Outliers and Bayesian Inference

Peter Sykacek
Austrian Research Institute for Artificial Intelligence
Schottengasse 3, A-1010 Vienna Austria
peter@ai.univie.ac.at

Abstract— **In this paper we report about an investigation in which we studied the properties of Bayes' inferred neural network classifiers in the context of outlier detection. The problem of misclassification due to outliers in the test data is seen as a serious problem in safety critical environments. We compare the usual way to deal with uncertainty in the Bayesian framework with a new approach based on the variance of the output layer activations and investigate the utility of both methods for outlier detection. The properties of both methods are visualized on a simple two dimensional classification problem. An investigation comparing both methods on some public data-sets with artificially constructed outlier patterns showed that a combination of the conventional method and the method proposed here should be used. These results where confirmed in a final experiment on real data, where a combination of both methods showed significantly better performance in rejecting outlying observations.**

I. Introduction

Neural networks are often used in safety-critical applications for regression or classification purpose. Since neural networks are unable to extrapolate into regions not covered by the training data (see [S. Roberts et. al.,1994]), one should not use their predictions in such regions. Consequently methods for outlier detection got a lot of attraction. In a recent publication [L. Tarassenko et. al., 1995] converted a classification problem to the problem of marking unusual inputs. Classification of tumors in mammograms was solved by marking unusual tissue for further inspection, a method known as novelty detection.

For conventional classification applications different methods have been used to be able to refuse classification of outlying patterns. In [S. Roberts et. al.,1994] outliers are rejected using an artificial class "outside" training data. Class "outlier" is assigned if the probability of a novel test pattern is largest for this artificial class. In [N. Schaltenbrand et. al., 1993] an instance based approach was used to represent the training inputs. The idea to allow for certainty of decisions based on whether the test input was well represented in the training data is very important, especially in a safety critical context like clinical diagnosis.

Suppose we are given k observations

$$(y_1, \underline{x}_1), (y_2, \underline{x}_2), ..., (y_k, \underline{x}_k)$$

as training data, where y_i denote the targets and $\underline{x}_i$ the corresponding inputs. Typical frequentist approaches assume that the test data is generated by the same distribution $p(y, \underline{x})$ the training data came from. An assumption which is misleading in many practical applications. There exist two sources of problems which may violate this assumption: Random or systematic errors may occur in the routine phase. The same is true if the problem is not sufficiently specified (That is: There exist additional possibilities of classification outcome, not specified in the training data). Especially the second problem has caused some prominent blunders [1]: Using linear discriminant analysis between teeth of *homo sapiens* and chimpanzees [J. Bronowski et. al., 1951] have tried to identify whether *Australopithecus africanus* is human or ape. Ignoring the possibility of "outliers" they found agreement with homo but not with ape. Later [C.R. Rao, 1960] pointed out that using the whole set of variables,both conclusions are implausible.

To overcome this problem, we should have knowledge of the whole distribution $p(y, \underline{x})$.

[1] The following example is a quote from [B.D. Ripley, 1996]

Probably the best way to do this is to use Bayesian inference to determine the model parameters. Bayesian inference regards the model parameters itself as random variables. In case of neural networks the solution is given by a probability distribution over network weights and biases.

During evaluation, this distribution over parameter space leads to predictive distributions over network outputs. In a classification task one typically integrates over this distribution, a method which is also known as marginalization. Test samples from regions with low density in the training data lead to distributions with large variance over output activations. Integration moderates the probabilities for all classes. Using doubt levels allows to refuse classifications not only in regions with overlapping class conditional densities but also from regions with low density of training data. This approach automatically incorporates confidence into probability estimates, it was used in [D.J.C. MacKay, 1992b] to get moderated probabilities for classes in outlying regions. In conjunction with doubt levels this should prohibit classification of outliers. The aim of this paper is to discuss marginalization and compare it to a method for outlier detection which uses a variance based measure. The effects of both methods are visualized using a simple artificial classification problem. Results of outlier rejection are presented in the final section.

