Which method learns most from the data?

Methodological issues in the analysis of comparative studies

A. Feelders
University of Twente
Department of Computer Science
P.O.Box 217, 7500 AE Enschede
The Netherlands
e-mail: feelders@cs.utwente.nl

W. Verkooijen
Tilburg University
Department of Economics
P.O. Box 90153, 5000 LE Tilburg
The Netherlands
e-mail: W.J.H.Verkooijen@kub.nl

Introduction

The mutual discovery of the statistical and artificial intelligence communities (see e.g. [Han93, CO94]) has resulted in many studies which compare the performance of statistical and machine learning methods on empirical data sets; examples are the StatLog project ([MST94]) and the Santa Fe Time Series Competition ([WG94]), as well as numerous journal articles ([KWR93, RABCK93, WHR90, TAF91, TK92, FG93]).

What has struck us is the casual manner comparisons are typically carried out in the literature. The ranking of k preselected methods is performed by training (estimating in statistical terminology) them on a single data set, and estimating their respective mean prediction errors (MPE) from a hold-out sample. The methods are, subsequently, ranked according to their estimated MPEs. When the total number of observations is small, usually cross-validation rather than a hold-out sample is used to estimate the mean prediction errors.

A more rigourous comparison of methods should include significance testing rather than giving a mere ranking based on the estimated MPEs. The statistical analysis of comparative studies, method ranking in particular, is addressed in this paper. Specifically, we address *methodological* issues of studies in which the performance of several regression or classification methods is compared on empirical data sets.

1 Ranking Methods by Significance Testing

The ranking of methods by simply ordering them by their estimated prediction errors should be extended by statistical significance testing. Appropriate tests are those for the difference between means (regression) and proportions (classification). The standard t-test for testing the difference between two sample means $\overline{Y}_1$ and $\overline{Y}_2$ which come from *independent* normal distributed populations, leads to the following confidence interval for the difference

$$\theta_1 - \theta_2 \in [(\overline{Y}_1 - \overline{Y}_2) \pm t_{(\alpha/2,\nu)}\, \hat{\sigma}_{\text{diff}}] \qquad (1)$$

where $\hat{\sigma}_{\text{diff}}$ equals $\sqrt{\hat{\sigma}^2_{\overline{Y}_1} + \hat{\sigma}^2_{\overline{Y}_2}}$. In the standard comparative experiment, however, the MPEs are all estimated from the *same* test sample, which makes them highly correlated. Therefore, a *paired sample* t-test should be used instead. The dependence within the pairs only changes the standard error of the difference $\hat{\sigma}_{\mathrm{diff}}$, which now becomes

$$\hat{\sigma}_{\mathrm{diff}} = \sqrt{\hat{\sigma}^2_{\overline{Y}_1} + \hat{\sigma}^2_{\overline{Y}_2} - 2\,\mathrm{cov}(\overline{Y}_1, \overline{Y}_2)} \qquad (2)$$

When the variables are positively correlated the covariance will have a positive value and thus the variance and standard error of a difference between means will be *less* for matched than for unmatched samples. Consequently, the confidence intervals become smaller (given the same α value), which results in more powerful tests. In conclusion, neglecting the dependence between the samples generally results in too conservative tests.

Often the estimated prediction errors of more than two, say k, methods are being compared. The first idea that comes to mind is to test each possible difference by a paired t-test with P(Type I error) of α. The problem with this approach is that the probability of making at least one Type I error in the whole family of t-tests exceeds α, by an amount that increases with the number of tests that are made. For J *statistically independent* tests the probability of making at least one Type I error, better known as the *familywise error* rate (FWE), is $1 - (1 - \alpha)^J$. When J is large, say 20, this can be a large probability; for $\alpha = 0.05$ there will be a probability of 0.64 for one or more Type I errors. When the tests are statistically dependent of each other, such as pairwise difference tests, the FWE becomes even larger. Thus, when enough pairwise tests are performed one will with high probability find one or more 'significant' differences. This problem is known as the *multiplicity effect* or *selection effect*. Statistical procedures have been designed to take into account and properly control for the multiplicity effect, they are called *multiple comparison procedures*.

A crude approach to deal with the multiplicity effect is the Bonferroni method, which rejects the pairwise null hypothesis $\theta_i - \theta_{i'} = 0$ when the p-value is less than α/J, where α is the preset FWE level and J is the number of tests. This method, neglects the dependency between the pairwise difference tests, and it further assumes normality of the data.

