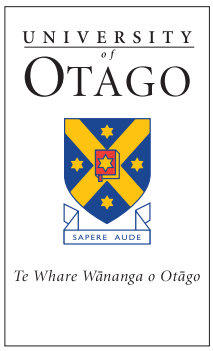


Comparing Huber's M-Estimator Function with the Mean Square Error in Backpropagation Networks when the Training Data is Noisy

David Clark

The Information Science Discussion Paper Series

Number 2000/19
December 2000
ISSN 1177-455X

University of Otago

Department of Information Science

The Department of Information Science is one of six departments that make up the School of Business at the University of Otago. The department offers courses of study leading to a major in Information Science within the BCom, BA and BSc degrees. In addition to undergraduate teaching, the department is also strongly involved in postgraduate research programmes leading to MCom, MA, MSc and PhD degrees. Research projects in spatial information processing, connectionist-based information systems, software engineering and software development, information engineering and database, software metrics, distributed information systems, multimedia information systems and information systems security are particularly well supported.

The views expressed in this paper are not necessarily those of the department as a whole. The accuracy of the information presented in this paper is the sole responsibility of the authors.

Correspondence

This paper represents work to date and may not necessarily form the basis for the authors' final conclusions relating to this topic. It is likely, however, that the paper will appear in some form in a journal or in conference proceedings in the near future. The authors would be pleased to receive correspondence in connection with any of the issues raised in this paper, or for subsequent publication details. Please write directly to the authors at the address provided below. (Details of final journal/conference publication venues for these papers are also provided on the Department's publications web pages: http://www.otago.ac.nz/informationscience/pubs/publications.htm). Any other correspondence concerning the Series should be sent to the DPS Coordinator.

Department of Information Science
University of Otago
P O Box 56
Dunedin
NEW ZEALAND

Fax: +64 3 479 8311
email: dps@infoscience.otago.ac.nz
www: http://www.otago.ac.nz/informationscience/

Comparing Huber's M-Estimator function with the mean square error in backpropagation networks when the training data is noisy

David Clark

Knowlegde Engineering Laboratory
Department of Information Science,
University of Otago,
Dunedin, New Zealand
davidc@ise.canberra.edu.au

Abstract. In any data set there some of the data will be bad or noisy. This study identifies two types of noise and investigates the effect of each in the training data of backpropagation neural networks. It also compares the mean square error function with a more robust alternative advocated by huber.

Introduction

The popularisation of the error backpropagation algorithm by Rumelhart et al (1986) in the late 1980s sparked the resurgence of interest in neural networks. And as Jang et al (1997, p. 234) comment *"Backpropagation MLPs are by far the most commonly used NN structures for applications in a wide range of areas, ..."*. There are many variations on the algorithm – the Matlab neural network toolkit alone has over ten training algorithms. These variations are often standard non-linear optimizations algorithms such as conjugate gradient or quasi Newton applied to the problem of training a feed forward neural network.

The backpropagation algorithm solves the non-linear optimization problem

$$\min \sum_d \sum_n (t_{dn} - o_{dn})^2$$

where t_{dn} = the target value of the d^{th} data point at the n^{th} output neuron Backp

o_{dn} = the output of the d^{th} data point at the n^{th} output neuron

ropagation is a gradient descent algorithm which adjusts each network weight by taking the partial derivative the sum of the squares of the errors $\sum_d \sum_n (t_{dn} - o_{dn})^2$ with respect to that weight. This is how errors propagate and the source of the algorithm's name.

Backpropagation as regression. In discussing regression, Huber (1996, p. 35) says "One wants to estimate the unknown true θ by a value $\hat{\theta}$ such that the residuals $\Delta_i = \Delta_i(\theta) = y_i - f_i(\theta)$ are made 'as small as possible'. Classically, this is interpreted (Gauss, Legendre) as $\sum_i \Delta_i^2 = \min!$... Unfortunately, this classical approach is highly sensitive to occasional gross errors." He suggests replacing Δ_i^2 with a less rapidly increasing function $\rho(\Delta_i)$.

