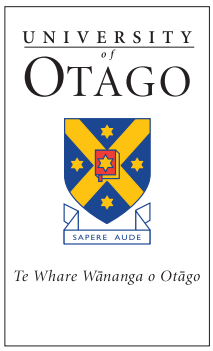


Using Consensus Ensembles to Identify Suspect Data

David Clark

The Information Science Discussion Paper Series

Number 2000/17
November 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/

Using Consensus Ensembles to Identify Suspect Data

David Clark

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

Abstract: In a consensus ensemble all members must agree before they classify a data point. But even when they all agree some data is still misclassified. In this paper we look closely at consistently misclassified data to investigate whether some of it may be outliers or may have been mislabeled.

1 Introduction

Using the results of several classifiers is a technique which has been shown to give more accurate classification than a single classifier [1], [2], [3]. The resulting classifier is known as an *ensemble*.

The most popular methods of constructing ensembles are *bagging* [4] and *boosting* [5]. Both methods generate multiple classifiers by resampling the training data. Bagging (bootstrapping aggregates) trains the component classifiers using independent samples drawn with replacement from the training data. Boosting creates a succession of classifiers by giving greater weight to data points misclassified by previous classifiers.

In an ensemble constructed by bagging, the ensemble may classify a data point by *averaging* or *voting*. When averaging is used, the predictions of the component classifiers are averaged to make the ensemble classification. With voting, each component classifier votes for a category and the ensemble category is the category with the most votes. These methods may be modified by weighting the classifiers according to their individual accuracy.

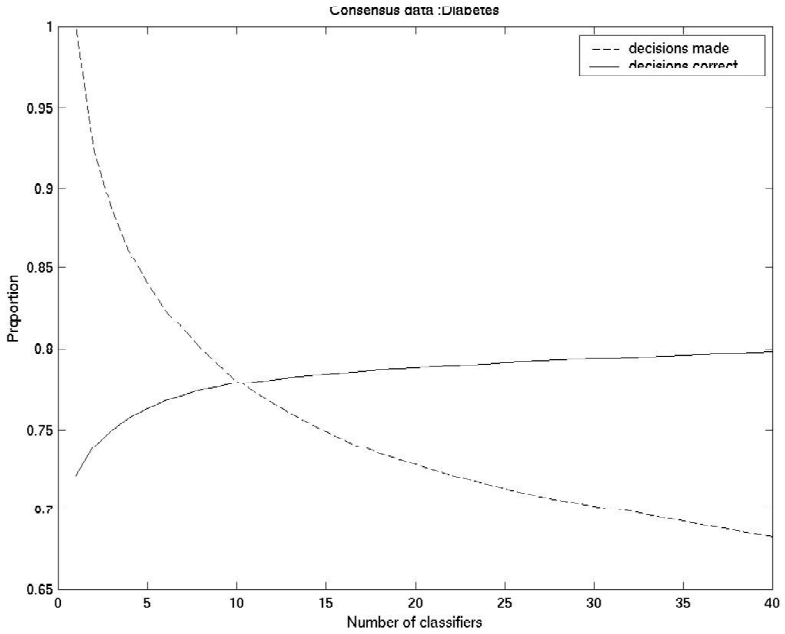
The most popular method for ensemble classification is unweighted averaging [4], [6], typically with the outputs of each component classifier being normalized. Part of the reason that voting is not as popular is that it does not use all of the information available. It does not distinguish between a weak and a strong preference by the component classifier. Voting, however, does give the opportunity not to make decisions where there is insufficient agreement. This can increase the accuracy of classification where a decision has been made. There is thus a trade off between the proportion of data for which a decision is made and the proportion of that data which is correctly classified. Cox, Clark and Richardson [7] explored this trade off. In particular, they investigated using *consensus ensembles* – ensembles which only make a decision when all members of the ensemble agree. Figure 1 shows the effect of the trade off for the Diabetes data [8].

Cox *et al* found that the data can be split into data on which all classifiers agree and data in which there is some disagreement. We refer to them

as consensus data and non-consensus data respectively. A further finding was that although the prediction rate on consensus data did indeed increase, it did not reach 100%. (The Cancer data set [8] was the exception.) That is, there are data on which each of 40 classifiers made the same mistake. For most of the data sets examined in their study, the proportion of this data was between 5 and 10%, although it was 17% for the difficult Abalone data [12] and 0% for Cancer.