II. The Bayesian view of confidence

In [D.J.C. MacKay, 1992b] the author uses Bayesian inference and marginalization to get moderated probabilities for classes in regions, where the classifier is uncertain about the class label. We may expect a trend towards equal probabilities in such regions and are able to refuse classification by flagging "doubt" if none of the probabilities is above a certain threshold. This method will in general lead to a high rate of correct results among all remaining guesses. Nevertheless the question arises in which regions of the input space classification is refused. The second question that should be discussed is, whether an approach to define an "error-bar" for classifiers is a possible alternative.

In the following section we assume that the classifier network is a two layer perceptron and that it models probabilities for classes. The network has one hidden layer with sigmoid activation and a linear output layer.

If we look at the well known expression for the posterior probability for weights, (1), we see, that the posterior probability depends on both the prior $p(\underline{w})$ and the likelihood term $p(\mathcal{D} \mid \underline{w})$.

$$p(\underline{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \underline{w}) p(\underline{w})}{p(\mathcal{D})} \tag{1}$$

The likelihood term will be small for all weight vectors which do not model the probabilities for classes properly. Consequently, if we move in input space in directions from a region with "large probability for a decision boundary" into a region with "large probability for one of the classes", we will come into regions far from any training patterns, where the "marginalization-doubt" approach will have troubles to detect outliers. The reason for this is that all weight vectors with a large value of $p(\underline{w} \mid \mathcal{D})$ will map these input patterns to large positive or negative output activations. In other words the mean value of the predictive distribution of linear network outputs will have large absolute values.

On the other hand [C.K.I. Williams et. al., 1995] have shown that the standard deviation of the predictive distribution in regression problems scales approximately proportional to $p(\underline{x})^{-1}$. Assuming that this relationship holds also in the case of classification for the linear output activation, we may conclude that moderation effects will depend on the direction we move into regions with low input data density $p(\underline{x})$.

In other words: The Bayesian solution describes an uncertainty of classification decisions. All regions which are separated from decision boundaries by a well defined region[2] with high probability for one class, are regions with high certainty of the class label - irrespective whether training data was available there or not. Hence the uncertainty about class label introduced by

[2] Well defined means with sufficient training data.

the predictive solution is *not* useful for novelty detection.

III. Error bar in the latent space

We have two possibilities to perform novelty detection:

- Build a model of the input data density $p(\underline{x})$ and assert *outlier* if the density estimate for a test input is below a certain threshold.
- Try to find an engineering solution that approximates the first. This is the method we will describe here.

In order to make the following part easier to understand, we will add figure 1, a multi layer perceptron (MLP) network that shows the relevant details.

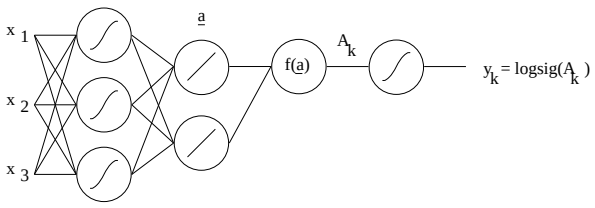

Fig. 1. A sketch showing a MLP-network with the correct output transformation for classification. The sketch shows all relevant expressions we need below in order to describe a method capable of preventing classification in such regions of input space that were not covered by training data.

The neural network classifier shown in figure 1, has one hidden layer with sigmoid activation and a linear output layer. In the two class case there is only one output node, the output-value before transformation is denoted with a. The output values are transformed by an appropriate data model, which requires "logsig" transformation in a two class case and the "softmax" transformation in a more general one-of-c class problem. We will use the simple retransformation, (2), which allows to view the "softmax" transformation as a k-fold "logsig" transformation.

$$P(C_k \mid \underline{x}, \mathcal{D}) = \frac{1}{1 + \exp(A_k)} \qquad (2)$$

$$A_k = a_k - \ln\left(\sum_{k' \neq k} \exp(a_{k'})\right)$$

The expression in (2) reduces to the simple logistic transformation in case of a two class problem with $P(C_1 \mid \underline{x}, \mathcal{D}) = \frac{1}{1+\exp(a)}$ and $P(C_2 \mid \underline{x}, \mathcal{D}) = 1 - P(C_1 \mid \underline{x}, \mathcal{D})$.