Aim of the study

The aim of this study is to investigate how sensitive standard backpropagation is to noisy training data, and to compare it with an alternative in which a more robust error function (HME – see below) is used in place of the mean square error (MSE). We identify two types of noise in data, namely outliers and mislabeled data. We make comparisons for each type of noise. The performance of the two functions is also compared on data to which no noise has been added.

Specifically, answers to the following questions will be attempted:

1. Does HME perform as well as MSE when no noise is added?
2. How is the performance of a classifier affected by the addition of outliers to the training data?
3. How is the performance of a classifier affected when training data is mislabeled?
4. For questions 2 and 3, is there a difference between classifiers trained using MSE and HME?

Noisy data

In any data set some of the data will be "bad". Hampel (1973) comments "Altogether 5-10% wrong values in a data set seem to be the rule rather than the exception".

Barnett and Lewis (1994, pp. 33, 34) identify three sources of variability in data sets, namely inherent variability, measurement error and execution error. Inherent variability depends on the distribution of the data. Some data sets are naturally more variable than others. For example, people's salaries are more variable than their height. Measurement errors are caused by inadequacies in the measuring instrument. It includes rounding and transcription error as well as instrument malfunction. In the case of a classification problem where the classifier is trained in a supervised mode, a further source of measurement errors is that observations may be mislabeled. Execution errors can arise if the selection of the data is imperfect, such as by the sample being biased in some way.

Outliers

Unrepresentative data are referred to as outliers. Barnett and Lewis define an outlier as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (1994, p. 7). They give two characteristics of an outlier, "engendering surprise owing to its extremeness and … being statistically unreasonable in terms of some basic model" (p. 269).

For much continuous data, the basic model is often normal or near normal. Huber (1996, p. 2) observes that "Typical 'good data' samples in the physical sciences appear to be well modeled by an error law of the form $F(x) = (1-\varepsilon)\Phi(x) + \varepsilon\Phi(x/3)$, where Φ is the standard normal cumulative, with ε in the range between 0.1 and 0.01." He further comments that "this may just be a convenient description of a slightly longer-tailed than normal distribution.". An outlier can thus be identified by its z score – the number of standard deviations from the mean.

With multivariate data the definition of outliers is not straightforward. An observation may indeed "stick out" in one or more of its components, but there may be other data which are outliers because of a combination of components, none of which would be sufficient of itself to warrant being considered an outlier. Unlike in univariate data, no unique total ordering is possible. Sub-orderings are possible, based on particular distance measures. Where the basic model is multivariate normal, Barnett and Lewis recommend $(x-\mu)^T V^{-1}(x-\mu)$, where μ is the mean and V is the variance covariance matrix. Other options include using the z score of a single component of the data, thereby treating it as univariate and using the maximum z score over all of components.

The situation is more difficult if some of the attributes are binary valued. For continuous data an approximately normal distribution is typical. This is not so for binary data. For example, for a binary valued attribute if 20% of the population has one value and 80% the other, the 20% will all be two standard deviations from the mean. These are by no means outliers. The problem of identifying outliers is exacerbated when the components are highly skewed. For instance, in the Card data 45 of the 51 components are binary. Of these, 19 have fewer than 0.2% "ones". It is far from clear what a basic model should be in a case like this. Any identification of possible outliers needs to take into account the pattern of values over all of the binary components.

Consistent or random noise?

