OUTLIER ESTIMATION AND DETECTION
APPLICATION TO SKIN LESION CLASSIFICATION

S. Sigurdsson, J. Larsen and L. K. Hansen*

Informatics and Mathematical Modelling
Technical University of Denmark
Richard Petersens Plads, Building 321
DK-2800 Kongens Lyngby, Denmark
{siggi,jl,lhansen}@imm.dtu.dk

P. A. Philipsen and H. C. Wulf

Department of Dermatology
Bispebjerg Hospital
University of Copenhagen
Bispebjerg Bakke 23
DK-2400 Copenhagen, Denmark

ABSTRACT

We extend MacKays Bayesian approach to neural classifiers to include an outlier detector mechanism. We show that the outlier detector can locate misclassified samples.

1. INTRODUCTION

Multi-layer perceptron networks posses powerful approximation capabilities and when used for classification they can adapt to arbitrarily complex posterior probability functions. Such extreme flexibility calls for careful control of overfit. In previous work we have designed resampling based tools for control of overfit and outlier detection. Overfit control is aimed at regularization, typically using weight decay, i.e., controlling the roughness of decision surfaces so that they not get too rough in the face of noise in finite samples. Outlier detection, on the other hand, is aimed at modeling and controling random label noise that can lead to wrong decision surface *topologies* by creating isolated “islands” of the wrong class. In this work we develop overfit and outlier control in a Bayesian (MLII) setting. The potential advances of the Bayesian approach are

- For limited data sets, as typically appear in medical applications, resampling based approaches are problematic, because of the poor statistics of outliers. If there are only a few outliers in the total sample, results will be highly dependent on the distribution of the outliers among the training and validation sets.
- The Bayesian approach requires less computation because it avoids multiple training sessions inherent to cross-validation procedures.
- The Bayesian approach avoids the open issue of resampling split ratio.

2. OUTLIER PROBABILITY IN CLASSIFICATION

We aim at modeling the posterior probability functions for multi-classification given by $p(\mathcal{C}_k|\mathbf{x})$, $k = 1, 2, \ldots, c$, where $\mathbf{x}$ is the input feature vector with dimension I, $\mathcal{C}_k$ is the corresponding class label and c is the number of classes.

Outliers are defined as an input pattern having the corresponding target class label erroneously “flipped” to another class. An example of this could be in skin lesion classification, where samples are labeled by histological examination. If the sample for some reason is erroneously registered, the label can have a random relation to the input pattern. Hence, we defined a probability ε of being assigned with random target label. The outlier probability $\varepsilon = [0, 1]$ is assumed to be independent of both “true” class label and input pattern value.

The posterior probability distribution has been previously formulated [1]

$$p(\mathcal{C}_l|\mathbf{x}) = p_0(\mathcal{C}_l|\mathbf{x})(1-\varepsilon) + \frac{\varepsilon}{c-1}\sum_{k=1,k\neq l}^{c} p_0(\mathcal{C}_k|\mathbf{x}) \qquad (1)$$

where $p_0(\mathcal{C}_l|\mathbf{x})$ is the posterior probability with zero outlier probability. The first term in equation 1 is the probability that the input pattern $\mathbf{x}$ is not an outlier, while the second term is the outlier contribution coming from classes other than $\mathcal{C}_l$. Defining a scaled outlier probability $\beta = \varepsilon/(c-1)$, equation 1 can be rewritten as

$$p(\mathcal{C}_l|\mathbf{x}) = p_0(\mathcal{C}_l|\mathbf{x})(1-\beta c) + \beta \qquad (2)$$

where $\beta = [0; 1/(c-1)]$.

