IDIAP RESEARCH REPORT

Are two classifiers performing equally? – a treatment using Bayesian Hypothesis Testing

David Barber [a]

IDIAP–RR 04-57

May 2004

[a] IDIAP

IDIAP Research Institute www.idiap.ch
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-57

Are two classifiers performing equally? – a treatment using Bayesian Hypothesis Testing

David Barber

May 2004

Abstract. We consider here how to assess if two classifiers, based on a set of test error results, are performing equally well. This question is often considered in the realm of sampling theory, based on classical hypothesis testing. Here we present a simple Bayesian treatment that is quite general, and also is able to deal with the (practically common) case where the errors that two classifiers make are *dependent*.

Introduction

This small note is an introduction to Bayesian statistics applied to the analysis of experimental results. In particular, the situation that we are interested in is how to tell if two machine learning classifiers are performing the same[1]. This is a standard problem in assessing the results of a machine learning experiment[2]. Let's say classifier A makes 20 errors, and 35 correct classifications, whereas classifier B makes 23 errors and 32 correct classifications – is classifier A really better than classifier B? Of course, intuitively, the uncertainty stems from the small amount of results. If we had rather that classifier A makes 20000 errors and 35000 correct classifications, classifier B makes 23000 errors and 32000 correct classifications, intuitively, we would be much more confident than in the previous scenario that classifier A is better than classifier B.

Initially, we will try to answer here the above question by appealing to a basic Bayesian analysis. In doing this, we assume that the errors/labels that two classifiers make do not depend on each other. Later we'll relax this to show how one can make a test to see if they are dependent.

This small note is essentially a classic exercise in Bayesian statistics. The interested reader may consult [Mac03] (chapter 37 is most closely related to this area, although the whole of part IV is relevant) and references therein for an introduction to the general area.

Error Analysis

Consider a situation where two classifiers A and B have been tested on some data, so that we have, for each example in the test set, an error pair

$$(e_a(\mu), e_b(\mu)), \mu = 1, \dots P$$

where P is the number of test data points, and $e_a \in \{1, \dots Q\}$ (and similarly for e_b). That is, there are Q possible types of error that can occur. This is useful in text classification, where TruePositive, FalseNegative, TrueNegative and FalsePositive might form four kinds of 'errors'. For notational simplicity we also call a TruePositive an 'error'. It might be more appropriate to use a term such as 'outcome label', although this should also not be confused with the class label of the classifier – it is their evaluation against the truth that we are interested in.

Let's call $\mathbf{e}_a = \{e_a(\mu), \mu = 1, \dots, P\}$, the sample set A, and similarly for $\mathbf{e}_b$.

We are here interested in the question :

'How much evidence is there supporting that the two classifiers are performing differently?'

Mathematically, our major *assumption* here is that this is the same question as :

'How much evidence is there in favour of two sample sets being from different multinomial distributions?'

The main question that we address here is to test whether or not two classifiers are essentially performing the same. To do this, we have two hypotheses :

1. H_{indep} : The sample sets are from different distribution.
2. H_{same} : The sample sets are from the same distribution.

[1] The theory is readily extensible to multiple classifiers, and is left as an exercise for the interested reader.

[2] Ideally, a true Bayesian will use a Bayesian Classifier, for which there will always, in principle, be a direct way to estimate the suitability of the model in explaining the experimental data. We consider here the less fortunate situation where two non-Bayesian classifiers have been used, and only their test performances are available for evaluating the classifiers.

We need then to formally mathematically state what these two hypotheses mean. In both cases, however, we will make the *independence of trials assumption*

$$p(\mathbf{e}_a, \mathbf{e}_b) = \prod_{\mu} p(e_a(\mu), e_b(\mu)).$$

H_{indep} : Independence of classifiers

For H_{indep}, we assume

$$p(e_a(\mu), e_b(\mu)|H_{indep}) = p(e_a(\mu)|H_{indep})p(e_b(\mu)|H_{indep})$$

Note that we don't *need* to make this independence-of-errors assumption, since it is often quite reasonable to assume that both classifiers will tend to perform well on the same 'easy' example, and perhaps poorly on the same 'difficult' example. We'll consider this case in a later section.

Since each classifier can make one of Q types of errors, we need to specify what the probability of making such an error could be. For classifier A, we write

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q), \qquad \sum_q \alpha_q = 1$$

and similarly for $\boldsymbol{\beta}$. (These are the values of the probability tables for generating errors).

