
Proceedings of the

# Twentieth Annual Symposium of the Pattern Recognition Association of South Africa

30 November – 1 December 2009
Stellenbosch, South Africa

Proceedings of the

# Twentieth Annual Symposium of the Pattern Recognition Association of South Africa

30 November – 1 December 2009
Stellenbosch, South Africa

http://www.prasa.org

Edited by: F. Nicolls
ISBN 978-0-7992-2356-9

Member of the International Association of Pattern Recognition (IAPR)

# Table of Contents

## Full papers


Learning Structured Representations of Data 1
*Etienne Barnard, Christiaan van der Walt, Marelie Davel, Charl van Heerden, Fred Senekal, and Thegaran Naidoo*

The challenges of ignorance 7
*Etienne Barnard*

Model Based Estimation for Multi-Modal User Interface Component Selection 11
*L. Coetzee, I. Viviers, and E. Barnard*

Applying Bayesian segmentation in volumetric silhouette-based reconstruction 17
*Willie Brink*

Rule-based Conversion of Closely-related Languages: A Dutch-to-Afrikaans Convertor 23
*Gerhard B. van Huyssteen and Suléne Pilon*

Characterization of a Small Helicopter UAV's Main Rotor Blade through Image Processing 29
*William J. Shipman, Yuko Roodt, and Francois du Plessis*

Comparing manually-developed and data-driven rules for P2P learning 35
*Linsen Loots, Marelie Davel, Etienne Barnard, and Thomas Niesler*

Real-time stereo reconstruction through hierarchical DP and LULU filtering 41
*François Singels and Willie Brink*

Characterisation and simulation of telephone channels using the TIMIT and NTIMIT databases 47
*Herman Kamper and Thomas Niesler*

Texture based classification of images using frequency estimated pairwise joint probability distributions on site labels from wavelet decomposed images 53
*Markus Louw and Fred Nicolls*

Combining Multiple Classifiers for Age Classification 59
*Charl van Heerden and Etienne Barnard*

Combining motion detection and hierarchical particle filter tracking in a multi-player sports environment 65
*Robbie Vos and Willie Brink*

A perceptual evaluation of corpus-based speech synthesis techniques in under-resourced environments 71
*D.R. van Niekerk, E. Barnard, and G. Schlünz*

Background Subtraction Survey for Highway Surveillance 77
*Zane Mayo and Jules R. Tapamo*

# The challenges of ignorance

*Etienne Barnard*

Human Language Technologies Research Group
Meraka Institute, CSIR, Pretoria, South Africa
ebarnard@csir.co.za

## Abstract

We have previously argued that the infamous "No Free Lunch" theorem for supervised learning is a paradoxical result of a misleading choice of prior probabilities. Here, we provide more analysis of the dangers of uniform densities as ignorance models, and point out the need for a framework that allows for prior probabilities to be constructed in a more principled fashion. Such a framework is proposed for the task of supervised learning, based on the trend of the Bayes error as a function of the number of features employed. Experimental measurements on a number of standard classification tasks confirm the representational utility of the proposed approach.

## 1. The No Free Lunch theorem in Pattern Recognition

The "No Free Lunch" (NFL) theorem for supervised learning [1, 2] is one of the most controversial results in all of pattern recognition. Taken at face value, the NFL theorem implies that learning can only succeed if the learning algorithm happens to make the correct assumption about the problem being solved. That is, "unless one can establish *a priori*, before seeing any of the data $d$, that the [function] $f$ that generated $d$ is one of the ones for which one's favourite algorithm performs better than other algorithms, one has no assurances that that algorithm performs any better than the algorithm of purely random guessing." [3] Or, to quote a popular text on pattern recognition [4]: "there are no context- or problem-independent reasons to favor one learning or classification method over another."

Now, to the regular user of pattern-recognition algorithms, these statements are quite unexpected: we certainly tend to prefer, say, support vector machines over naïve Bayesian classifiers, and both of those algorithms over random guessing! This conflict between theory and practice has prompted a number of responses [5, 6, 7, 8], including a recent analysis [9] which characterises the NFL theorem as a logical paradox - that is, as a counter-intuitive result that is correctly proved from apparently incontestable assumptions. In particular, that paper demonstrates that the uniform prior used in the proof of the theorem has a number of unpalatable consequences besides the NFL theorem. However, [9] does not propose alternative prior distributions that can be used in place of the uniform prior; such distributions would be very useful for a number of theoretical and practical reasons (e.g. the abstract comparison of different learning algorithms or the generation of "representative" data sets). In the current paper, we investigate some of the properties of such a "generic" prior for pattern recognition, and demonstrate a descriptive framework that may be useful in that regard.