The presence of data misclassified by a large number of classifiers raises the question of whether the classifiers were all inaccurate, or whether the data itself was atypical. This study attempts to answer this question.

Figure 1 : Consensus data, Diabetes



Aim of the study

This study focuses on incorrectly classified consensus data. That is, data which all 40 classifiers in an ensemble have misclassified. It identifies these data to see what proportion of them are potential outliers or potentially mislabeled. It compares these proportions to those of the remainder of the data. If the proportions of “bad” data in the incorrectly classified consensus data are significantly higher than those in the remainder of the data, then it is suspect. Once suspect data is identified it can be examined for patterns such as a high proportion originating from a particular instrument or labeled by a particular labeler.

What this study does not do. This study does not suggest that identifying incorrectly classified consensus data is a tool for statistical analysis, or that it should replace normal preprocessing of data. For instance a statistical analysis of data would use a cutoff level appropriate to a particular data set's distribution rather than using the same value for all data sets. Instead, this study acknowledges that some of the data presented to a classifier to use in training may be suspect, and it provides a means of identifying suspect data for closer scrutiny.

Terminology

We shall use the following terms in this paper.

Misclassified data are data which have been misclassified by a trained classifier. That is, the category determined by the classifier is not the label category.

Potential outliers are data which appear statistically inconsistent with the remainder of the data in their labeled categories. (See "Outliers" below.)

Potentially mislabeled data are data which appear statistically inconsistent with their labeled categories, but which appear consistent with another category.

Suspect data are data which are potential outliers or potentially mislabeled.

Typical data are data which are neither potential outliers or potentially mislabeled. Thus all data are either typical or suspect (potential outliers or potentially mislabeled).

Data may also be classified (correctly) by some classifiers or consistently misclassified by all classifiers. Table 1 shows these concepts as applied to the Diabetes data.

2 Outliers

In any data set some of the data will be "bad". Hampel [9] comments "Altogether 5-10% wrong values in a data set seem to be the rule rather than the exception".

Barnett and Lewis [10, 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. Execution errors can arise if the selection of the data is imperfect, such as by the sample being biased in some way. 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.

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" [10, 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" [10, p 269].

For much continuous data, the basic model is often normal or near normal. Huber 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." [11, p 2].

The identification of outliers in continuous univariate data is relatively straightforward. Where the underlying distribution is normal, an observation which is two standard deviations from the mean occurs in less than 5% of the population.

With multivariate data the identification 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. We will refer to this as the inverse covariance measure. Other possibilities include using a single component of the data, thereby treating it as univariate and using the maximum of the single components.

If the data is continuous then an approximately normal distribution is typical. This is not the case with 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. Carelessly applying a "two standard deviations" rule could result in up to 20% of the data being labeled as outliers. Where the data are binary multivariate, 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.

Measurements used in this study

For the purpose of our study we identify an observation as a possible outlier if one or more of its components is more than three standard deviations from the component mean for its labeled category. If an observation has been identified as a potential outlier, we identify it as being potentially

mislabeled if none of its components is three standard deviations from the mean for an alternative category. We will refer to this as the maximum z score measure. We use three rather than two standard deviations because of the Huber's comment that physical data can be modeled by a slightly longer-tailed than normal distribution. But the choice of three standard deviations is not critical. The same pattern of results occurs over a range of values.

We also analysed the data using the inverse covariance measure, but found that it was highly correlated with the maximum z score measure.

The Heart and the Card data have binary valued attributes. We did not use these in our identification of suspect data as in both data sets the pattern of binary values was unique for about half of the data. There did not seem to be any measure that would not suggest that a disproportionate amount of the data was suspect.

3 Methodology

The members of the ensembles were standard feedforward neural networks trained using the backpropagation algorithm. Matlab was used to analyse the data and its backpropagation neural network algorithm was used to populate the ensembles. The ensembles were bagging ensembles.

The analysis for each data set is as follows:

1. Train an ensemble of 40 classifiers. Classify all of the data with each member of the ensemble.
2. Split data into consensus and non-consensus data.
3. Combine non-consensus data with correct consensus data. This is the data which has been correctly classified by some classifiers. The remainder of the consensus data is the data which has been consistently misclassified.
4. For each datum in the two subsets of the data, use the maximum z score measure to categorise it as typical, a potential outlier or potentially mislabeled. Aggregate these statistics.