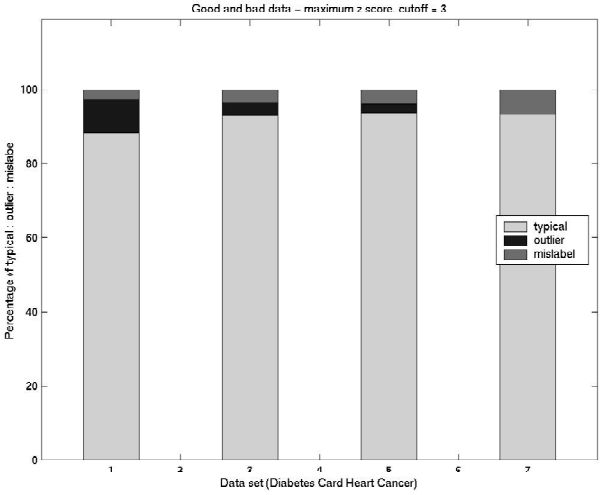
When errors are introduced into data they may be random or consistent. Transcription errors, for example, are likely to be random whilst a misfunctioning or failing instrument will produce consistent errors. Mislabeling errors are also more likely to be consistent. Consistent errors will have more effect on training as random errors will tend to cancel one another. This study, therefore, will focus on consistent errors.

Noise in standard data sets

Noise in data can be due to outliers and to mislabeled data. Figure 1 shows the proportions of noisy data in four standard sets, where the measure used is the maximum z score of any continuous attribute and the cutoff value is 3 standard deviations. This seems a reasonable value given the somewhat longer tailed normal distributions referred to by Huber. Using the inverse covariance measure advocated by Barnett and Lewis gives similar results, although the cutoff value should vary according to the dimensions of the data. In Figure 1 data are deemed to be mislabeled if they are statistically inconsistent with the remainder of the data in their labeled category but appear statistically consistent with the data in another category. Data are deemed to be outliers if they are statistically inconsistent with the data in every category.

The results displayed in Figure 1 indicate that there is noise, both in the form of outliers and mislabeled data, in standard data sets. Changing the cutoff value or the measure used would change the amounts, but unless the cutoff value was made very large, there would still be indication of noise. That there are outliers and mislabeled data in standard data sets is also supported by a recent study by Clark (2000).

Figure 1:
Good and bad data, standard data sets



Robustness

If the data used to train a classifier may contain outliers or mislabeled data, this may have an effect on classifying "good" data. Robust techniques can help to make the training less sensitive to the presence of "bad" data. According to Huber (1996, p. 1) "'*robustness' signifies insensitivity against small deviations from the assumptions*. ... Primarily, we shall be concerned with distributional robustness: the shape of the underlying distribution deviates slightly from the assumed model (usually the Gaussian law)." He further comments (p. 3) that "for most practical purposes 'distributionally robust' and 'outlier resistant' are interchangeable." He goes on to discuss a debate started by Fisher and Eddington in about 1920 on the relative merits of mean square deviation and mean absolute deviation. Huber points out that although mean square deviation is 12% more efficient that mean absolute deviation for exactly normal distributions, as few as 2 bad observations in 1000 suffice to offset the advantage of the mean square deviation. Given this sensitivity of mean square deviation to a small amount of bad data, it is worth while considering more robust alternatives. As Huber comments (p. 3) "I am inclined to agree with Daniel and Woods (1971, p. 84) who prefer technical expertise to any statistical criterion for straight outlier rejection. But even the thus cleaned data will not exactly correspond to the idealized model, and robust procedures should be used to process them further."

An alternative to mean square error

Again we follow Huber (p. 13) "We are particularly interested in location estimates $\sum \rho(x_i - T_n) = \min!$ or $\sum \psi(x_i - T_n) = 0.$
Our favourite choices will be of the form

$$\rho(x) = x^2/2 \qquad \text{for } |x| \le c,$$
$$= c|x| - c^2/2 \quad \text{for } |x| > c,$$
$$\psi(x) = -c \qquad \text{for } x < c,$$
$$= x \qquad \text{for } -c \le x \le c,$$
$$= c \qquad \text{for } x > c.$$

"

This function, which we refer to as Huber's M-Estimator function (HME), is what we use in this study.