A number of researchers have attempted to provide descriptions of the systematic regularities that appear in pattern-recognition problems. De Villiers and Barnard [10] as well as Van Der Walt and Barnard [11] required a systematic way to generate "typical" pattern-recognition problems, and proposed meta-density functions from which parameters of Gaussian mixtures could be drawn for this purpose. Brazdil, Guma and Henery [12] and Ho and Basu [13] introduced several measures that can be used to characterise classification problems, primarily to understand performance of different classification algorithms. Below, we add to this collection of characterisations by focusing on characteristics that are shown to be important from an analysis of the NFL theorem.

Our main aim is to propose an expression of the generic prior in terms of the *classification power* inherent in successively more encompassing subspaces of feature space. In order to motivate this proposal, we briefly introduce the Extended Bayesian Framework (EBF), which makes it possible to speak about these concepts with precision (Section 2), and review the factors that indicate the unacceptability of the uniform prior used to prove the NFL theorem. In Section 3, these insights are used as basis for the development of a theoretical tool that can be used to describe pattern-recognition problems at an appropriate level of abstraction. Section 4 then applies this tool to a number of standard pattern-recognition benchmarks, and in Section 5 we relate these conceptual and practical results to the overall goal of establishing a suitable functional prior.

## 2. NFL: statement and perspectives

The EBF as defined in [2] extends conventional probability theory by treating the *hypothesis* that is output by a learning algorithm as a random variable $h$. In addition to the probabilistic relationships between the *training data* $d$ and the function $f$ that represents the *input-output relationships*, one is led to also consider such relationships between $h$ and $d$, and the generalisation error is expressed precisely by conditioning the off-training set error on all three these variables ($h$, $f$ and $d$). To make practical progress, one generally assumes that the distributions of the hypothesised and underlying functions ($h$ and $f$, respectively) are independent given the data $d$; this allows the expected error rate given a training set to be expressed as a non-Euclidean inner product between the distributions of these functions, conditioned on the data. NFL then follows straightforwardly by choosing a particular (uniform) prior for the functions $f$. (The original papers on NFL also contain an alternative perspective to NFL which does not rely on the adoption of this prior - see [2]. We return to this issue in Section 5 below.)

The manner in which the EBF leads to NFL can be grasped straightforwardly by considering a deterministic two-class classification task defined over $n$ binary variables $x_i$. There are $2^n$ combinations of these variables, and therefore $2^{2^n}$ deterministic functions that can be defined by making all possible assignments of classes to these combinations. Wolpert suggests that

each of these functions should be given the same prior probability in the absence of further information.

In this context, an inductive learning algorithm is a function that takes a subset of $k$ training samples (that is, combinations of variables along with their classification) and produces hypotheses for the classification of the remaining $2^n - k$ variable combinations. Now consider the behaviour of two different learning algorithms $c_1$ and $c_2$, and the prediction that they make for an unseen (test) sample $x$. For each possible target function in which $c_1$ outperforms $c_2$, there is a corresponding target function for which the converse is true. Hence, since all target functions have the same prior probability, the expected values for the accuracies of the two classification algorithms are exactly equal. Since this is true for any $x$, one is led to conclude that any two such learning algorithms are equivalent if one does not make additional assumptions about the distribution of the target functions, which is the NFL theorem.

In [9] it is shown that the "uninformative" prior at the basis of this proof is not as innocuous as it seems. In fact, the uniform prior is shown to represent an extreme lack of determination: any amount of training data is expected to produce a negligible amount of information about the input-output function to be learnt. That is, under this assumption, any accurate prediction of off-training set data is extremely unlikely (with probability decreasing exponentially in the size of the test set). It is also shown that this behaviour results from the extreme symmetry that the uniform prior imposes on all variables, and on all values of those variables. Any group of variables can be permuted with any other group of variables on an arbitrary training or test sample without affecting any likelihoods, and the value of any binary variable can similarly be inverted arbitrarily - thus entirely destroying the concept of identity for any variable.

