ON ERROR CRITERIA IN HUMAN CLASSIFICATION MODELING

Louis ten Bosch[1], Roel Smits[2]
[1] *Lernout & Hauspie Speech Products N.V., Ieper, Belgium;*
Institute of Phonetic Sciences/IFOTT, University of Amsterdam, NL
tenbosch@brussels.lhs.be, louistb@fon.let.uva.nl
[2] *Max Planck Institute, Nijmegen, The Netherlands*

ABSTRACT

A number of error measures will be discussed that are used for training and testing classification algorithms designed to simulate human classification behaviour. These are (1) the error criterion based on multinomial decision strategy (average log likelihood ratio), (2) the mean squared error (MSE) based on the L_2 (Euclidean metric), (3) an error criterion based on similarity, (4) a novel one, the average log likelihood ratio. We will not focus on particular minimalization methods that are inherent to specific numerical minimization schemes, such as the back propagation method, stochastic annealing, etc., but rather on the conceptual differences of the error measures mentioned.

The classifiers are implemented by means of a feedforward network. The training will be considered to be supervised in all cases.

1. INTRODUCTION

This paper aims at a discussion and evaluation of distance functions used in numerical models simulating human classification. In recent models aiming at the simulation of human phoneme classification (see e.g. Smits et al. (1996), Nearey (1997)), the training algorithm uses an error measure ('distance function') during training and test. Commonly used examples of distance measures are the root-mean-squared distance, the Mahalanobis distance, nearest prototype distance, or measures related to log-likelihood (G^2, Kullback divergence). Evidently, the choice for a distance measure depends on the purpose of the model and on the need for a possible interpretation of the results in terms of human decision behaviour.

In the literature, many aspects of classification have been thoroughly discussed in a technical fashion (Fukunaga, 1972; Duda & Hart, 1973), and recent advancements have been suggested to adapt classically known error criteria (see e.g. Delvroye et al., 1997). In this paper, we will focus on the definition of a cost function that is used for the minimization on a training set of training samples, with special emphasis on the possibility of an *interpretation* in terms of human classification performance. Such an interpretation is in many cases not the focus of the technically and statistically oriented discussions.

The background of this research has been the modeling (simulation) of human categorization data. In human classification experiments, subjects are normally asked to label stimuli drawn from a continuous stimulus space by giving for each stimulus one unique response within a closed response set (N-alternatives, forced choice paradigm). Here, it is assumed that the human decision strategy is invariant over time; the aim of the modeling is to simulate the static performance of (a group of) well-trained subjects. Not necessarily each subject makes use of the same decision strategy. The assumption of invariance over time does not mean that a subject responds the same class with each repeated presentation of a single stimulus, but only that the decision strategy (which may include a biased die) is invariant.

2. CLASSIFICATION MODELS

A number of theoretical frameworks have been presented for a model for human classification (Nosofsky & Palmeri, 1997; cf. Ashby & Maddox, 1994). The discussed models of classification can be divided in two types. *Generalized contect models* stress the use of exemplars, i.e. class specific prototypes. During training, exemplars are trained such that the distance between an incoming unknown stimulus and the nearest prototype determines the probability of the stimulus being a member of a class. On the other hand, *decision bound models* emphasize the class boundaries, and the classification of an incoming unknown stimulus is related to the position of the stimulus relative to the class boundaries surrounding the incoming stimulus. Potentially, both model types have a capability of training arbitrarily complicated class distributions. From the input of *ideal* human observers, it is hard to distinguish both types of models on performance.

The entire simulation model consists of the following steps:

- a representation step, in which an input X is represented by a representation vector $(x_1, x_2, \ldots, x_m)$. The x_i are the (input) features of X.
- a mapping from the representation vector to a vector consisting of *a posteriori* categorical probabilities $(P(C_i|X)_{i=1,\ldots,n})$
- a choice model, in which the choice of a category C_j on the basis of the probabilities $(P(C_i|X)_{i=1,\ldots,n})$

The second step in this scheme is implemented as a multi layer perceptron (feed forward network, ANN) without time recurrence (cf. Richard & Lippmann, 1991).