Many alternative tests ranging from slight adjustments to the Bonferroni method to very sophisticated techniques have been developed [HT87, WY93] and still are being developed. The characteristics of a particular experimental design often prescribe adjustments to general tests for differences or make special purpose tests necessary. The experimental design that captures the subject of this study is the *one-way repeated measures design*, which is displayed in Table 1. In such designs blocks consisting of a random sample of, say, n experimental units drawn from a large population constitute the random factor. Each unit is measured under k different conditions. The conditions of measurements are fixed in advance, and constitute the treatment factor. In the terminology of this study experimental units correspond to the observations from the test set, and the treatment factor corresponds to the regression or classification model type.

2 Pairwise Comparisons for Regression

The general setting of this section is the one-way repeated measures design with k different (prediction) models which predict the observations from the *same* random test set of size n. The deviation of the predicted value from the true value is assumed to be measured as squared error, but any other error measure could be inserted equally well. When the observations are not randomly drawn from a population but result from a (highly) autocorrelated time series,

	Functions						
Observations	f_1	f_2	$\dots$	f_i	$\dots$	f_k	Total
1	Y_{11}	Y_{12}	$\dots$	Y_{1i}	$\dots$	Y_{1k}	$Y_{1.}$
2	Y_{21}	Y_{22}	$\dots$	Y_{2i}	$\dots$	Y_{2k}	$Y_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
j	Y_{j1}	Y_{j2}	$\dots$	Y_{ji}	$\dots$	Y_{jk}	$Y_{j.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_{n1}	Y_{n2}	$\dots$	Y_{ni}	$\dots$	Y_{nk}	$Y_{n.}$
Total	$Y_{.1}$	$Y_{.2}$	$\dots$	$Y_{.i}$	$\dots$	$Y_{.k}$	
Means	$\overline{Y}_{.1}$	$\overline{Y}_{.2}$	$\dots$	$\overline{Y}_{.i}$	$\dots$	$\overline{Y}_{.k}$	

Table 1: One-way repeated measures lay-out.

the subsequent approach seems not to be justified. Diebold [DM94] discusses the comparison of predictive accuracy of two time series models; he leaves the multiple comparison problem for further research.

Let $\mathbf{Y}_j = (Y_{j1}, Y_{j2}, \dots, Y_{jk})$ denote the vector of prediction errors for the jth observation $(1 \le j \le n)$. The following model is assumed:

$$\mathbf{Y}_j = \mathbf{M}_j + \mathbf{E}_j \;\; (1 \le j \le n) \tag{3}$$

where all the $\mathbf{M}_j = (M_{j1}, M_{j2}, \dots, M_{jk})$ and $\mathbf{E}_j = (E_{j1}, E_{j2}, \dots, E_{jk})$ are distributed independently of each other as k-variate normal vectors, the former with mean vector $\boldsymbol{\theta}$=$(\theta_1, \theta_2, \dots, \theta_k)$ (the vector of model effects) and variance-covariance matrix $\boldsymbol{\Sigma}_0$ and the latter with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}$. Thus the $\mathbf{Y}_j$'s are independent and identically distributed (i.i.d.) $N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ random vectors where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + \sigma^2\mathbf{I}$.

Exact procedures for making pairwise comparisons among the θ_i's can be constructed if we impose special restrictions on the form of $\boldsymbol{\Sigma}$. The least restrictive of such models for a one-way repeated measures design is the *spherical model*. In words, this model assumes that all pairwise differences of the sample means of the regression models have the same variance (for more details see [HT87, CH90, WBM91, Hay88]). This assumption, however, will rarely be satisfied in practice [HT87, Hay88].

Hochberg and Tamhane [HT87, page 215] propose a test that is to be preferred in case one is unsure about the sphericity assumption being satisfied. They propose the following approximate $100(1-\alpha)\%$ simultaneous confidence intervals for the pairwise differences $\theta_i - \theta_{i'}$:

$$\theta_i - \theta_{i'} \in \left[\overline{Y}_{.i} - \overline{Y}_{.i'} \pm |M|^{(\alpha)}_{k^*,n-1}\sqrt{\frac{S_{ii} + S_{i'i'} - 2S_{ii'}}{n}}\right] \;\; (1 \le i < i' \le k) \tag{4}$$

where $|M|^{(\alpha)}_{k^*,n-1}$ is the upper α point of the Studentized maximum modulus distribution (see [HT87, Table 6]) with parameter $k^* = k(k-1)/2$ and degrees of freedom $n-1$; and where

$$S_{ii'} = \frac{\sum_{j=1}^{n}(Y_{ji} - \overline{Y}_{.i})(Y_{ji'} - \overline{Y}_{.i'})}{n-1} \;\; (1 \le i, i' \le k) \tag{5}$$