In real pattern-recognition tasks, on the other hand, variables have distinct characteristics which are responsible for the properties of feature spaces that we take for granted. For example, different classes have distinct class-conditional densities as a function over the different variables; these variables correlate with one another to a greater or lesser extent, and are also in various degrees able to separate the various classes from one another. Each of these characteristics implies systematic regularities in any real pattern-recognition problem, whereas the uniform functional density assigns equal weight to regular and to highly irregular functions.

We are therefore lead to believe that the uniform prior does not serve well as an expression of ignorance, and it is interesting to note that this same observation has been made in a number of different contexts.

- In one version of De Mere's paradox, the famous gambler is said to argue that three dice should with equal likelihoods sum to either 11 or 12, since both sums result from 6 different combinations of single-die values [14]. However, empirical observation had shown him that a sum of 11 was in fact notably more frequent. Pascal pointed out that this assumption of uniformity over combinations is fallacious it is, in fact, permutations that need to be assigned equal likelihoods, thus explaining the empirical observations. (Note, also, that De Mere was not tempted to simply assign equal likelihoods to all sum values though that could also be motivated as a "uniform assumption"!)
- Betrand's paradox asks us to compare the length of a randomly chosen chord of a circle with the length of a side of an equilateral triangle inscribed in the circle [15]. One way to answer this question is to assume that a random chord is obtained by choosing two points on the circumference of the circle with uniform probabilities, whereas another is to assume that the chord is chosen to intersect its orthogonal radius with uniform probability along its length. These options are easily shown to answer our question with values $1/3$ and $1/2$, respectively yet both model ignorance with uniform distributions. In fact, Betrand shows yet another reasonable construction which leads to an answer of $1/4$.

These examples demonstrate that the assignment of equal probabilities to all outcomes is not in general a valid way to model ignorance in a probabilistic fashion. In fact, it is not even a well-defined prescription, since a uniform distribution generally becomes non-uniform under a non-linear transformation of variables. In each case, the derivation of a suitable prior requires that the actual processes from which the measurements are derived be understood in sufficient detail for plausible assumptions to be made. We now turn to a proposal for such an analysis in pattern recognition.

## 3. Determination curves

From the discussion above, it is clear that a realistic functional prior should reflect the fact that feature variables are not arbitrary collections of numbers. One way to do so would be to express the prior in terms of geometric smoothness measures in feature space: functions with unrealistically low smoothness could then be given appropriately small prior probabilities. However, any specific set of smoothness measures chosen amounts to a parametric assumption about the data distribution, which we wish to avoid in this general setting. Therefore, we choose to focus on the input-output relationship between features and classes, rather than the geometry of feature space itself.

In particular, we ask how effective various subsets are in distinguishing the different classes from one another realistic pattern-recognition feature sets invariably have the property that small subsets of features have limited discriminatory power, with increasingly large sets leading to improved classification up to some limit (corresponding to the Bayes error of that overall feature set). The determination curve of a classification problem in a given $D$-dimensional feature space, then, is a sequence of error rates as a function of $d \leq D$: for each $d$, it equals the lowest Bayes error rate for any $d$-dimensional subspace of the full feature space.

The determination curve has a number of properties that reveal important characteristics of the classification task under investigation: its initial value (for $d = 1$) reflects the discriminative power of the single most informative feature, and its slope is a measure of the incremental benefit of features added to the set of active features. In principle, this descriptor also has a number of problem-independent characteristics: it is not affected by a reparametrisation of the feature space, and is monotonically non-increasing. (Since each subspace for $d_1 < d_2$ is included in the subspace for $d_2$, the Bayes error rate cannot increase when $d$ increases.) For a practical estimator of the Bayes error, neither of these statements may be strictly true, but they will remain as tendencies for reasonable estimators, as we see below.

For our current purposes, we do not consider how the curve changes with the size of the training set, though that is obviously also a rich source of information.