3. NETWORK ARCHITECTURE AND INFERENCE

In following we will represent probabilities with a two-layer feed-forward neural network with I inputs given by

$$h_j(\mathbf{x}) = \tanh\left(\sum_{i=1}^{I} w_{ji}x_i + w_{j0}\right) \qquad (3)$$

where w_{j0} is the bias and $h_j(\mathbf{x})$ is the output of the jth sigmoidal activation function of the hidden layer. Network output k of the output layer is given by

$$y_k(\mathbf{x}) = \sum_{j=1}^{H} w_{kj}h_j(\mathbf{x}) + w_{k0} \qquad (4)$$

where H is the number of units in the hidden layer. To be able to interpret the outputs as estimates of the posterior probabilities $\hat{p}(\mathcal{C}_k|\mathbf{x})$ we use a modified version of SoftMax [2]. The standard SoftMax [3] has dependency between the weights as the outputs

*This work is supported by the Danish Research Councils through the project Signal and Image Processing for Telemedicine (SITE).

always sum to one, which causes problems in the evaluation of the inverse Hessian. The modified SoftMax solves the problem by using only $c-1$ outputs and is given by

$$\hat{p}_0(\mathcal{C}_k|\mathbf{x}) = \frac{\exp(y_k(\mathbf{x}))}{1+\sum_{k'=1}^{c-1} y_{k'}(\mathbf{x})}, \quad k=1,2,\ldots,c-1 \tag{5}$$

and the probability for the last class is easily evaluated with

$$\hat{p}_0(\mathcal{C}_c|\mathbf{x}) = 1 - \sum_{k=1}^{c-1} \hat{p}_0(\mathcal{C}_k|\mathbf{x}). \tag{6}$$

An estimate of the outlier modified posterior probability is given by $\hat{p}(\mathcal{C}_l|\mathbf{x}) = \hat{p}_0(\mathcal{C}_l|\mathbf{x})(1-\beta c)+\beta$ from equation 2.

The data set for the supervised training of the model is given by the input-output pairs $\mathcal{D} = \{\mathbf{x}^{(n)}, \mathbf{t}^{(n)}\}, \ n=1,2,\ldots,N$ where $\mathbf{t}^{(n)}$ is the one-of-c coded target value vector given by

$$t_k^{(n)} = \begin{cases} 1 & \text{if } \mathbf{x}^{(n)} \in \mathcal{C}_k \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $k=1,2,\ldots,c$. To simplify notation we define the network weight vector as $\mathbf{w}$, holding all weights.

To estimate the weights we invoke the approach proposed by David MacKay [4, 5]. The posterior probability of the parameters $\mathbf{w}$ can be written as

$$p(\mathbf{w}|\mathcal{D},\alpha,\beta) = \frac{p(\mathcal{D}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)}{p(\mathcal{D}|\alpha,\beta)} \tag{8}$$

where $p(\mathcal{D}|\mathbf{w},\beta)$ is the likelihood, $p(\mathbf{w}|\alpha)$ is the prior, $p(\mathcal{D}|\alpha,\beta)$ is the evidence. The α and β are hyperparameters, i.e., regularization parameter and scaled outlier probability respectively, both assumed to be known when inferring the weights. For a classification problem with multiple classes the choice of likelihood is $p(\mathcal{D}|\mathbf{w},\beta) = \exp[-E_D(\mathbf{w},\beta)]$ where

$$E_D(\mathbf{w},\beta) = -\sum_{n=1}^{N}\sum_{k=1}^{c} t_k^{(n)} \ln(\hat{p}(\mathcal{C}_k|\mathbf{x})) \tag{9}$$

is the cross-entropy error function [6]. The prior is given by

$$p(\mathbf{w}|\alpha) = \frac{\exp[-\alpha E_W(\mathbf{w})]}{Z_W(\alpha)} \tag{10}$$

where $Z_W(\alpha) = \int \exp -\alpha E_W(\mathbf{w}) d\mathbf{w}$ is a normalization factor and $E_W(\mathbf{w})$ is a regularization function, given by

$$E_W(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{W} w_i^2 \tag{11}$$

where W is the number of weights in the network. This is a zero mean Gaussian prior, better known as weight decay.