Hence, under the independence assumption a probability for generating an error on the μ^{th} example,

$$p(e_a(\mu)|\boldsymbol{\alpha}, H_{indep})$$

and likewise for classifier B.

The data likelihood

Since the data is assumed generated by a multinomial distribution, the likelihood of generating the training data is

$$p(\mathbf{e}_a|\boldsymbol{\alpha}) = \prod_{q=1}^{Q} \alpha_q^{c_q^a}$$

where c_q^a is the number of times that classifier A makes error q[3]. A similar expression holds for classifier B.

Dirichlet Prior

Since we are dealing with multinomial distributions, it is convenient to use the Dirichlet prior, which is conjugate to the multinomial distribution:

[3] The kinds of errors are assumed mutually exclusive and exhaustive. Clearly, the exhaustive condition means that there are dependencies in the errors produced at one time – this is taken care of by the constraint that the probabilities sum to one. Also, a potential source of confusion is whether or not we view the dataset errors as a sample from a distribution with a fixed total number of errors, say 50 TruePositve, 76 FalseNegative, etc, which would add combinatorial prefactors to the data likelihoods. Clearly, also in this case, if we know the total number of datapoints, then the errors are also not independent. Essentially, in the context here, using such a system is incorrect, since we only have a *single* dataset, and our interest is in the likelihood of generating this single dataset, and not in the likelihood of generating error counts.

$$p(\boldsymbol{\alpha}) = \frac{1}{Z(\mathbf{u})} \prod_q \alpha_q^{u_q - 1}$$

where

$$Z(\mathbf{u}) = \frac{\prod_{q=1}^{Q} \Gamma(u_q)}{\Gamma\left(\sum_{q=1}^{Q} u_q\right)}$$

The prior parameter $\mathbf{u}$ controls how strongly the mass of the distribution is pushed to the corners of the simplex. Setting $u_q = 1$ for all q corresponds to a uniform prior. The uniform prior assumption is reasonable, although there may be situations where it would preferable to use non-uniform priors[4].

H_{indep} : Model Likelihood

In the Bayesian framework, we want to find how likely it is that a model/hypothesis is responsible for generating the data:

$$p(H_{indep}|\mathbf{e}_a, \mathbf{e}_b) = p(\mathbf{e}_a, \mathbf{e}_b|H_{indep})p(H_{indep})/p(\mathbf{e}_a, \mathbf{e}_b)$$

where $p(H_{indep})$ is our prior belief that H_{indep} is the correct hypothesis. Note that the normalising constant $p(\mathbf{e}_a, \mathbf{e}_b)$ does not depend on the hypothesis. Then

$$\begin{aligned} p(\mathbf{e}_a, \mathbf{e}_b)p(H_{indep}|\mathbf{e}_a, \mathbf{e}_b) &= \int p(\mathbf{e}_a, \mathbf{e}_b|\boldsymbol{\alpha}, \boldsymbol{\beta}, H_{indep})p(\boldsymbol{\alpha}, \boldsymbol{\beta}|H_{indep})p(H_{indep})d\boldsymbol{\alpha}d\boldsymbol{\beta} \\ &= p(H_{indep}) \int p(\mathbf{e}_a|\boldsymbol{\alpha}, H_{indep})p(\boldsymbol{\alpha}|H_{indep})d\boldsymbol{\alpha} \int p(\mathbf{e}_b|\boldsymbol{\beta}, H_{indep})p(\boldsymbol{\beta}|H_{indep})d\boldsymbol{\beta} \end{aligned}$$

where we assumed (pretty innocuously) $p(\boldsymbol{\alpha}, \boldsymbol{\beta}|H_{indep}) = p(\boldsymbol{\alpha}|H_{indep})p(\boldsymbol{\beta}|H_{indep})$. Let's calculate

$$\int p(\mathbf{e}_a|\boldsymbol{\alpha}, H_{indep})p(\boldsymbol{\alpha}|H_{indep})d\boldsymbol{\alpha} = \frac{1}{Z(\mathbf{u})} \prod_q \alpha_q^{c_q^a + u_q - 1} = \frac{Z(\mathbf{u} + \mathbf{c}^a)}{Z(\mathbf{u})}$$

Hence

$$p(H_{indep}|\mathbf{e}_a, \mathbf{e}_b) = p(H_{indep}) \frac{Z(\mathbf{u} + \mathbf{c}^a)}{Z(\mathbf{u})} \frac{Z(\mathbf{u} + \mathbf{c}^b)}{Z(\mathbf{u})}$$

