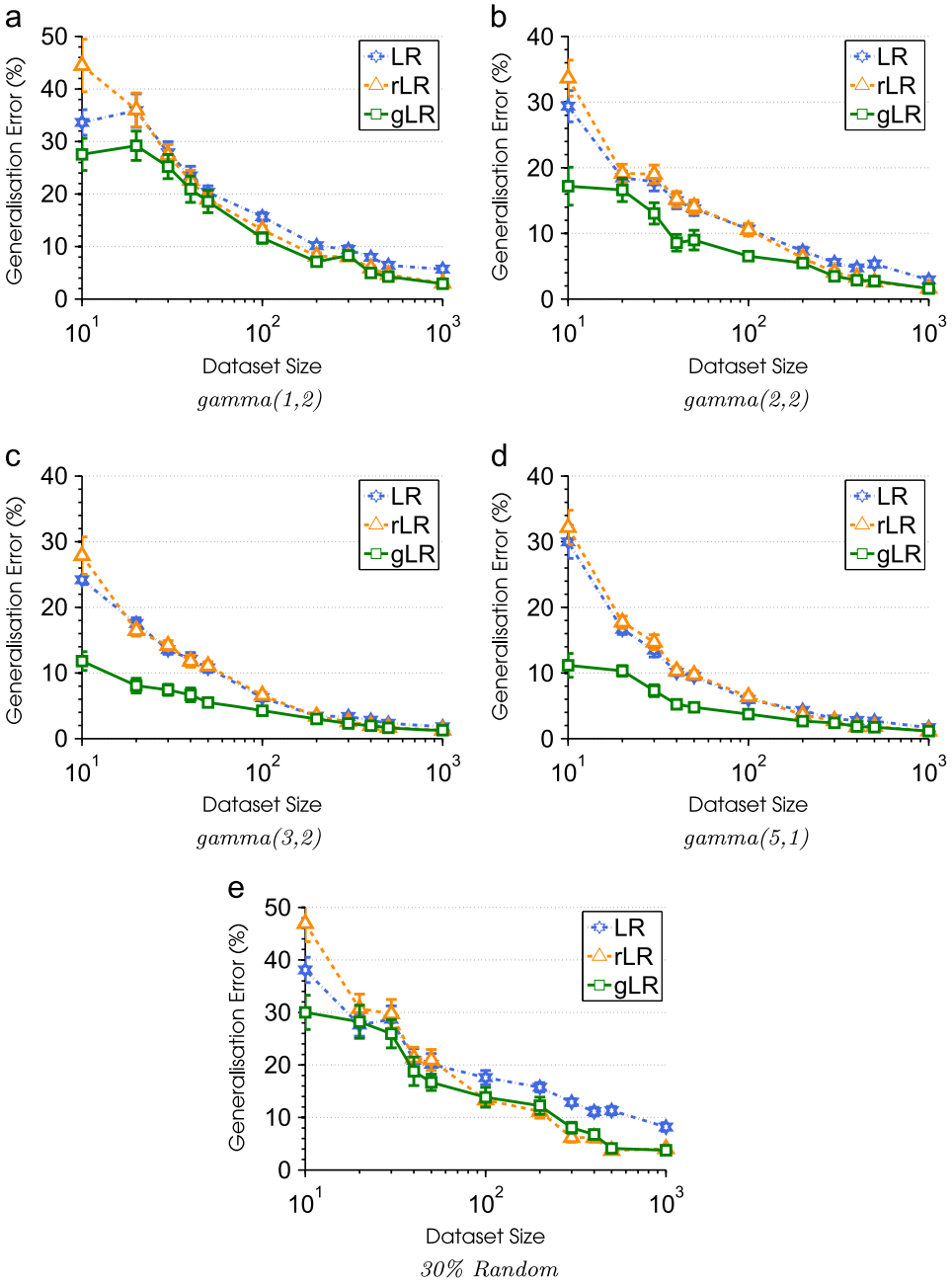

Fig. 5. Mean classification errors and their standard errors on 10-dimensional, 1.5-separated generatively generated data on five noise types with varying dataset size. The proposed gLR clearly outperforms traditional LR and is superior to rLR in almost all data configurations.

SVM. We considered possible values of those hyper-parameters in the range of 2^i where $i \in \{-10, 9, \ldots, 0, \ldots, 9, 10\}$.