The optimization of the weights is done by minimizing a cost function, $S(\mathbf{w}) \propto -\ln p(\mathbf{w}|\mathcal{D},\alpha,\beta)$, given by

$$S(\mathbf{w}) = E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}). \tag{12}$$

where weight independent terms have been omitted. The weights are optimized using a Gauss-Newton scheme [7] given by

$$\mathbf{w}^{new} = \mathbf{w}^{old} - \eta \mathbf{A}^{-1}(\mathbf{w}^{old})\mathbf{g}(\mathbf{w}^{old}) \tag{13}$$

where $\mathbf{g}(\mathbf{w}) = \partial S(\mathbf{w})/\partial \mathbf{w}$ is the gradient of the cost function with respect to the weights, $\mathbf{A}(\mathbf{w})$ is the Gauss-Newton approximation of the Hessian matrix and η is the step size, determined by line search. See [1] for details.

3.1. Adapting the hyperparameters

The posterior distribution for the hyperparameters is given by

$$p(\alpha,\beta|\mathcal{D}) = \frac{p(\mathcal{D}|\alpha,\beta)p(\alpha,\beta)}{p(\mathcal{D})}. \tag{14}$$

We assign a uniform prior over hyperparameters $p(\alpha,\beta)$ and thus make the so-called evidence approximation using the evidence $p(\mathcal{D}|\alpha,\beta)$ to evaluate $p(\alpha,\beta|\mathcal{D})$. For details on this approximation see [8]. The evidence can be evaluated with the Laplace approximation

$$\begin{aligned} p(\mathcal{D}|\alpha,\beta) &= \int p(\mathcal{D}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w} & (15)\\ &= \frac{1}{Z_W(\alpha)}\int \exp[-S(\mathbf{w})]d\mathbf{w} & (16)\\ &\approx \frac{e^{-S(\mathbf{w}_{\text{MP}})}(2\pi)^{W/2}|\mathbf{A}|^{-1/2}}{Z_W(\alpha)} & (17)\end{aligned}$$

where $\mathbf{w}_{\text{MP}}$ is maximizes the product $p(\mathcal{D}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)d\mathbf{w}$.

Finding $\hat{\beta}$, an estimate of β, is done by minimizing

$$\begin{aligned} C(\beta) &\propto -\ln p(\mathcal{D}|\alpha,\beta) & (18)\\ &= S(\mathbf{w}_{\text{MP}}) + \frac{1}{2}|\mathbf{A}| & (19)\end{aligned}$$

where terms independent of β have been omitted. We suggest using Brent's method [9], approximating $C(\beta)$ as a quadratic function to find $\hat{\beta}$. This is possible as $C(\beta)$ is a smooth function and we have an upper and lower bound on β setting the range for the search of $\hat{\beta}$. As Brent's method does not use gradient information we avoid evaluating $\partial S(\mathbf{w})/\partial\beta$ which has the unpleasant property of zero denominator when $\beta=0$.

The α is computed as in [4], by maximizing $\ln p(\mathcal{D}|\alpha,\beta)$, evaluating $\partial \ln p(\mathcal{D}|\alpha,\beta)/\partial\alpha$ which gives the following update formula

$$\alpha^{new} = \frac{\gamma}{2E_W(\mathbf{w}_{\text{MP}})} \tag{20}$$

where $\gamma = W - \alpha \text{Trace}\mathbf{A}^{-1}$ is the effective number of weights in the network.

A practical approach to adapting the regularization parameter would be to train the weights and update the α and β when the weights have converged. This is repeated cyclically until the regularization parameters have converged.

3.2. Outlier detection

Having estimated the network from the data, it is possible to evaluate the outlier probability labeled examples

$$p_{outlier} = \frac{\beta(1-p_0(\mathcal{C}_l|\mathbf{x}))}{p_0(\mathcal{C}_l|\mathbf{x})(1-\beta c)+\beta} \tag{21}$$

where $\mathcal{C}_l$ is the target label class, see [1] for details. This leads to the estimate $\hat{p}_{outlier} = \hat{\beta}(1-\hat{p}_0(\mathcal{C}_l|\mathbf{x}))/\hat{p}(\mathcal{C}_l|\mathbf{x})$. To make decisions we threshold the value of $\hat{p}_{outlier}$ at 0.5.