An empirical analysis of the *boston housing data*[1] illustrates this procedure. We compare the performance of four regression models, designated f_1 through f_4. The first model f_1 is a linear model trained with OLS; models f_2 through f_4 are feed-forward neural networks with respectively 2, 4 and 6 hidden units. The data set is split into two parts: the first part (400) is used to estimate the parameters of the model; the second (106) to measure the model's performance. The observed average squared prediction error $\overline{Y}_{.i}$ for method f_i $(i = 1, 2, 3, 4)$ are $\overline{Y}_{.1} = 6.38e - 3$; $\overline{Y}_{.2} = 3.49e - 3$; $\overline{Y}_{.3} = 2.94e - 3$; and $\overline{Y}_{.4} = 3.00e - 3$.

Suppose that it is of interest to make all pairwise comparisons among the prediction accuracy of the different methods with Type I familywise error rate $\alpha = 0.10$. In Table 2 the $S_{ii'}$ values are displayed in the cells, calculated according to (5). We use $|M|^{(0.1)}_{6,105}$, which equals 2.135 ([HT87, Table 6]), to construct the confidence intervals for the pairwise differences $\theta_i - \theta_{i'}$ according to (4). Table 3 shows that the linear model performs significantly worse than the three neural network models. Among the neural network models no significant differences can be observed; there is no statistical evidence for preferring neural network models with more than two hidden units.

	f_1	f_2	f_3	f_4
f_1	1.68e-04	3.76e-05	3.46e-05	4.02e-05
f_2	-	3.01e-05	1.54e-05	1.27e-05
f_3	-	-	2.24e-05	2.67e-05
f_4	-	-	-	4.15e-05

Table 2: The $S_{ii'}$ matrix.

	θ_2	θ_3	θ_4
θ_1	[5.9e-4 , 5.2e-3]	[1.1e-3 , 5.7e-3]	[1.1e-3 , 5.8e-3]
θ_2	-	[-4.3e-4 , 1.5e-3]	[-9.2e-4 , 1.9e-3]
θ_3	-	-	[-7.2e-4 , 6.2e-3]

Table 3: All pairwise confidence intervals. The cell $(\theta_i, \theta_{i'})$ contains the confidence interval for the pairwise difference $\theta_i - \theta_{i'}$.

3 Pairwise Comparisons for Classification

In this section we discuss significance testing for the comparison of two or more *classification* methods. Again we notice that it is not appropriate to use a standard test based on the assumption of independent samples. Instead, we use - as suggested by Ripley ([Rip93]) - McNemar's test ([MM77]), when only two classification methods are compared. This test is normally used to test for differences between proportions in paired sample designs. The comparisons performed in this section should not be considered as serious evaluations of the methods involved, they are purely illustrative.

A hypothesis test involving the application of linear discriminant analysis and a feed-forward neural network to the *diabetes* data set[2] illustrates the use of McNemar's test. The 768 observations in this dataset were divided in a training and test set of 384 observations each.

[1] This dataset is publically available by ftp from `lib.stat.cmu.edu` with user `statlib`

[2] This data set can be obtained by anonymous ftp from `ics.uci.edu` in the directory `pub/machine-learning-databases`

	I_{nn}	C_{nn}	Total
I_{lda}	61	23	84
C_{lda}	32	268	300
Total	93	291	384

Table 4: Incorrect and Correct classifications of lda and nn

Table 4 summarizes the result of using the linear discriminant function and a neural network estimated on the training set to classify the observations in the test set. The cells (C_{lda}, C_{nn}) and (I_{lda}, I_{nn}) respectively contain the number of cases classified correctly and incorrectly by both the linear discriminant function and the neural network. Since we want to test

$$H_0 : \text{MPE}_{lda} = \text{MPE}_{nn} \qquad \text{against} \qquad H_1 : \text{MPE}_{lda} \neq \text{MPE}_{nn},$$

only the cells (I_{lda}, C_{nn}) and (C_{lda}, I_{nn}) of this table are of interest. Here MPE is defined as the proportion of misclassifications f_i makes on the population of interest. When observations falling in the (I_{lda}, C_{nn})-cell of this table are defined as a success, then the number of successes is binomially distributed with $n = (I_{lda}, C_{nn}) + (C_{lda}, I_{nn}) = 55$ and $p = 0.5$, under the null hypothesis. Since $n > 10$, a normal or chi-square approximation of the binomial distribution is sufficiently accurate, and is thus employed. Application of the chi-square version of McNemar's test to the data in Table 4 yields

$$X^2 = \frac{(|23 - 32| - 1)^2}{23 + 32} = 1.1636$$

where the X^2 statistic has a chi-square distribution with 1 degree of freedom. This value of X^2 has a p-value of approximately 0.28. According to any conventional significance level, we should conclude that H_0 cannot be rejected.