3.4. Results: Real-world data

The experimental results are all summarised in Table 2. We first discuss the 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 Table 2 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 on non-random noise cases but still lags behind gLR in general. The LR+shift is more robust than rLR on all types of noises studied. However, it struggled in the case of random noise compared to gLR. The SVM performed surprisingly well on this dataset. Its performances closely match with those of gLR. The rNDA performed rather poorly with classification errors over 20%. One of the reasons might be that the distribution of the data does not fit the Gaussian assumption made by the model. The result also hints that as far as the classification is concerned, a discriminative model is preferable to a generative model.

Next, we shall discuss the results from the datasets from the bio-medical domain namely *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 in Table 2, the gLR employing the proposed generalised label noise model performs better than the other robust classifiers and the SVM, which ranked second in the previous dataset, in almost all

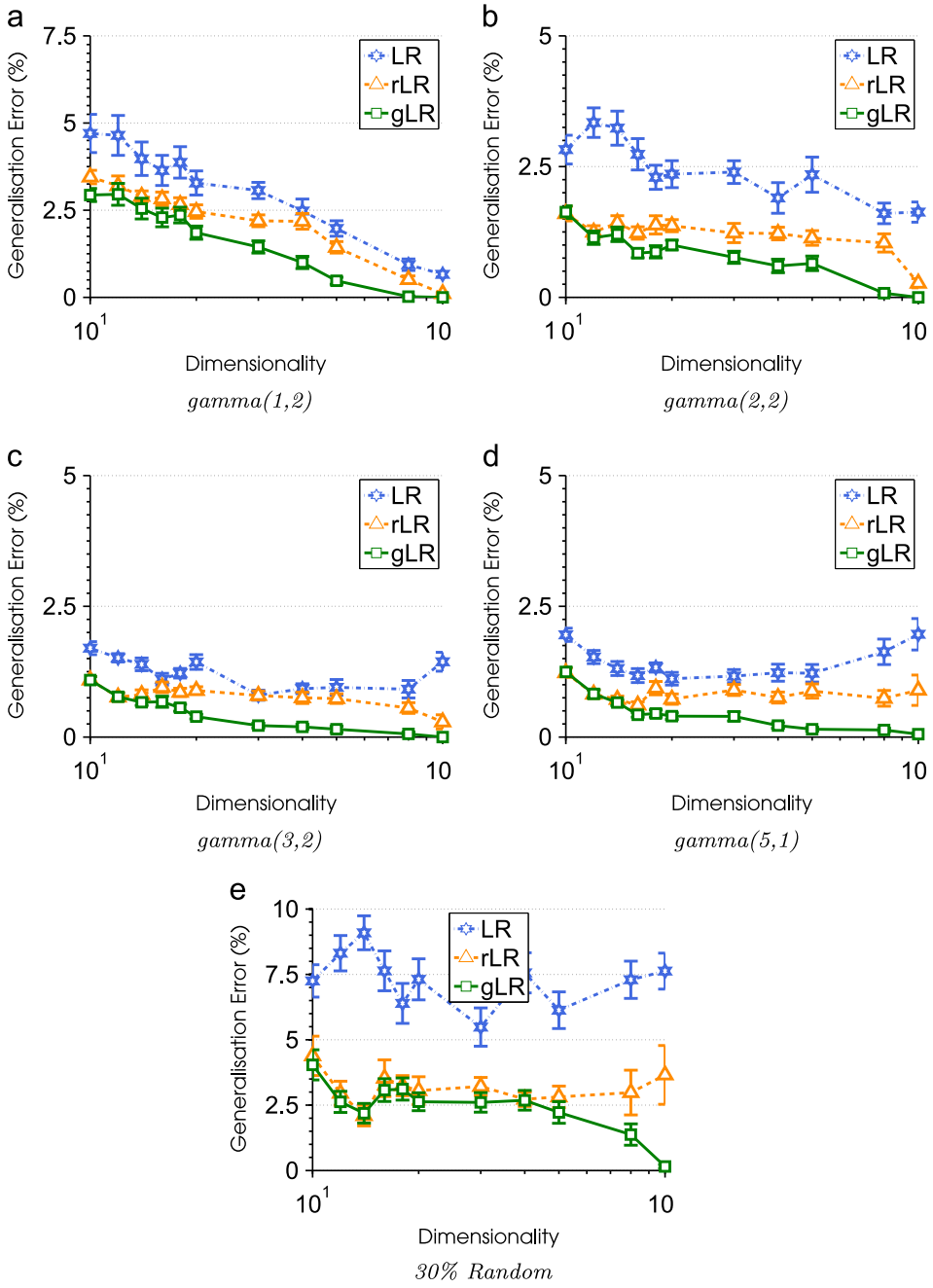

Fig. 6. Mean classification errors and their standard errors on 1.5-separated generatively generated data with 1000 data points on five noise types with varying data dimensionality. The proposed model again show better generalisation performance in all cases tested.