4. EVALUATION

The performance of the outlier model is first tested on a toy problem. The toy problem has $c = 3$ classes defined in a 2D input space. The class conditional probabilities are $p(\mathcal{C}_k|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{I})$ where $\boldsymbol{\mu}_1 = [0, 2]$, $\boldsymbol{\mu}_2 = [-1.5, -1]$, $\boldsymbol{\mu}_3 = [1.5, -1]$ and $\mathbf{I}$ is the identity matrix. The prior class probabilities are $p(\mathcal{C}_k) = 1/3$. The number of training examples is $N = 300$ and also 3000 independent data points for testing are generated. Outliers are introduced to the training data by flipping the labels at random with flip rate r. In Figure 1 we show results averaged over 100 independently generated data sets. The network is initialize with $H = 3$ hidden units. The outlier model presents considerably better performance for large outlier rates.

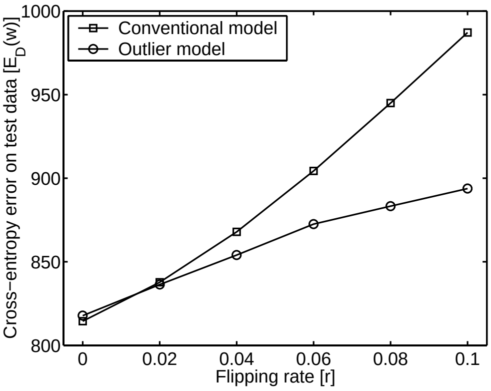

Fig. 1. The figure shows the average cross-entropy error on test data using the toy data with different flipping rates. The cross-entropy error for the outlier model was evaluated with outlier probability $\hat{\beta} = 0$ which gives the same error function. The outlier model has significantly less cross-entropy error when the data is corrupted when compared to a conventional model.

In Figures 2 and 3 we take a closer look at the outlier assignment for the toy problem. Some of the flipped labels belong to patterns near decision boundaries (see Figure 3), where we find low posterior probabilities for the target label class. These patterns are accounted for by the noise model of the underlying decision problem (overlapping classes), hence, should not be detected as outliers. With this in mind we expect the estimated outlier rate to be less than the flip rate. This is confirmed in Figure 4.

5. SKIN LESION CLASSIFICATION

Malignant Melanoma (MM) is a lethal skin cancer developing from pigmented skin lesions. The cancer is most lethal if it enters the bloodstream, hence it is important to diagnose MM at an early stage. Diagnosing MM is not trivial, as many common skin lesions resemble MM visually. Two studies show that trained dermatologists diagnoses clinically 63% [10] and 75% [11] of cancerous skin lesions correctly. A study at *Karolinske Hospital, Stockholm, Sweden* shows that dermatologists with less than one year experience diagnose only 31% correctly.

New incidences of MM in Denmark, e.g., has increased 5- to 6-fold from 1942 to 1982 while the mortality rate has been doubled from 1955 to 1982 [12]. Currently, approximately 800 cases of malignant melanoma are reported in Denmark every year. In Germany $9000 - 10000$ new cases are expected every year with an annual increase of $5 - 10\%$ [13].

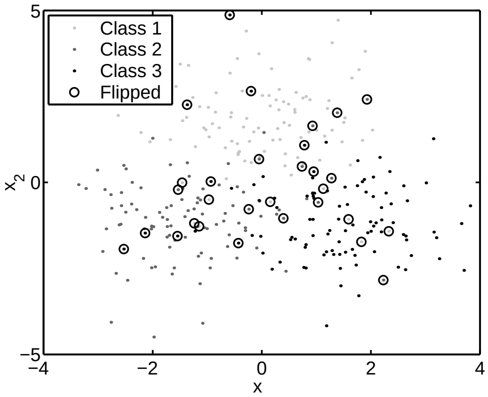

Fig. 2. The figure shows one of the generated data sets with data points flipped at a rate of $r = 0.1$.