H_{same} : Model Likelihood

In H_{same}, the hypothesis is that the errors for the two classifiers are generated from the *same* multinomial distribution. Hence

$$\begin{aligned} p(\mathbf{e}_a, \mathbf{e}_b)p(H_{same}|\mathbf{e}_a, \mathbf{e}_b) &= p(H_{same}) \int p(\mathbf{e}_a|\boldsymbol{\alpha}, H_{same})p(\mathbf{e}_b|\boldsymbol{\alpha}, H_{same})p(\boldsymbol{\alpha}|H_{same})d\boldsymbol{\alpha} \\ &= p(H_{same}) \frac{Z(\mathbf{u} + \mathbf{c}^a + \mathbf{c}^b)}{Z(\mathbf{u})} \end{aligned}$$

[4]We could also use $p(\theta)$ to specify a distribution of priors $p(\mathbf{u}|\theta)$, over which one can then integrate.

Bayes Factor

If we assume that we have no prior preference for either hypothesis ($p(H_{indep}) = p(H_{same})$), then

$$\frac{p(H_{indep}|\mathbf{e}_a, \mathbf{e}_b)}{p(H_{same}|\mathbf{e}_a, \mathbf{e}_b)} = \frac{Z(\mathbf{u} + \mathbf{c}^a)Z(\mathbf{u} + \mathbf{c}^b)}{Z(\mathbf{u})Z(\mathbf{u} + \mathbf{c}^a + \mathbf{c}^b)}$$

This is the evidence to suggest that the data were generated by two different multinomial distributions. In other words, this is the evidence in favour of the two classifiers being different.

Examples

In the experiments that I demonstrate here and elsewhere, I'll assume that there are three kinds of 'errors', $Q = 3$.

- We have the two error counts $\mathbf{c}^a = [39, 26, 35]$ and $\mathbf{c}^b = [63, 12, 25]$

  Then, the above Bayes factor is 20.7 – strong evidence in favour of the two classifiers being different. (This is consistent with the model I used to generate the data – they were indeed from different multinomial distributions).

- Alternatively, if we have the two error counts $\mathbf{c}^a = [52, 20, 28]$ and $\mathbf{c}^b = [44, 14, 42]$

  Then, the above Bayes factor is 0.38 – weak evidence against the two classifiers being different. (This is consistent with the model I used to generate the data – they were indeed from the same multinomial distributions)

- As a final example, consider counts $\mathbf{c}^a = [459, 191, 350]$ and $\mathbf{c}^b = [465, 206, 329]$. This gives a Bayes factor of 0.008 – strong evidence that the two classifiers are statistically the same (Indeed, the errors were in this case generated by the same multinomial).

These results show that the Bayesian analysis performs in a way that is consistent with the intuition that the more test data we have, the more confident we are in our statements about which is the better model.

Dependent Error Analysis

Here we consider the (perhaps more common) case that errors are *dependent*. For example, it is often the case that if classifier A works well, then classifier B will also work well. Similarly, if one classifier performs poorly, then often the other will too. Here, we assume that dependencies exist, but we make no preferences for one to another (of course, such preferences would be straightforward to include if desired). (There may be some interest in situations where if classifier A performs poorly, then classifier B is likely to perform well). Thus we want to consider the Hypothesis

H_{dep} : the errors that the two classifiers make are dependent.

For convenience, let's write

$$\mathbf{e} = (\mathbf{e}_a, \mathbf{e}_b)$$

Mathematically, we need to specify a distribution:

$$p(e_a(\mu), e_b(\mu)|H_{dep}) = p(e_a(\mu), e_b(\mu)|\mathbf{P}, H_{dep})$$

here $\mathbf{P}$ is a $Q \times Q$ matrix of probabilities:

$$[P]_{ij} = p(e_a = i, e_b = j)$$

namely that the ij element of $\mathbf{P}$ is the probability that A makes error i, and B makes error j.

Then, as before

$$\begin{aligned}\frac{1}{p(H_{dep})}p(\mathbf{e})p(H_{dep}|\mathbf{e}) &= p(\mathbf{e}|H_{dep}) \\ &= \int p(\mathbf{e},\mathbf{P}|H_{dep})d\mathbf{P} \\ &= \int p(\mathbf{e}|\mathbf{P},H_{dep})p(\mathbf{P}|H_{dep})d\mathbf{P}\end{aligned}$$

Assuming a Dirichlet prior on $\mathbf{P}$, with parameters $\mathbf{U}$, we have

$$p(\mathbf{e})p(H_{dep}|\mathbf{e}) = p(H_{dep})\frac{Z(vec\,(\mathbf{U}+\mathbf{C}))}{Z(vec(\mathbf{U}))}$$

where $vec(\mathbf{D})$ simply forms a vector by concatenating the rows of the matrix $\mathbf{D}$. Here $\mathbf{C}$ is the count matrix, with $[C]_{ij}$ equal to the number of times that joint error $(e_a = i, e_b = j)$ occurred in the P datapoints. As before, we can then use this in a Bayes factor calculation. For the uniform prior, $[U]_{ij} = 1, \forall i, j$.