Methodology

Adding noise in the form of mislabeled data is simple. All that is required is to change the label and hence the target values of a proportion of the data.
Adding noise in the form of outliers requires more care. Where the data is multivariate difficulties arise as described above. Binary valued attributes cause more difficulties. We attempt to adapt to these difficulties by adding noise to data preprocessed using principal component analysis. This has the effect of orthogonalizing components of the data and eliminating components which contribute the least to variation in the data set, including components which are linearly dependent on others. The data is also normalised so that each component has zero mean and standard deviation of one. Noise is added to a data point by setting one of its components to a large positive value.
The methodology is therefore.

- Add noise in the form of mislabeled data to the training data. In order to make the noise consistent only one category will have its label changed.
- Train classifiers using MSE and HME.
- Compare their results on the validation data.
- Repeat for noise in the form of outliers. In order to make the noise consistent only one component's value will be changed and it will be changed to a large positive value.

The specific experimental details are.

- Noise is added to 0%, 5%, 10%, 15%, 20% and 25% of the training data.
- The value of *c* in Huber's M-estimator function is 0.5. (This is half of the range between the "yes" and "no" target values.)
- When a component is changed to make a data point an outlier its value is set to 4.0.
- Each experiment is repeated 11 items to give both a mean value and a standard deviation.
- The networks were constructed in MSE / HME pairs with the same architecture and initial weights.

Result for Diabetes

In the Diabetes data set (UCI), 8 measurements are used to predict whether a Pima Indian individual is diabetes positive. A single backpropagation classifier correctly classifies about 75% of the data. There were 578 points in the data set, 433 of which were used for training and the remainder for validation.
Table 1 shows the effect on classifier performance as noise is added where the classifiers are trained using MSE and HME.

Table 1
Mean classifier performance - Diabetes data with noise

Noise	Function	Percentage of noise added					
		0%	5%	10%	15%	20%	25%
Outlier	MSE	74.8	73.7	72.3	72.3	72.0	71.7
	HME	75.4	74.5	73.9	73.5	72.9	72.9
Mislabel	MSE	74.8	73.9	74.0	71.4	60.6	52.1
	HME	75.4	74.5	74.2	72.2	63.2	53.0

The results in Table 1 indicate that as outliers are added to the training data, the performance of the classifiers is only slightly affected. When training data is mislabeled, however, the behaviour is quite different. There is only a small deterioration if up to 10% of the data is mislabeled, but when more than 10% of data is mislabeled the performance deteriorates markedly.
Figures 2 and 3 illustrate the results summarised in Table 1.