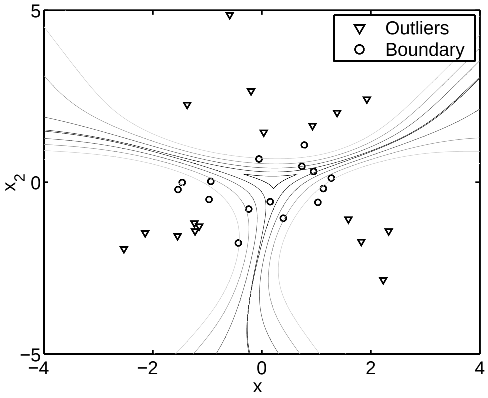

Fig. 3. The contour shows the maximum posterior probability for all the classes, estimated with an outlier model using the data in figure 2. The outlier model easily detects the flipped data points that are not too close to the decision boundary. Flipped data points near the boundary are confused with noisy data points and cannot detected as outliers.

Taking a biopsy of every suspicious lesion and using a histological analysis is not acceptable for patients with multiple lesions and is also costly and time consuming. Taking biopsies in sensitive places like the face could produce scars. Automatic classification of skin cancer using machine learning techniques could help both dermatologist and non-dermatologist to diagnose an early stage of MM. A non-invasive method like Raman spectroscopy can probe the tissue biochemistry in the lesion and discriminate between lesion types. Raman spectra are obtained by pointing a laser beam in vitro or in vivo. The laser can excite molecular vibrations causing reflected beam with a spectrum of frequencies. This is called the Raman effect. The frequency shifts are dependent on the type of molecules in the sample and the Raman spectra holds therefore information about the local biochemistry.

The outlier model was applied to classification of skin lesions. The data set consists of $c = 5$ different classes of skin lesions where each of the $N = 177$ lesions is represented with a Ra-

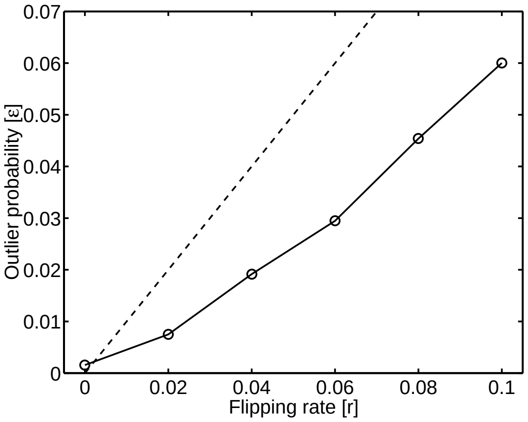

Fig. 4. The figure shows the outlier probability $\hat{\varepsilon}$ of the outlier model as a function of the flipping rate of the toy data. When the data has no outliers the outlier model estimates a low outlier probability, but in general the outlier probability is estimated lower than the flipping rate.

man spectrum of 1711 points. The input space is reduced with principal components analysis where the $I = 25$ largest principal components are used as network inputs. The network is initialized with $H = 20$ hidden units. In Figure 5 we show the Raman spectrum for an example detected as an outlier. The class label is pigmented nevi, however, it seems likely that this is an error since the the spectrum is more similar to the normal class.

6. CONCLUSION

We have extended the Bayesian MLII approach for neural network classification to incorporate an outlier model with an estimate of the outlier probability. The estimate of the outlier rate in a toy problem was shown to be conservative. In the context of skin lesion classification, the new scheme seems promising; a detected outlier seems indeed to be a misclassified example.