It seems obvious to use a measure that is related to the "error bars" commonly used for regression problems (see [D.J.C. MacKay, 1992a]). In this paper we use the standard deviation of the predictive distribution of network outputs before applying the final logistic transformation. In accordance with the terms used in the statistic community, we refer to this as the standard deviation of the distribution over latent space. In a two class problem this is simply the standard deviation σ_a of
$p(a \mid \underline{x}, \mathcal{D})$, where a denotes the linear output activations. In the more general 1-of-c classification problem we have to calculate the standard deviation for all $p(A_k \mid \underline{x}, \mathcal{D})$. The largest standard deviation for any decision, $\sigma_A = \max(\sigma_k) \forall k$, is finally used as a measure for uncertainty. This expression only depends on the distribution of training inputs. A contour plot of both values, the probability for class and the standard deviation of the predictive distribution in the latent space is shown in figure 2 for an artificial three class problem. For the sake of brevity we will denote this standard deviation as "retransformed error-bar".

An upper threshold for the "retransformed error-bar" can be used to suppress classification of outlying patterns.

IV. Experiments and results

In order to see how the proposed method for outlier detection compares to the conventional method, we performed some experiments using real datasets and artificially generated outliers. To find out, whether the resulting classifier is also useful for real problems, we looked at a biomedical problem, where we aimed at using an artificial neural network for sleep staging. In real world scenarios one typically faces the problem that data are contaminated with samples not belonging to either specified class. In the case of sleep staging the *rouge* data are samples of movement periods, where the EEG to be classified into one of six stages shows so called

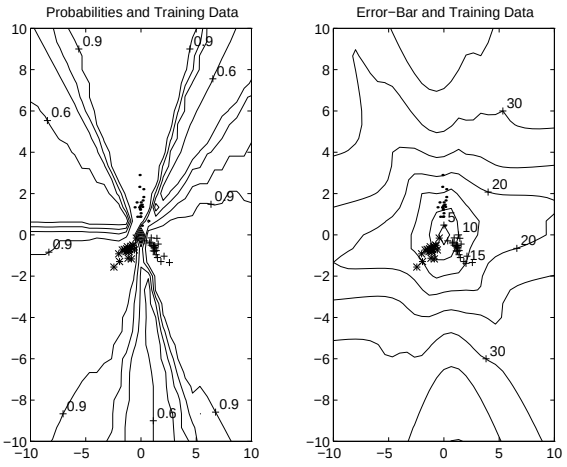

Fig. 2. Contour plot of probabilities and the standard deviation of the predictive distribution of linear activations and the inputs used during training. The left sub-plot shows that, depending on the doubt level, large outlying regions would be classified as belonging to either class. On the right side we see that the predictive distribution over latent space is more reliable in delimiting regions with training data. Note that for reasons of better visualization only parts of the training patterns were plotted.

movement artifacts. These artifacts are muscle activity superimposed on to the EEG activity due to movements of the sleeping subject.

A. Real data and artificial outliers

The experiments were performed with 5 datasets provided online by B.D. Ripley. The data was used throughout his recent book [B.D. Ripley, 1996], a description of the data may be found there. The experiments are based on Cushing's Syndrome data, Virus data, Pima-Diabetes data, Glass data and on the synthetic two class data. These data sets cover both problems with small number of inputs and classes and such problems, where the number of inputs and classes is large. We transformed each feature in the training data to zero mean and unit variance. If the inputs are not uncorrelated, then the high dimensional input vectors will have larger norms. As there are only two outlier samples in Cushing's Syndrome data and none in any of the other data, it was necessary to generate artificial outlier samples.