4 Result for Diabetes

In this section we examine closely the results of the Diabetes data. In the Diabetes data set [8], 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.

Of the 145 points used for validation, 99 were consensus and 46 non-consensus. Of the consensus data 79 were correct. Hence 20 data points were misclassified by all classifiers, while the remaining 125 data points were classified by at least one classifier.

Table 1 below summarises the results for the Diabetes data.

Table 1: Diabetes data
maximum z score measure, cutoff 3 standard deviations

	Typical data	Suspect data		Total
		Potential outlier	Potentially mislabeled	
Classified by some classifiers	107 (86%)	12 (10%)	6 (5%)	125
Misclassified by all classifiers	12 (60%)	2 (10%)	6 (30%)	20

The difference between the consistently misclassified data and the remainder is evident from the table, and is clearly significant. 40% of data which was consistently misclassified is suspect, as opposed to 14% in the remainder. The hypothesis "the proportions of suspect data misclassified by all classifiers and classified by some classifiers are equal" was tested. The numbers in Table 1 give a Chi Square value of 7.7 which (with one degree of freedom) is large enough to reject the hypothesis at the 1% level.

To test how sensitive the results were to the choice of outlier cutoff, we repeated the experiments using values of 1.5, 2, 2.5, 3 and 3.5 standard deviations. The results are shown in Table 2. They show the same pattern as in Table 1, but with the number of suspect observations decreasing with the cutoff level.

Table 2: Diabetes data, varying cutoff level
maximum z score measure

Cutoff level	Classified by some classifiers			Misclassified by all classifiers		
	Typical	Suspect		Typical	Suspect	
		Potential outlier	Potential mislabel		Potential outlier	Potential mislabel
1.5	68	49	8	1	9	10
2.0	96	23	6	8	3	9
2.5	105	16	4	11	2	7
3.0	107	12	6	12	2	6
3.5	111	9	5	15	1	4

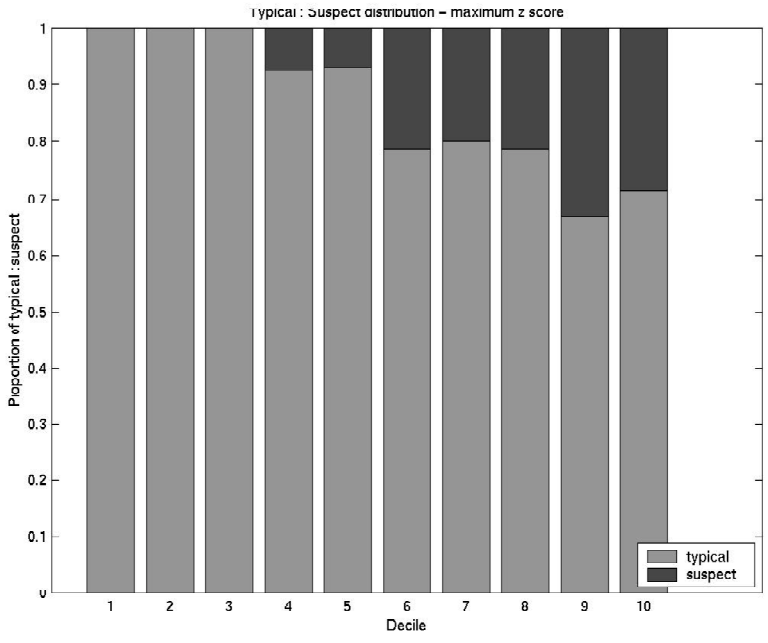
To test whether the results depended on the choice of outlier measure, the inverse covariance measure was used at varying cutoff levels. Table 3 is the equivalent of Table 2, using the inverse covariance measure in place of the maximum z score measure. The results in Table 3 show the same general pattern as those in Table 2.