types of noise except the 'Original' and 'Random' *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. Lastly, we present the results from *Leukaemia* dataset. We note that the only mislabelling in this dataset is however not backed up by biological evidence but rather was identified by a consensus of noise detecting algorithms in the literature. Since the contamination rate for this dataset is quite low, it is expected that this will not interfere much with the learning of the competing algorithms. Indeed, from the results in Table 2, we see that on the original data all of the algorithms except rLR perform reasonably well. Still, the proposed gLR ranks first among all. The proposed method continues to show more stability as artificial noises are injected into the dataset, except in the presence of random label noise. LR+shift which has a potential to counteract non-random noise also performed very well as expected. Compared to the classifiers employing the original latent variable model (rLR and rNDA), we see again the superiority of the generalised label noise model over the existing model in non-random noise cases.

3.5. Detecting noisy labels

Finally, in our series of empirical evaluations we test the classifiers' abilities to detect mislabelled examples in the datasets. The detection method is based on calculating a mislabelling probability, $p(\hat{y} \neq \tilde{y}|\mathbf{x}, \mathbf{w})$, where $\hat{y}$ denotes the predicted label. The probability can be used in a hard threshold rule, i.e., to predict that the example was mislabelled if $p(\hat{y} \neq \tilde{y}|\mathbf{x}, \mathbf{w}) > 0.5$ or it can be used