3. ERROR CRITERIA

Throughout the discussion, the k^{th} stimulus will be denoted by X_k. Each stimulus is presented M times. Let $y_{ij} = p_{obs}(C_j|X_i)$ and $\tilde{y}_{ij} = p_{pred}(C_j|X_i)$ denote the probability of responding class C_j to stimulus X_i as defined by the training set (observed) and by the model (predicted), respectively. The number of classes is denoted n.

We will discuss:

(1) the (log)likelihood measure of the data given the model;

(2) the Euclidean or Mahalanobis (covariance-based) measure;

(3) a statistical similarity measure;

(4) the average likelihood ratio measure.

AVERAGE LOG LIKELIHOOD RATIO (ALLR)

First, we consider the likelihood of a response Y on the basis of a stimulus X and the estimated categorical probabilities $P(C_i|X) = \tilde{y}_i$. For each individual stimulus X, this likelihood $L(X)$ (depending on X) is based on the multinomial function and reads

$$L(X) = \frac{M!}{\prod m_i!} \prod_{i=1}^{n} \tilde{y}_i^{m_i} = \frac{M!}{\prod m_i!} \prod_{i=1}^{n} \tilde{y}_i^{y_i M} \qquad (1)$$

and the resulting likelihood ratio reads

$$LR = \frac{L(X)}{L_{max}(X)} = \prod_{i=1}^{n} \left(\frac{\tilde{y}_i}{y_i}\right)^{y_i M}$$

For each stimulus X, the distance between model and observation is now defined as the log likelihood ratio $LLR(X) = \log(LR)$. The sum $\sum_X LLR(X)$ is the first error criterion to be considered; twice its value is also known as the G^2 measure.

EUCLIDEAN DISTANCE

In this case, the error d between observed probability vector $Y = (y_i)$ and predicted probability vector $\tilde{Y} = (\tilde{y}_i)$ is for each stimulus specified by

$$d^2(X) = (Y - \tilde{Y})^t G (Y - \tilde{Y})$$

where G is the corresponding Mahalanobis matrix (which may be the inverse of an appropriate non-singular covariance matrix). Here, $G = I$.

If the classifier is trained to minimize the Euclidean (or Mahalanobis) error, then the optimum classifier is independent of the representation of the data, that is whether the vector $(y_i)_{i=1,\ldots,N}$ represents probabilities or represented by a number of 0-1 vectors. That property is only shared by the similarity measure.

SIMILARITY MEASURE

A similarity measure is based upon the following argument. Let $Y = (y_i)$ and $\tilde{Y} = (\tilde{y}_i)$ denote the observed and predicted class assignment probability vector, respectively. If one assumes the third step of the decision model to be based on a biased die with the adequate probabilities, the probability p that model and subject predict the same class reads

$$\sum_i y_i \tilde{y}_i$$

The average probability of the model predicting the correct class for all stimuli simultaneously relates to $\sum_X \sum_i y_i \tilde{y}_i$ which is to be optimized.

AVERAGE LIKELIHOOD RATIO (ALR)

This measure has been used successfully by Smits et al. (1996) to model the response behaviour of subjects in a stop classification experiment. In contrast with the log likelihood ratio and log likelihood, the average likelihood ratio does not show the property of resulting in a low score for the whole training set due to just one bad response ('outlayer').

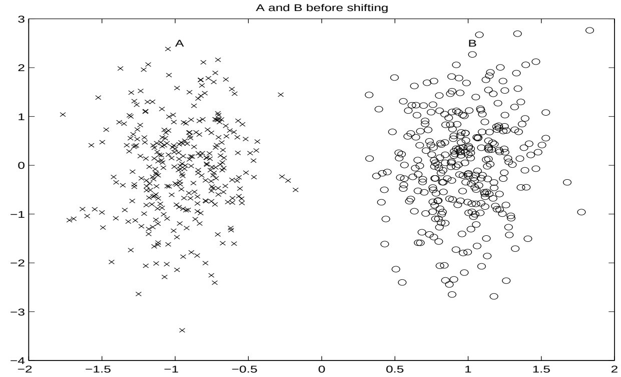

Figure 1: Sets A and B used in experiment 1, before the shift. $\lambda = 0$.