Testing for dependencies in the classifiers

Imagine that we wish to test whether or not the errors of the classifiers are dependent H_{dep}, against the hypothesis that they are independent H_{indep}.

$$\frac{p(H_{indep})}{p(H_{dep})} = \frac{Z(\mathbf{u}+\mathbf{c}^a)}{Z(\mathbf{u})}\frac{Z(\mathbf{u}+\mathbf{c}^b)}{Z(\mathbf{u})}\frac{Z(vec(\mathbf{U}))}{Z(vec\,(\mathbf{U}+\mathbf{C}))}$$

Examples

- Consider the error count matrix $\mathbf{C}$

  $$\begin{pmatrix} 98 & 7 & 93 \\ 168 & 13 & 163 \\ 245 & 12 & 201 \end{pmatrix}$$

  so that $\mathbf{c}^a = [511, 32, 457]$, and $\mathbf{c}^b = [198, 344, 458]$. Then

  $$\frac{p(H_{indep})}{p(H_{dep})} = 3020$$

  – very strong evidence that the classifiers perform independently (indeed, this is consistent with the way I generated the data).

- Consider the error count matrix $\mathbf{C}$

  $$\begin{pmatrix} 82 & 120 & 83 \\ 107 & 162 & 4 \\ 170 & 203 & 70 \end{pmatrix}$$

  so that $\mathbf{c}^a = [359, 485, 156]$, and $\mathbf{c}^b = [284, 273, 443]$. Then

  $$\frac{p(H_{indep})}{p(H_{dep})} = 2 \times 10^{-18}$$

  – extremely strong evidence that the classifiers perform dependently (again, this is consistent with the way I generated the data).

Dependent vs the Same

Perhaps the most useful test that can be done practically is between the H_{dep} versus H_{same}. This is because, in practice, it is reasonable to believe that dependencies are quite likely in the errors that classifiers make (both classifiers will do well on 'easy' test examples, and badly on 'difficult' examples). In this sense, it is natural to believe that dependencies will most likely exist in practice. The relevant question is : are these dependencies strong enough to make us believe in fact that the errors are coming from the *same* process? In this sense, we want to test

$$\frac{p(H_{same})}{p(H_{dep})} = \frac{Z(\mathbf{u} + \mathbf{c}^a + \mathbf{c}^b)}{Z(\mathbf{u})} \frac{Z(vec(\mathbf{U}))}{Z(vec\,(\mathbf{U} + \mathbf{C}))}$$

- Consider an experiment which gives the test error count matrix $\mathbf{C}$

  $$\begin{pmatrix} 105 & 42 & 172 \\ 42 & 45 & 29 \\ 192 & 203 & 170 \end{pmatrix}$$

  so that $\mathbf{c}^a = [339, 290, 371]$, and $\mathbf{c}^b = [319, 116, 565]$. Then

  $$\frac{p(H_{same})}{p(H_{dep})} = 4.5 \times 10^{-38}$$

  – extremely strong evidence that the classifiers are performing differently (this is consistent with the way I generated this data set).

- Consider an experiment which gives the test error count matrix $\mathbf{C}$

  $$\begin{pmatrix} 15 & 8 & 10 \\ 5 & 4 & 8 \\ 13 & 12 & 25 \end{pmatrix}$$

  so that $\mathbf{c}^a = [33, 24, 43]$, and $\mathbf{c}^b = [33, 17, 50]$. Then

  $$\frac{p(H_{same})}{p(H_{dep})} = 42$$

  – strong evidence that the classifiers are performing the same (this is consistent with the way I generated this data set).

Discussion

We presented a simple application of Bayesian analysis to assessing if the results from two classification experiments can be considered to be equivalent. The formulae derived are straightforward, since they are based on a simple Dirichlet prior over error distributions, and readily applicable in practical situations. An extension of this work would be to summarise the posterior distribution of the errors by computing standard confidence intervals around the given values, under various assumptions such as independence or dependence of the classifiers.

References

[Mac03] D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2003.