## 4. Experimental determination curves

In order to investigate the behaviour of determination curves on practical problems, we have computed approximate curves for a number of standard problems from the UCI database [16]. These problems are summarised in Table 1. Two approximations were required to ensure computational feasibility:

- As an approximation of the Bayes error rate, we consistently use the 1-nearest neighbour error estimate obtained with leave-one-out cross validation.
- To avoid the combinatorial explosion that would result if all d-dimensional subspaces of $D$ dimensional space were evaluated, we perform sequential forward selection: we first select the single feature with the lowest Bayes error rate, and then successively add the feature that results in the lowest (estimated) error rate to the set of selected features.

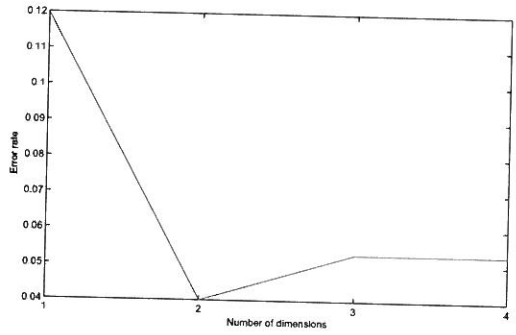
Although both approximations are known to over-estimate the Bayes error rate systematically, they are sufficiently accurate to allow us to deduce general trends.

Table 1: *Summary of classification problems used in experimental investigation*

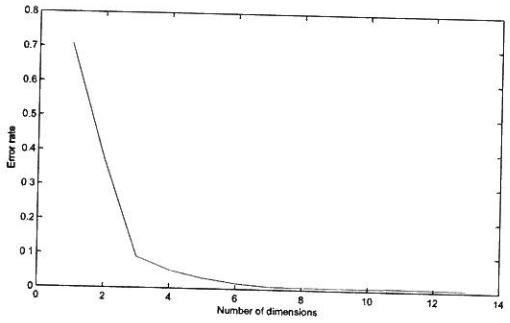
| Problem | Number of samples | Number of dimensions | Number of classes |
|---|---|---|---|
| Iris | 150 | 4 | 3 |
| Vowel-context | 990 | 13 | 15 |
| Yeast | 800 | 8 | 10 |
| Wdbc | 569 | 30 | 2 |

The determination curve for the widely-used *iris* data set is shown at the top of Fig. 1. This is a typical trend for an "easy" problem: the asymptotic error rate is close to zero, and only a small number of features (two, in this case) are required to attain that level of performance. For the *vowel-context* set in Fig. 1, the asymptotic error rate is even lower; however, the number of features required to reach that level of performance is somewhat larger. The other two problems are substantially harder, with asymptotic error rates of approximately 20% and 45% for the *yeast* and *wdbc* data sets, respectively. However, they differ in the sense that each additional feature (except the last) adds to the accuracy of *yeast*, whereas *wdbc* reaches asymptotic accuracy with fewer than half of all the features.

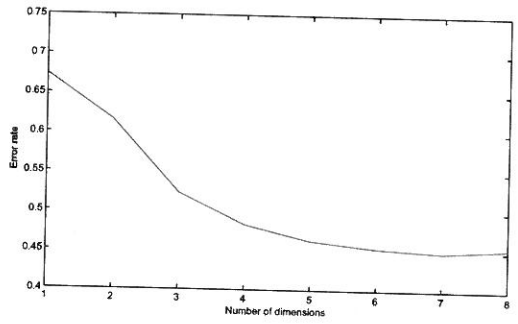
These trends are indicative of what we expect for classification problems: a smooth curve that descends from the best single-feature Bayes error to the asymptotic Bayes rate, and then either stays at that value, or gradually increases if the classifier is not able to treat additional features appropriately. Such regularities capture what we mean by a "feature" or "variable" in pattern recognition namely, that it (to a greater or lesser degree) provides information relevant to the classification task. Therefore, the determination curves provide a way to characterise the expected behaviour of a classification task in realistic terms, in contrast to the uniform functional prior described above. We therefore propose that this is a sensible basis for the construction of functional priors, with the prior probability of a given hypothesis function being determined by the likelihood of the corresponding determination curve.