Let $P = \left(\frac{M!}{\prod m_i!}\prod_{i=1}^{n} \tilde{y}_i^{m_i}\right)$ denote the actual probability $P(data|model)$. If $L_i = P/P_{opt}$, $-2\log(L_i)$ is χ^2-distributed met degree of freedom df = number of classes - 1. As a consequence, all L_i (independent of the stimulus index i) have identical distributions, not given however by χ^2 but rather by its exponential $\exp(-\chi^2/2)$. This new distribution is only dependent on χ^2, which is in turn only dependent on df, which is in turn the same for each stimulus. If we can assume that the L_i are independent across stimuli, we can finally apply the Central Limit theorem (Stuart & Ord, 1993, ch. 8), and conclude that $\sum_i L_i$ is approximately gaussian.

4. OUTLAYER BEHAVIOUR

From the expressions for the ALLR and ALR, it is evident that there are two problematic cases: (1) $y_i = 0$ and $\tilde{y}_i > 0$, and $y_i > 0$ and $\tilde{y}_i = 0$. Case (1): the fact that there are no obervations is sometimes to be considered as positive information. However, in the expression for the ALLR, the corresponding term contributes 0 to the overall amount, irrespective of the predicted $\tilde{y}_i$. Agresti (1990) suggests to get rid of empty cells by smoothing the observed counts after discounting large counts. Case (2): This occurs when the subjects do not classify ideally and the model is crisp, i.e. the predicted probability $\tilde{y}_i$ equals zero. The ALR can be justified on statistical grounds by arguing that this measure reflects the goodness-of-fit in case of a *randomly chosen* stimulus from the stimulus set, rather than the goodness-of-fit of the model for the *entire collection* of stimuli as a set.

5. EXPERIMENTS

In order to show the differences between the similarity,

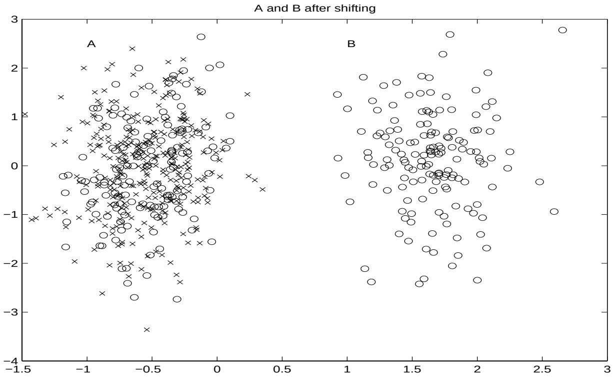

Figure 2: Sets A and B used in experiment 1, after the shift. $\lambda = 1$.

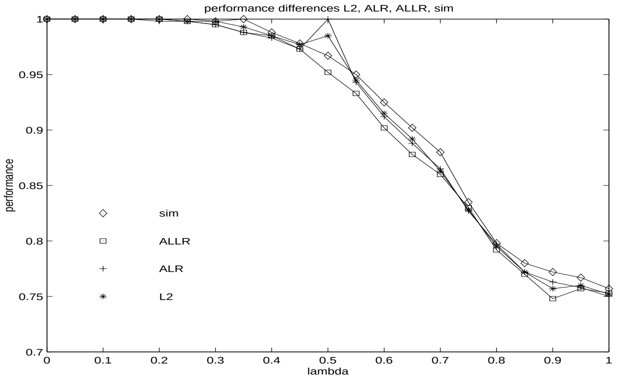

Figure 3: Experiment 1. Fraction correct as a function of the shift λ, after training with sim (diamond), ALLR (square), ALR (+), and L_2 (star).

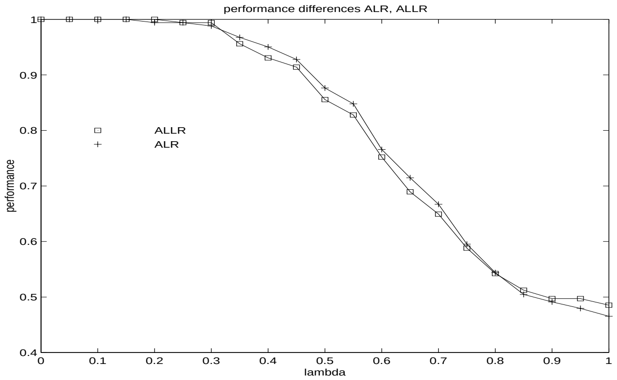

Figure 4: Experiment 2. Fraction correct as a function of the factor λ.