Figure 2 : Diabetes data with outliers

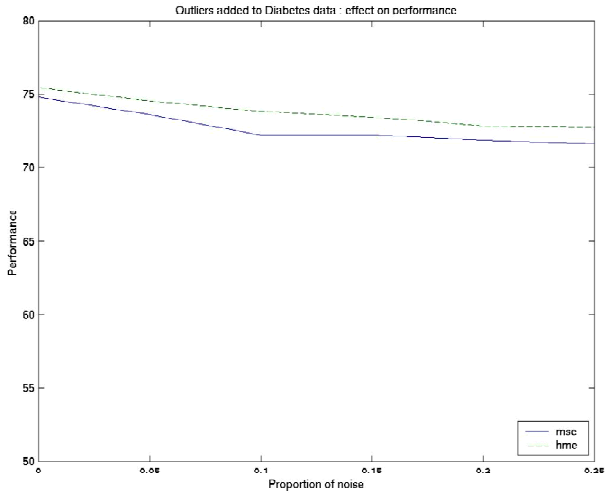


Figure 3 : Diabetes data with mislabeled data

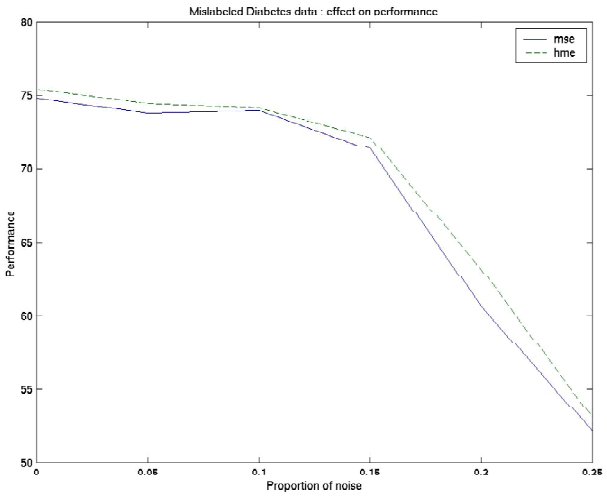


Table 2 shows the effects on the standard deviations of the performance are given in Table 2. There is a tendency to a smaller standard deviation for HME than MSE. Table 2 also shows a markedly higher standard deviation as the performance deteriorates as more than 10% of the training data is mislabeled.

Table 2
Standard deviation of classifier performance - Diabetes data with noise

		Percentage of data to which noise is added					
	Func-tion	0%	5%	10%	15%	20%	25%
Outlier	MSE	1.8	2.4	2.3	2.0	2.3	2.5
	HME	2.1	1.9	1.7	2.1	2.2	2.0
Mislabel	MSE	1.8	2.4	1.9	3.1	3.7	6.0
	HME	2.1	1.9	1.4	3.0	4.9	5.4

The best and worst performances (not shown) were also better for HME than MSE - the worst performances by about 0.5 to 1% and the best performances by about 0 to 0.5%.
A final experiment with the Diabetes data was to add noise randomly to the training data rather than in a consistent manner. For both outliers and mislabeled data there was very little affect on performance. One explanation of this is that there is a canceling out of the random effects.

Results on other data sets

The experiments of the Diabetes data described above was repeated with to several data sets, namely Diabetes, Card and Heart (UCI). Table 3 summarises the results over these data sets.

Table 3
Mean classifier performance – Other data sets with noise

Data set	Function	Percentage of data to which noise is added					
		0%	5%	10%	15%	20%	25%
Card	MSE	84.8	84.3	84.6	85.0	84.9	84.8
Outlier	HME	84.4	84.7	84.4	84.5	84.5	84.6
Mislabel	MSE	84.8	84.3	81.0	77.3	73.2	61.3
	HME	84.4	84.4	81.5	78.0	70.6	57.8
Heart	MSE	80.9	81.1	80.9	80.9	80.8	80.9
Outlier	HME	80.5	80.8	81.0	80.8	80.9	80.7
Mislabel	MSE	80.9	80.7	81.0	77.2	70.0	61.5
	HME	80.5	80.9	82.0	77.8	70.2	60.7
Cancer	MSE	99.3	99.3	99.4	99.3	99.0	98.7
Outlier	HME	99.5	99.4	99.5	99.5	99.6	99.4
Mislabel	MSE	99.3	98.8	98.0	96.3	89.0	82.2
	HME	99.5	99.3	98.5	96.5	89.4	82.0

Two of the patterns of the Diabetes data are repeated, namely the effect on performance as outliers are added and as data is mislabeled. There is only a small deterioration in performance as outliers are added, even when up to 25% of the data is affected. Mislabeling less than about 10% of training data has little effect, but beyond about 10% the deterioration becomes more pronounced.

In the Diabetes data there was a small improvement in the performance of classifiers trained with HME. This is not present in the other data sets.

Finally, the standard deviation of the performance (not shown) was again smaller for HME than MSE over the other data sets.

The results over all data sets can be summarised as

- the performance of classifiers trained with HME is little different to that of classifiers trained with MSE, irrespective of how much noise is added,
- the effect performance of classifiers as outliers are added to the data is small, even when 25% of the data is corrupted,
- up to about 10% of the training data can be mislabeled without having much effect on performance, but thereafter performance deteriorates quite sharply,
- there is less variance in classifier performance with HME than MSE.