The generation of artificial outlier vectors was performed in a way, that in the first step uniform distributed data vectors with a length between one and four were generated as as outlying candidates. To be sure, that the outliers are from subspaces not covered by training data, we edited this database. All samples in the outlier data, where at least one of two nearest neighbors was a sample in the training data, were removed. This process is repeated until all samples in the outlier data fulfill this requirement. A similar approach was used in [S. Roberts et. al.,1994] to generate the artificial outlier class. As we use non whitened training data, the high dimensional problems will contain more outlier patterns in regions between class conditional clusters. The low dimensional problems will contain more outlier patterns in regions far from decision boundaries. Therefore the number of correctly assigned outliers may be used to decide which type of outlying patterns can be detected by either approach.

B. Analysis of both methods

The number of hidden units of all neural network architectures used in the experiments were set to twenty. Similar results were achieved with networks with five hidden units. The number of inputs and outputs depend on the data set. Bayesian inference was performed with R. Neals hybrid Monte-Carlo algorithm. As we were not interested in getting any unbiased estimate of the classification performance, we used all available training data. The number of recognized outliers were estimated with a confidence threshold, where 15% of the training data would be declared outlying. As the necessity of outlier removal is motivated by the fact that we only want to classify extremely confident cases, the doubt threshold was set to 0.9. The following table shows as results the number of correctly removed outliers. We performed three experiments: In the first we used only the confidence threshold, in the second rejection was based on the doubt threshold alone, and in the final experiment both thresholds together were used to decide upon whether to declare a sample as an outlier or not. If we compare the re-

TABLE I
PERCENTAGE OF RECOGNIZED OUTLIERS

Data Set	Error Bar	Doubt Level	Both
Cushing's Syndrome	93.4%	49.3%	97.3%
Pima Diabetes	45.7%	84.4%	97.0%
Virus	5.7%	91.6%	93.0%
Glass	14.9%	93.6%	93.6%
Synthetic	80.2%	22.2%	86.5%

sults in table I with the dimension of the problems, we see that there is a strong dependence of the results on the number of network inputs and classes. From the way how we generated the outlier patterns, we may conclude that the "marginalization-doubt" approach is not a good way to get rid of outliers in regions far from decision boundaries, a result which is also illustrated in figure 2. On the other hand, we see that such outliers which are in regions between class conditional clusters are not detected by the second proposed method. To get confident decisions, both methods should be used in combination.

C. Rejection of movement periods in an all night sleep stager

The experiments performed so far show that the method works for artificial problems. In order to see whether this method can also help to suppress classification of outlying patterns in a real problem, we look at a phenomenon that occurs during sleep staging. From time to time the sleeping subject starts moving. The EEG is then contaminated by muscle artifacts and hence can not be assigned to one of the six sleep stages. The number of movement periods is usually very small and it is impossible to model movements as a separate class. What we may try, is to reject movements based on a doubt threshold for the probabilities, based on an upper threshold for the variance in the latent space and finally try a combination of both. In order to cross test the rejection of movements, we used five classifiers trained on different training-samples and then looked at the number of rejections of movement periods.

TABLE II
REJECTION OF MOVEMENT PERIODS

both	$\mathcal{U}$	dbt.	cnf.
46.1%	9%	23.1%	53.8%
61.5%	8.1%	46.1%	53.8%
38.5%	9%	46.1%	0%
69.2%	9%	46.1%	69.2%
69.2%	9.3%	23.1%	69.2%

The results shown in table II suggest that rejection based on a variance measure can improve the overall reliability. Nevertheless one has to be careful in applying this method. The experiment was designed to give a fair comparison of both methods and a combination of them. We started by finding a lower threshold value for the probabilities for classes and an upper threshold for the "retransformed error-bar" and used both thresholds together to reject the movement periods. The first column in table II shows the percentage of correctly rejected movements if the combined method id used. Both thresholds were set to a value where each method alone would reject five percent of all training patterns. The second column lists the true percentage of training samples rejected with both thresholds combined. In order to make the comparison fair, we used the corresponding thresholds when using both rejection methods separately. The fraction of movements rejected by the lower *doubt* level for probabilities is shown in the third column. The last column in table II shows the fraction of movements rejected by an upper threshold for the "retransformed error-bar" alone.