ALR, ALLR, and the L_2, a number of numerical experiments have been set up. All experiments show the same tendency for various shapes of data sets. Two experiments are discussed in more detail below. The observed tendencies are less clear if the data sets show a great amount of overlap.

EXPERIMENT 1

In this experiment, the behaviour of the four error measures (L_2, ALLR, ALR, and similarity) is studied in a simple setting. Two 2-dimensional datasets A and B (fig. 1) have been generated by randomly sampling two different Gaussian pdfs with covariance matrix close to 0.14 times the unity matrix. A and B are linearly separable and consist each of 300 points. The right hand set (set B) is modified such that a subset of 150 randomly chosen points from B shift in the direction of A, and finally end in A. (figure 2, $\lambda = 1$). The shift is parametrized by a parameter λ. Figure 3 shows the results of the classification training with the different error measures. The shift parameter λ is plotted along the horizontal axis, the performances (in terms of fraction correct) is plotted along the vertical axis. As long as A and B do not overlap, minimization of the four error measures yields a percentage correct of close to 100 percent. As soon as the two sets A and B start overlapping, the optimization by L_2, LAR and the similarity yields similar results in terms of correct class membership. For λ between 0.5 and 0.7, where the two sets start overlapping and the number of outlayers is low, the ALLR results are comparable, but slightly worse (1.0 percent) than ALR or L_2, which is significant at the 1% level ($N = 600$). The difference disappears for larger overlap between A and B, when there are many intruders of B inside A. Finally, the separation is only 75 percent, which is the theoretically best result.

One observes that the similarity performs well compared to the other measures; this finding, however, depends on the global shape of the sets.

EXPERIMENT 2

This experiment is based on more elaborate sets. Four different gaussian datasets A, B, C, and D have been created, with 300, 200, 100 and 100 points respectively. The dimension is 3, and the means are fixed to $(0, 0, 0, (1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$ respectively. A random subset of A (150 points) is linearly scaled up with a factor λ until this new set overlaps B, C or D. Figure 4 shows the classification performance after a training with ALLR and ALR. The scaling factor λ is plotted along the horizontal axis, the performances along the vertical axis. As soon as there is a small overlap the ALLR optimization shows a robust slightly worse performance (1.5 percent) than ALR. The difference disappears for larger overlap between the classes ($\lambda = 0.8$). For larger overlaps, the ALLR outperforms the ALR.

CONCLUSION

The two numerical experiments show a slightly but significantly different behaviour a classifier after training with different error measures. Of all four measures tested, L_2 and ALR perform equally well (also due to the shapes of the classes considered). ALLR and ALR show a difference if there is a small overlap between the classes; a difference that turns into the opposite for larger overlaps. In general, the ALR is preferable over ALLR when the amount of overlap is small; for larger overlaps the ALLR tends to be better. This can all be explaned by considering the expressions in the case of outlayers. In general, the behaviour of the error measures discussed depend on the pathologies of the class shapes and the topological properties of the underlying pdfs.

REFERENCES

Ashby, F.G., Todd Maddox, W. (1994). A response time theory of separability and integrality in speeded classification. J. of Math. Psych. 38 (4).

Delvroye, L., Györfi, L, and Lugosi, G. (1997). A probabilistic theory of pattern recognition. Springer Verlag, Berlin, Heidelberg.

Duda, R.O., and Hart, P.E. (1973). Pattern classification and scene analysis. John Wiley and Sons, New York. (1st ed.)

Fukunaga, K. (1972). Introduction to statistical pattern recognition. Academic Press, New York.

Nearey, T.M. (1997). Speech perception as pattern recognition. JASA 101(6), pp. 3241.

Nosofski, R.M., Palmeri, T.J. (1997). Comparing exemplar-retrieval and decision-bound models of speeded perceptual classification. Perception & Psychophysics 59(7), pp. 1027-1048.

Richard, M.D., and Lippmann, F.F. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Computation 3, p. 461-483.

Smits, R., ten Bosch, L.F.M., Collier, R. (1996). Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. II. Modelling and evaluation. JASA 100(6), p. 3865-3881.

Stuart, A. and Ord, J.K. (1993). Kendall's Advanced Theory of Statistics. Vol 1, Distribution Theory.