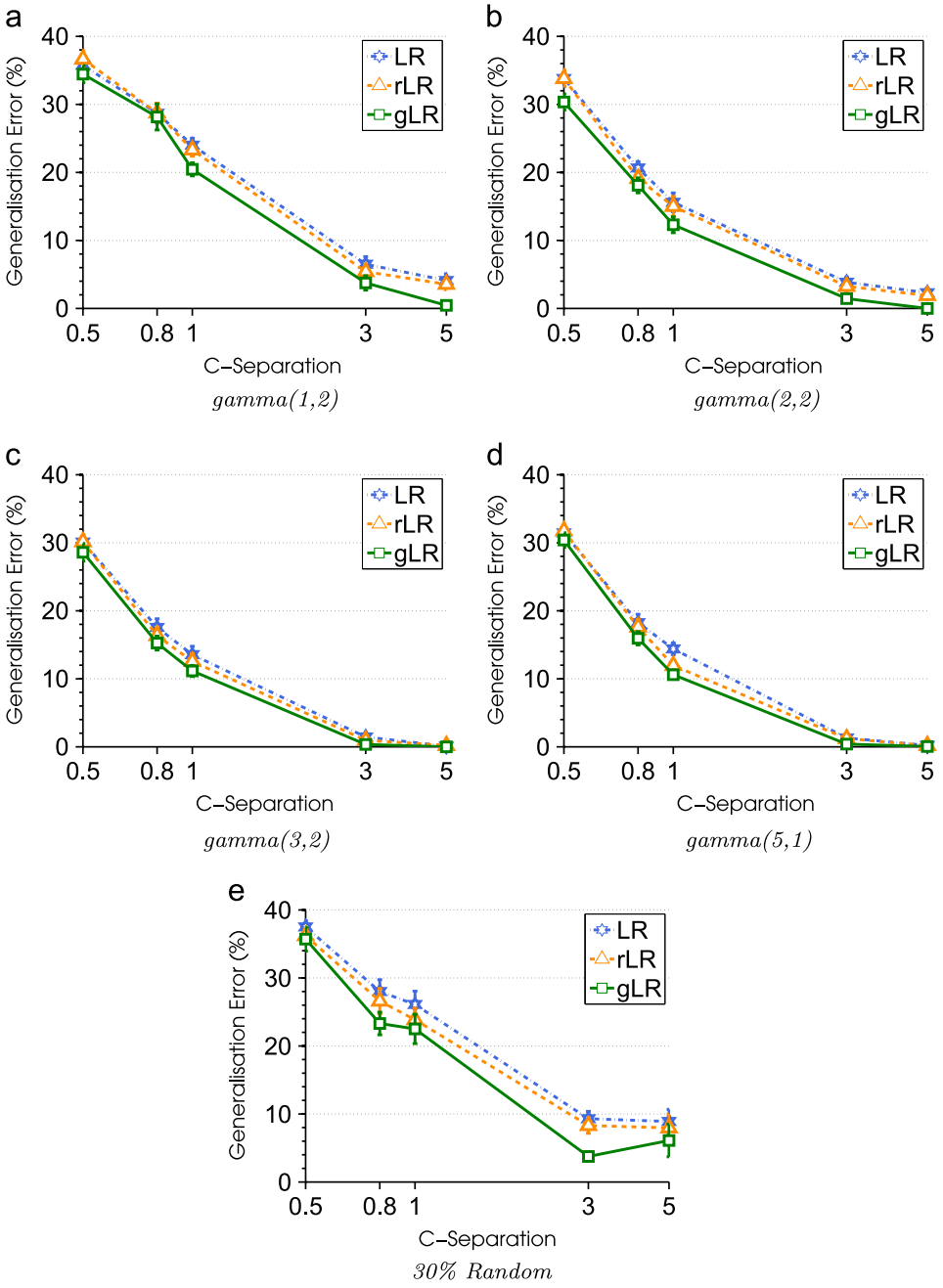

Fig. 7. Mean classification errors and their standard errors on 10-D generatively sampled datasets with 100 datapoints on five noise types with varying class separation. The proposed gLR is superior to the competing algorithms in almost all cases. The only scenario where the benefit of the proposed model is less significant is when the classes are highly overlapped.

Table 1
The characteristics of the datasets employed in this study. All of the identified mislabellings are backed up by biological/supporting evidence except for *Leukaemia* dataset.

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
Leukaemia	25(AML)/47(ALL)	1/0	7129
Websearch	515(bike)/515(not bike)	100/83	1318

as a degree of belief that the example was mislabelled. A good way to summarise the detection performance is by constructing the Receiver Operating Characteristic (ROC) curves. The area under ROC curves indicates the probability that a randomly drawn and mislabelled example would be flagged by the proposed algorithm.

Fig. 8 presents the ROC curves for the original *Websearch*, the original *Colon* and the original *Breast* datasets. Superimposed for reference are the ROC curves that correspond to LR+shift, the algorithm which ranked second overall in the previous experiment and rLR, a robust logistic regression employing the latent variable model. It is clear from the results that the newly proposed model is superior to the existing one. The gap between gLR and rLR is well apparent in all cases tested. The LR+shift algorithm performs reasonably well on *Websearch* and *Colon* but it lags far behind gLR in *Breast* dataset. These all demonstrate the advantage of our modelling approach over existing methods.

Overall, we can conclude based on these empirical evidences that the proposed generalised label noise model helps counteracting the

Table 2
Mean generalisation errors (%) on 4 real-world datasets with inherent label noise together with their standard errors. A statistical test was performed using Friedman test at 5%-level. Underlined entries rank first among the competing algorithms. The proposed method tops the evaluation with an average rank of 1.67.