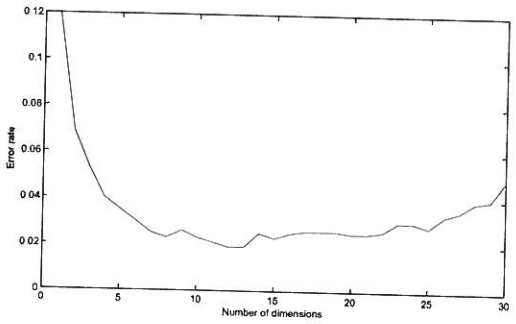

(a) Iris




(b) Vowel-context




(c) Yeast




(d) Wdbc

Figure 1: *Determination curves for four classification tasks from the UCI database.*

## 5. Summary and outlook

We have motivated an analytic tool that can serve as an alternative basis for the construction of prior probabilities in the Extended Bayes Framework (or similar environments, where one wishes to assign likelihoods to entire classification problems). Although this tool is not intuitively straightforward, is does manage to describe some important regularities of classification problems without the need for an explicit geometrical parametrisation.

As mentioned in Section 2, the original derivation of NFL treated the uniform prior $P(f)$ as a tool for calculating expectation values that reflect the intuition of ignorance [1]. That research also suggests that this intuition could be captured in other ways for example, by uniformly averaging over choices for $P(f)$. Although we suspect that arguments similar to those offered here will apply to that perspective as well, it is important to note that our discussion has been focused on the specific assumption for $P(f)$.

For practical applications, one would need to parametrise the determination curve, and assign a probability density function over the space of allowed parameters. This process is simplified by the smoothness and monotonicity of this curve; one may even be tempted to assign a uniform density over an acceptable range of its curvature, initial values and asymptotic values. However, as pointed out in Section 2, care should be exercised when ignorance is modelled with uniform distributions! A more detailed generic model of data generation processes would probably be required to make these choices in a principled way.

By expressing our problem characteristics in terms of Bayes error rates, we were able to avoid choosing a particular parametrisation of feature space. This is useful for theoretical analysis, but prevents us from using this tool directly to compare classifiers with one another. It would therefore be practically important to investigate the relationship between this description and descriptors that are more directly tied to the geometry of feature space.

## 6. References


[1] D. Wolpert, "The relationship between the various supervised learning formalisms," in *The Mathematics of Generalization*, 1994.

[2] ——, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, pp. 1341 – 1390, 1996.

[3] ——, "Bayesian and computational learning theory," in *Encyclopedia of Cognitive Science*, 2006.

[4] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley, 2nd edition, 2001.

[5] T. Roos, P. Grünwald, P. Myllymäki, and H.Tirri, "generalization to unseen cases," in *Advances in Neural Information Processing Systems*, vol. 18, 2005, pp. 1129–1136.

[6] H. Zhu and R. Rohwer, "No free lunch for cross-validation," *Neural Computation*, vol. 8, no. 7, pp. 1421–1426, 1996.

[7] C. Goutte, "Note on free lunches and cross-validation," *Neural Computation*, vol. 9, pp. 1245 – 1249, 1997.

[8] R.Vilalta and Y.Drissi, "A perspective view and survey of metalearning," *Artificial Intelligence Review*, vol. 18, no. 2, pp. 77 – 95, 2002.

[9] E. Barnard, "Determination and the "no free lunch" paradox," Submitted for publication, 2009.

[10] J. De Villiers and E. Barnard, "Backpropagation neural nets with one and two hidden layers," *IEEE Transactions on Neural Networks*, vol. 4, no. 1, pp. 136 –141, 1993.

[11] C. Van der Walt and E. Barnard, "Data characteristics that determine classifier performance," *SAIEE Africa Research Journal*, vol. 98, no. 3, pp. 87–92, 2007.

[12] P. Brazdil, J. Gama, and B. Henery, "Characterizing the applicability of classification algorithms using meta-level learning," in *Proceedings of the European Conference on Machine Learning*, 1994, pp. 83–10.

[13] T. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE. Trans. on Pattern Analysis and Mach. Intell.*, vol. 24, no. 3, pp. 289–300, 2002.

[14] O.Ore, "Pascal and the invention of probability theory," *American Mathematical Monthly*, pp. 409–19, 1960.

[15] M. Forster, "How do simple rules 'fit to reality' in a complex world?" *Minds and Machines*, vol. 9, no. 4, pp. 543 – 564, 1999.

[16] A. Asuncion and D. Newman, "UCI machine learning repository," http://www.ics.uci.edu/ mlearn/MLRepository.html, 2007.