Further analysis of the results requires to define precisely, which question we want to ask. When discussing the different possibilities for outlier rejection, we asserted that when combining both the classical and the proposed method, the classification results should be more reliable compared to the standard "marginalization-doubt" approach. This allegation requires to compare the number of correctly detected movements in column "both" with the number of cor-

rectly detected outliers in the third column. Using a t-test, we got a significantly higher reliability of the combined method at a significance level of 4.04% [3]. If we allow moderate significance levels (5%), then we may conclude that using outlier detection based on a combination of both rejection measures is more reliable than relying on moderation effects alone.

When looking at the numbers in table II and at the result of the test, it becomes clear that both methods together are better than one method alone. Nevertheless the results suggest that the problem of confident decisions is a difficult one. Relying on the combination of both methods one risks in the worst case as many as 60% missed rejections.

V. Conclusion

From our investigations we may conclude that a combined method for outlier detection should be used if classification decisions are only allowed in regions covered by training data. The well known "marginalization-doubt" approach is a good method to get rid of all probable wrong decisions in regions between class conditional clusters, the "error-bar" like approach is better in outlying regions which are not close to any decision boundary.

The work presented here raised several problems that remain to be investigated. The first problem is the problem of proper setting the threshold level for the "retransformed error-bar". If this is done on the worst sample in the training set, then one risks that this is too small if outliers are present in the training data. We could do so when using robust methods during model inference. Finally the method needs to be compared with methods based on the probability density function over input data, which could be done by using classification in the *sampling paradigm*.

Acknowledgements

I want to acknowledge the work of R. Neal from the Departments of Statistics and Computer Science at the University of Toronto, who's software was used to sample from the posterior probability for weights. Furthermore I want to express gratitude to B.D. Ripley from the Department of Applied Statistics at the University of Oxford, who made some interesting data sets available electronically and P. Rappelsberger from the Department of Neurophysiology at the University Vienna for providing several all night recordings.

Peter Sykacek is currently funded by the European Commission (Biomed-2 project SIESTA grant BMH4-CT97-2040). The Austrian Research Institute for Artificial Intelligence is funded by the Austrian Federal Ministry of Science and Transport.

[3] Note that we may do such an analysis, since we restrict ourselves to one paired test and the risk of observing such a result by chance remains 4.04%

References

[J. Bronowski et. al., 1951] J. Bronowski and W. M. Long. Statistical methods in anthropology. *Nature*, 1168:794, 1951.

[D.J.C. MacKay, 1992a] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[D.J.C. MacKay, 1992b] D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4:720–736, 1992.

[C.R. Rao, 1960] C. R. Rao. Multivariate analysis: an indispensable statistical aid in applied research. *Sankhyā*, 22:317–338, 1960.

[B.D. Ripley, 1996] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[S. Roberts et. al.,1994] S. Roberts, L. Tarassenko, J. Pardey, and D. Siegwart. A confidence measure for artificial neural networks. In *International Conference Neural Networks and Expert Systems in Medicine and Healthcare*, pages 23–30, Plymouth, UK, 1994.

[N. Schaltenbrand et. al., 1993] N. Schaltenbrand, R. Lengelle, and J.P. Macher. Neural network model: application to automatic analysis of human sleep. *Computers and Biomedical Research*, 26:157–171, 1993.

[L. Tarassenko et. al., 1995] L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *Proceedings of the Fourth International IEE Conference on Artificial Neural Networks (Cambridge, 1995).*, 1995.

[C.K.I. Williams et. al., 1995] C. K. I. Williams, C. Quazaz, C. M. Bishop, and H. Zhu. On the relationship between bayesian error bars and the input data density. In *Fourth International Conference on Artificial Neural Networks, Churchill College, University of Cambridge, UK. IEE Conference Publication No. 409*, pages 160–165, 1995.