Dataset	NoiseType	rLR	rNDA	SVM	LR+shift	gLR
Colon	Original	5.77 ± 0.35	25.00 ± 0.67	12.69 ± 0.38	9.23 ± 0.41	$\underline{3.08 \pm 0.29}$
	g(1,2)	8.85 ± 0.49	$\underline{6.54 \pm 0.34}$	13.08 ± 0.50	8.46 ± 0.48	6.92 ± 0.43
	g(2,2)	10.38 ± 0.47	11.15 ± 0.29	13.46 ± 0.46	7.69 ± 0.28	$\underline{6.15 \pm 0.30}$
	g(3,2)	8.08 ± 0.34	10.77 ± 0.52	14.23 ± 0.55	10.77 ± 0.44	$\underline{5.38 \pm 0.25}$
	g(5,1)	7.69 ± 0.25	10.38 ± 0.44	14.23 ± 0.31	7.31 ± 0.44	$\underline{5.77 \pm 0.30}$
	Random	18.85 ± 0.56	37.69 ± 0.54	27.31 ± 0.36	20.38 ± 0.63	$\underline{12.69 \pm 0.64}$
Breast	Original	32.00 ± 0.60	13.00 ± 0.67	$\underline{9.00 \pm 0.69}$	18.00 ± 0.64	14.00 ± 0.73
	g(1,2)	33.00 ± 0.75	24.00 ± 0.77	14.00 ± 0.80	12.00 ± 0.68	$\underline{5.00 \pm 0.55}$
	g(2,2)	36.00 ± 0.41	19.00 ± 0.83	17.00 ± 0.88	15.00 ± 0.64	$\underline{13.00 \pm 0.75}$
	g(3,2)	29.00 ± 0.69	24.00 ± 0.70	19.00 ± 0.83	13.00 ± 0.81	$\underline{10.00 \pm 0.69}$
	g(5,1)	33.00 ± 0.59	27.00 ± 0.88	19.00 ± 0.83	20.00 ± 0.56	$\underline{15.00 \pm 0.79}$
	Random	31.00 ± 0.69	26.00 ± 0.80	$\underline{23.00 \pm 0.81}$	27.00 ± 0.75	25.00 ± 0.79
Leukaemia	Original	32.33 ± 0.54	12.33 ± 0.38	15.67 ± 0.55	6.67 ± 0.37	6.67 ± 0.29
	g(1,2)	32.67 ± 0.44	12.00 ± 0.38	12.67 ± 0.30	$\underline{6.33 \pm 0.33}$	8.33 ± 0.37
	g(2,2)	34.00 ± 0.49	11.00 ± 0.31	13.00 ± 0.57	10.33 ± 0.48	$\underline{8.67 \pm 0.38}$
	g(3,2)	34.33 ± 0.45	12.33 ± 0.39	11.67 ± 0.42	$\underline{11.33 \pm 0.53}$	12.33 ± 0.29
	g(5,1)	32.00 ± 0.48	12.00 ± 0.49	12.67 ± 0.47	$\underline{8.33 \pm 0.42}$	10.33 ± 0.54
	Random	29.33 ± 0.49	24.67 ± 0.62	25.33 ± 0.54	$\underline{24.00 \pm 0.74}$	28.33 ± 0.56
Websearch	Original	17.65 ± 0.18	20.05 ± 0.16	15.22 ± 0.13	14.81 ± 0.11	$\underline{14.73 \pm 0.12}$
	g(1,2)	14.20 ± 0.09	19.66 ± 0.18	$\underline{13.28 \pm 0.12}$	13.62 ± 0.11	13.33 ± 0.09
	g(2,2)	15.24 ± 0.15	22.52 ± 0.25	$\underline{13.74 \pm 0.14}$	14.78 ± 0.14	13.81 ± 0.10
	g(3,2)	15.19 ± 0.15	22.74 ± 0.22	13.91 ± 0.11	14.08 ± 0.11	$\underline{13.33 \pm 0.11}$
	g(5,1)	16.63 ± 0.20	23.93 ± 0.27	14.05 ± 0.19	$\underline{13.20 \pm 0.10}$	13.25 ± 0.13
	Random	27.65 ± 0.38	38.69 ± 0.32	18.25 ± 0.24	21.97 ± 0.22	18.25 ± 0.25
Average rank		4.17	3.75	3.08	2.33	1.67
P-value		4.61×10^{8}				