Table 3: Diabetes data, varying cutoff level
inverse covariance measure

Cutoff level	Classified by some classifiers			Misclassified by all classifiers		
	Typical	Suspect		Typical	Suspect	
		Potential outlier	Potential mislabel		Potential outlier	Potential mislabel
8.5	85	23	17	8	5	7
9.0	105	13	7	10	3	7
9.5	118	5	2	11	2	7
10.0	122	2	1	12	1	7
10.5	124	1	0	16	0	4

Finally the Diabetes data was examined further by ranking it according to the maximum z score measure. Figure 2 shows the proportion of consistently misclassified data in each decile. It shows that the consistently misclassified data is over-represented in the higher deciles. This does not of itself indicate that they are outliers. After all some data has to come at the ends in any ranking. Nevertheless it does illustrate that the consistently misclassified data tend to be extreme, even though the extremeness may not be statistically unreasonable.

Figure 2: Distribution of suspect Diabetes data by decile
maximum z score



Note that not all data with a high z score will have been misclassified. This is to be expected. One of the strengths of neural networks as classifiers is their ability to construct highly non-linear non-convex discriminants. If statistical tests such as the maximum z score or the inverse covariance measure were sufficient to classify data, there would be no point in using neural networks.

5 Results on several data sets

The analysis of the Diabetes data described above was applied to several data sets, namely Diabetes, Card and Heart [8], Abalone [12] and Glass, Wine and Mortgage [13].
Table 4 summarises the results over these data sets. The binary attributes in the Card and Heart data were used in training and classification but not in the identification of suspect data. The cutoff for the Wine data was 2.5.

Table 4: Several data sets, maximum z score, cutoff level = 3

Data	Classified by some classifiers			Misclassified by all classifiers		
	Typical	Suspect		Typical	Suspect	
		Potential outlier	Potential mislabel		Potential outlier	Potential mislabel
Diabetes	107	12	6	12	2	6
Card	224	9	9	9	2	7
Heart	287	7	19	24	1	7
Abalone	547	3	10	170	1	64
Glass	34	1	1	11	2	3
Mortgage	22	0	0	0	1	0
Wine	76	10	1	1	0	1

The results in Table 4 show a pattern of data which has been consistently misclassified having a higher proportion suspect than data which has been correctly classified by at least one of the classifiers in the ensemble. The results also indicate that the misclassification may have been due to the data being mislabeled. Whilst the numbers in the Wine and Mortgage sets are too small to stand alone, they are included as they show the same behaviour.

6 Conclusions

The evidence in this paper suggests that data which is consistently misclassified by a large number of classifiers is not typical of data in its labeled category. The results indicate that a significant proportion of the data may be outliers or may have been mislabeled. These results are not sensitive to the threshold for deciding on whether or not data is typical.

Nor are they dependent on the particular measure used to identify data which is suspect. Consistently misclassified data is suspect and should be identified for closer external scrutiny. The use of consensus ensembles provides a tool for identifying suspect data.

References

1. Brieman, L.: Bagging predictors. Machine Learning 24(2) (1996) 123-140
2. Clemen, R.: Combining forecasts: A review and annotated bibliography. Journal of forecasting 5 (1989) 559-583
3. Wolpert, D.: Stacked generalization. Neural Networks 5 (1992) 241-259
4. Brieman, L.: Stacked regressions. Machine Learning 24(1) (1996) 49-64
5. Freund, Y. and Schapire, R.: Experiments with a new boosting algorithm. Proceedings of the Thirteenth International Conference on Machine Learning (1996) Morag Kaufmann 148-156
6. Alpaydin, E.: Multiple networks for function learning. Proceedings of the 1993 IEEE International conference on Neural Networks, I (1993) 27-32
7. Cox, R., Clark, D. and Richardson, A. An investigation into the effect of ensemble size and voting threshold on the accuracy of neural network ensembles. The 12th Australian Joint Conference on Artificial Intelligence (AI'99), Sydney, Dec, 1999, 268-277.
8. UCI machine learning repository at:
   http://www.ics.uci.edu/~mlearn/MLRepository.html
9. Hampel, F. R. Robust Estimation: A condensed partial survey, Z. Wahrsch. Verw. Geb., 27, pp. 87-104.
10. Barnett, V. and Lewis, T. Outliers in Statistical Data, 3rd edition. Wiley, Chichester, England. 1994.
11. Huber, P.J. Robust Statistical Procedures, 2nd edition. SIAM, Philadelphia, 1996.
12. Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J. Ford, W. B., "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of the Bass Strait, SFD, Tasmania. Technical Report # 48 (1994)
13. Knowledge Engineering Laboratory, Department of Information Science, University of Otago at http://divcom.otago.ac.nz/infosci/kel/