7. REFERENCES

[1] J. Larsen, L.N. Andersen, M. Hintz-Madsen, and L.K. Hansen, “Design of Robust Neural Network Classifiers,” in *Proceedings of the 1998 International Conference on Acoustics, Speech and Signal Processing*, New York, New York, 1998, pp. 1205–1208.

[2] L.N. Andersen, J. Larsen, L.K. Hansen, and M. Hintz-Madsen, “Adaptive Regularization of Neural Classifiers,” in *Proceedings of the 1997 IEEE Workshop on Neural Networks for Signal Processing VII*, J. Principe, L. Gile, N. Morgan, and E. Wilson, Eds., New York, New York, 1997, pp. 24–33.

[3] J.S. Bridle, “Probabilistic Interpretation of Feedforward Classification Network Outputs with Relationships to Statistical Pattern Recognition,” in *Neurocomputing - Algorithms, Architectures and Applications*, F. Fougelman-Soulie and J. Herault, Eds., vol. 6, pp. 227–236. Springer-Verlag, Berlin, 1990.

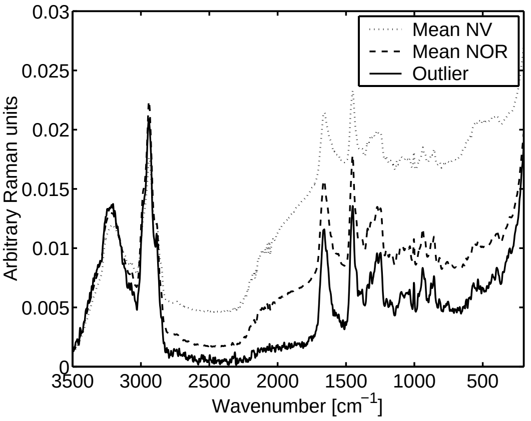

Fig. 5. The outlier model used on a practical medical problem, classifying skin lesions. The figure shows the mean Raman spectra over available training data for classes pigmented nevi (NV) and normal skin (NOR). The Raman spectra detected as an outlier is labeled to the NV class while the estimated posterior probability for the NOR and NV class is 0.98 and 0.02 respectively. Visual inspection shows that the background bias from 200cm^{-1} to 2000cm^{-1} for the outlier spectra is much similar to the NOR spectra than the NV spectra. This bias increases with the amount of pigmentation of the lesion and a NV lesion with almost no pigmentation is quite unlikely.

[4] D.J.C. MacKay, “A Practical Bayesian Framework for Backpropagation Networks,” *Neural Computation*, vol. 4, no. 3, pp. 448–472, 1992.

[5] D.J.C. MacKay, “The Evidence Framework Applied to Classification Networks,” *Neural Computation*, vol. 4, no. 5, pp. 720–736, 1992.

[6] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

[7] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, John Wiley & Sons, New York, New York, 1995.

[8] D.J.C. MacKay, “Comparison of Approximate Methods for Handling Hyperparameters,” *Neural Computation*, vol. 11, no. 5, pp. 1035–1068, 1999.

[9] W. Press, B. Flannery, S. Teukolsky, and W. Vettering, *Numerical Recipes in C: The Art of Scientic Computing*, Cambridge University Press, Cambridge, 1988.

[10] B. Lindelöf and M.A. Hedblad, “Accuracy in the Clinical Diagnosis and Pattern of Malignant Melanoma at a Dermatologic Clinic,” *The Journal of Dermatology*, vol. 21, no. 7, pp. 461–464, 1994.

[11] H.K. Koh, R.A. Lew, and M.N. Prout, “Screening for Melanoma/Skin Cancer,” *Journal of American Academy of Dermatology*, vol. 20, no. 2, pp. 159–172, 1989.

[12] A. Østerlind, *Malignant Melanoma in Denmark*, Ph.D. thesis, Danish Cancer Registry, Institute of Cancer Epidemiology, Denmark , 1990.

[13] G. Rassner, “Früherkennung des malignen Melanoms der Haut,” *Hausartz*, vol. 39, pp. 396–401, 1988.