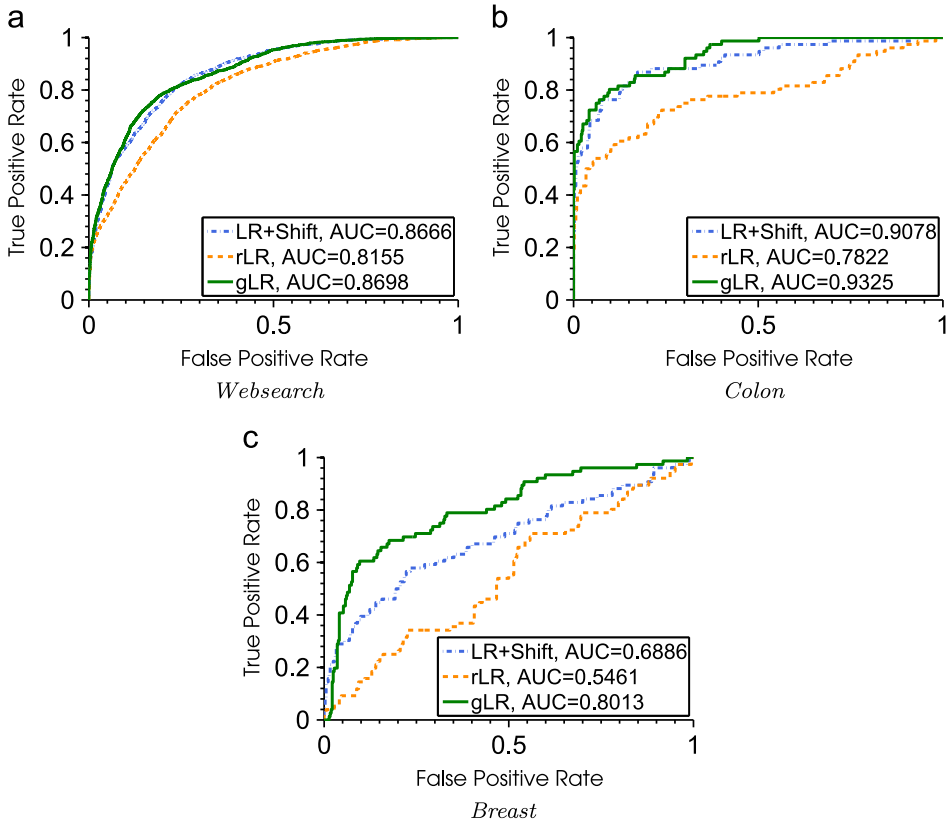

Fig. 8. Mislabelling detection capability of LR+Shift, rLR and the proposed gLR as summarised by the Receiver Operating Characteristics curves. The area under ROC curves quantified the superiority of the proposed method over the existing robust classifiers.

negative effect of various types of noises including random noise, and that the existing random noise model is less suitable for classification tasks where non-random label noise is present.

4. Conclusion

We presented a novel label noise model for classification where the training labels are inaccurate. The model is a generalisation of the existing latent variable model developed for random label noise. Unlike the existing methods, 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.