That the performance did not deteriorate as outliers were added may have been due to the fact that outlier noise was only added to one component of the data (after pre-processing with principal component analysis). The number of such components is 8, 20, 17 and 8 for Diabetes, Card, Heart and Cancer respectively. Thus in each case only a small proportion of the com-

ponents are corrupted. The information in remaining unaffected components may be sufficient to obviate the effect of the misinformation in the one components. A further experiment was done to explore this possibility. In this experiment outlier noise was added to one quarter of the components rather than just one component. The results are summarised in Table 4. Although there is a little more deterioration than when only one component was changed, it is only small even when 25% of the data points are changed.

Table 4
Mean classifier performance – outlier noise applied to 25% of components

Data set	Function	Percentage of data to which noise is added					
		0%	5%	10%	15%	20%	25%
Diabetes	MSE	74.8	72.8	71.7	70.6	69.5	70.0
	HME	75.4	74.2	73.9	73.1	71.8	71.9
Card	MSE	84.8	85.5	85.2	84.6	84.7	85.0
	HME	84.4	85.5	85.0	85.0	84.5	84.3
Heart	MSE	80.9	78.8	78.1	77.9	78.9	78.8
	HME	80.5	79.8	77.8	77.6	78.9	78.9
Cancer	MSE	99.3	99.3	99.1	98.8	98.2	97.6
	HME	99.5	99.4	99.4	99.1	98.8	97.8

When data is mislabeled, the effect is that all of the components have the same misinformation, and so the effect is greater. The surprise is not that the performance deteriorates, but that it is as resilient as it is for as long as it is.
The lack of improvement when HME is used in place of MSE is rather unexpected. The surprise is not that HME was not effective – it was – but that MSE was equally resilient to outliers. Huber's analyses of MSE and its reputation in other contexts gave rise to the opposite expectation. The explanation may be in the non-linear nature of the backpropagation algorithm, and in particular to the "squashing" activation functions used in individual nodes. This would have had the effect of reducing the effects large input values as they propagate through the network.
A possible explanation of the smaller standard deviation of HME is that the size of the gradient of the HME function is limited to $\pm c$. This would cause the error surface to be flatter so that local minima are not as steep sided.

Conclusions

Some bad or noisy data is to be expected in any data set, and standard data sets are no exception. This study sought the answers to two related questions: how the performance of feedforward neural networks was affected by noise in the training data; and whether replacing the mean square error function with Huber's M-estimator function would improve the performance. The results of the study indicate that: outliers in the training data have little effect on performance; up to about 10% of the training data being mislabeled has little effect on performance, but beyond 10% the perform-

ance deteriorates markedly; and replacing the mean square error function with Huber's M-estimator function has very little effect on performance, whether or not the training data is made noisy.

References

Barnett, V. and Lewis, T. *Outliers in Statistical Data*, 3rd edition. Wiley, Chichester, England. 1994.

Clark, D. I. *Using Consensus Ensembles to Identify Suspect Data*. Discussion Paper 2000/17, Department of Information Science, University of Otago, Dunedin, New Zealand.

Daniel, C. and Wood, F.S., *Fitting Equations to Data*, John Wiley, New York, 1971.

Hampel, F. R. *Robust Estimation: A condensed partial survey*, Z. Wahrsch. Verw. Geb., 27, 1973, pp. 87-104.

Huber, P.J. *Robust Statistical Procedures*, 2nd edition. SIAM, Philadelphia, 1996.

Jang, J.-S., R., Sun, C.-T. and Mizutani, E. *Neuro-Fuzzy and Soft Computing.* Upper Saddle River, N.J.: Prentice Hall, 1997.

Rumelhart, D.E. and McClelland, J.L. (eds) *Parallel Distributed Processing*, vol I, M.I.T. Press, Cambridge, MA, 1986.

UCI machine learning repository at http://www.ics.uci.edu/~mlearn/MLRepository.html