Acknowledgements

The author would like to thank Ata Kabán for helpful discussions. This work is supported by the Faculty of Science, Chiang Mai University. Department of Computer Science at Chiang Mai University provides computing facilities.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gishdagger, S. Ybarradagger, D. Mackdagger, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. USA 96 (12) (1999) 6745–6750.
[2] E. Beigman, B.B. Klebanov, Learning with annotation noise, in: ACL 2009, in: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2–7 August 2009, Singapore, 2009.
[3] J. Bootkrajang, A. Kabán, Multi-class classification in the presence of labelling errors, in: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, (ESANN 2011), 2011.
[4] J. Bootkrajang, A. Kabán, Label-noise robust logistic regression and its applications, in: ECML-PKDD'12, 2012.
[5] J. Bootkrajang, A. Kabán, Classification of mislabelled microarrays using robust sparse logistic regression, Bioinformatics 29 (7) (2013) 870–877.
[6] J. Bootkrajang, A. Kabán, Learning kernel logistic regression in the presence of class label noise, Pattern Recognit. 47 (11) (2014) 3641–3655.
[7] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27, ⟨http://www.csie.ntu.edu.tw/cjlin/libsvm⟩.
[8] R.S. Chhikara, J. McKeon, Linear discriminant analysis with misallocation in training samples, J. Am. Stat. Assoc. 79 (388) (1984) 899–906.
[9] E. Cohen, Learning noisy perceptrons by a perceptron in polynomial time, in: 38th Annual Symposium on Foundations of Computer Science, FOCS '97, Miami Beach, Florida, USA, October 19–22, 1997, 1997.
[10] E. Côme, L. Oukhellou, T. Denoeux, P. Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern Recognit. 42 (3) (2009) 334–348.
[11] S. Dasgupta, Learning mixtures of Gaussians, in: Proceedings of the 40th Annual Symposium on Foundations of Computer Science, (FOCS'99), 1999.
[12] T. Denoeux, A k-nearest neighbor classification rule based on dempster-shafer theory, IEEE Trans. Syst. Man and Cybern. 25 (5) (1995) 804–813.
[13] T. Denoeux, Analysis of evidence-theoretic decision rules for pattern classification, Pattern Recognit. 30 (7) (1997) 1095–1107.
[14] T. Denoeux, A neural network classifier based on dempster-shafer theory, IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. 30 (2) (2000) 131–150.
[15] B. Frénay, G. de Lannoy, M. Verleysen, Label noise-tolerant hidden Markov models for segmentation: Application to ECGs, in: ECML-PKDD'11, 2011.
[16] B. Frénay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Trans. Neural Netw. Learn. Syst. 25 (5) (2014) 845–869.
[17] M.J. García-Zattera, T. Mutsvari, A. Jara, D. Declerck, E. Lesaffre, Correcting for misclassification for a monotone disease process with an application in dental research, Stat. Med. 29 (30) (2010) 3103–3117.
[18] A. Ghosh, N. Manwani, P.S. Sastry, Making risk minimization tolerant to label noise, Neurocomputing 160 (2015) 93–107.
[19] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[20] B.B. Klebanov, E. Beigman, From annotator agreement to noise models, Comput. Linguist. 35 (4) (2009) 495–503.
[21] A. Kolcz, G.V. Cormack, Genre-based decomposition of email class noise, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28–July 1, 2009, 2009.
[22] P.A. Lachenbruch, Discriminant analysis when the initial samples are misclassified, Technometrics 8 (4) (1966) 657–662.
[23] P.A. Lachenbruch, Discriminant analysis when the initial samples are misclassified ii: non-random misclassification models, Technometrics 16 (3) (1974) 419–424.
[24] N.D. Lawrence, B. Schölkopf, Estimating a Kernel Fisher Discriminant in the Presence of Label Noise, in: ICML'01, Morgan Kaufmann, Williamstown, Massachusetts, United States, 2001.
[25] G. Lugosi, Learning with an unreliable teacher, Pattern Recognit. 25 (1992) 79–87.
[26] A. Malossini, E. Blanzieri, R.T. Ng, Detecting potential labeling errors in microarrays by data perturbation, Bioinformatics 22 (17) (2006) 2114–2121.
[27] M. Martin-Merino, A kernel svm algorithm to detect mislabeled microarrays in human cancer samples, in: IEEE 13th International Conference on Bioinformatics and Bioengineering (BIBE), 2013, 2013.
[28] N. Natarajan, I.S. Dhillon, P.K. Ravikumar, A. Tewari, Learning with noisy labels, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 26, Curran Associates, Inc., Lake Tahoe, Nevada, United States, 2013.
[29] V.C. Raykar, S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, L. Moy, Learning from crowds, J. Mach. Learn. Res. 11 (2010) 1297–1322.
[30] R. Snow, B. O'Connor, D. Jurafsky, A.Y. Ng, Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks., in: EMNLP, 2008.
[31] J. Tibshirani, C.D. Manning, Robust logistic regression using shift parameters, CoRR abs/1305.4987.
[32] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson Jr., J.R. Marks, J.R. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, Proc. the Natl. Acad. Sci. USA 98 (20) (2001) 11462–11467.
[33] W. Zhang, Y.W. Wan, G. Allen, K. Pang, M. Anderson, Z. Liu, Molecular pathway identification using biological network-regularized logistic models, BMC Genom. 14 (Suppl 8) (2013) S7.
[34] L. Zhenqiu, J. Feng, T. Guoliang, W. Suna, S. Fumiaki, S.J. Meltzer, T. Ming, Sparse logistic regression with Lp penalty for biomarker identification, Stat. Appl. Genet. Mol. Biol. 6 (1) (2007) 1–22.

Jakramate Bootkrajang is a Lecturer in the Department of Computer Science of Chiang Mai University. He received his B.Sc. (2007) and M.Sc. (2009) degrees in Computer Science from Seoul National University, Republic of Korea and Ph.D. (2013) degree in Computer Science from the University of Birmingham. His interests concern statistical machine learning and probabilistic modelling of data. His current research focuses on supervised learning from unreliable annotated data.