We will now consider the case where $k > 2$ classification functions are compared. In this comparison we use the same training and test set as in the above example. We performed a comparative study, including linear discriminant analysis (f_1), quadratic discriminant analysis (f_2), a classification tree (f_3), and two feed-forward neural networks (f_4 and f_5 respectively) which only differ in the value of the weight decay parameter used. All pairwise comparisons are performed which amounts to a total of $k^* = 10$ pairwise comparisons. Table 1 presents the general lay-out of a study which compares k classification functions. In this matrix Y_{ji} is one if f_i classifies observation j correctly, and zero otherwise.

For our comparative study we have: $n = 384$, $k = 5$, $Y_{.1} = 300$, $Y_{.2} = 295$, $Y_{.3} = 265$, $Y_{.4} = 303$ and $Y_{.5} = 296$. Consequently, $\overline{Y}_{.1} = 0.781$, $\overline{Y}_{.2} = 0.768$, $\overline{Y}_{.3} = 0.69$, $\overline{Y}_{.4} = 0.789$, $\overline{Y}_{.5} = 0.771$.

The pooled variance of any pairwise difference $\overline{Y}_{.i} - \overline{Y}_{.i'}$ for this design can be written as ([MM77], p. 180)

$$\hat{\sigma}^2_{\text{diff}} = \frac{2(k\Sigma_{j=1}^{n} Y_{j.} - \Sigma_{j=1}^{n} Y_{j.}^2)}{n^2 k(k-1)}$$

We can now construct $100(1-\alpha)\%$ simultaneous confidence intervals for all pairwise differences $\theta_i - \theta_{i'}$ as follows

$$\theta_i - \theta_{i'} \in \left[\overline{Y}_{.i} - \overline{Y}_{.i'} \pm Z^{\nu}_{k^*:1-\alpha/2}\hat{\sigma}_{\text{diff}}\right] \quad (1 \leq i < i' \leq k)$$

	θ_2	θ_3	θ_4	θ_5
θ_1	[–0.045,0.071]	[0.033,0.149]	[–0.066,0.05]	[–0.048,0.068]
θ_2	–	[0.02,0.136]	[–0.079,0.037]	[–0.061,0.055]
θ_3	–	–	[–0.157,–0.041]	[–0.139,–0.023]
θ_4	–	–	–	[–0.04,0.076]

Table 5: 95% confidence intervals for all pairwise differences.

where θ_i denotes the population proportion of correct classifications of f_i, and $\nu = n-1$ denotes degrees of freedom. The distribution of Z is based on the Student t distribution, adjusted for the number of comparisons k^* involved ([Dun61]). Tables for this statistic can be found in ([MM77],[Dun61]). As one would expect, the value of $Z^{\nu}_{k^*:1-\alpha/2}$ increases with the number of comparisons k^*, leading to wider confidence intervals.

In Table 5, 95% confidence intervals for all pairwise differences are provided. In computing these intervals we used $Z^{\infty}_{10:0.975} = 2.81$ and $\hat{\sigma}_{\text{diff}} = 0.0205$, leading to confidence intervals that are $2 \times (2.81 \times 0.0205) = 0.116$ wide. If the interval of $\theta_i - \theta_{i'}$ contains 0, then there is no significant evidence that classification functions f_i and $f_{i'}$ differ in their true prediction error. From Table 5 we conclude that f_3 (the classification tree) performs significantly worse than *all* other functions; among these other functions no significant difference has been found.

4 Conclusion

In this paper we proposed a first step towards a sound methodology for performing and analysing studies that compare the predictive accuracy of several regression or classification functions. Rather than providing a mere ranking, hypothesis testing is used to determine whether a significant difference between functions has been found. The formal methods required to perform the appropriate hypothesis tests originate primarily from the field of experimental design. This paper selected parts of these formal methods and showed their relevance to a type of study that is encountered frequently in the recent AI and Machine Learning literature.

Although the general difficulties induced by the multiplicity effect and by the dependency among observations are easy to grasp, finding "the right" testing procedure is much more difficult. The literature on the subject is somewhat ambiguous, and requires a high entrance level of statistical knowledge–which AI-researchers don't always possess. This may explain why comparative experiments are often performed in a rather casual way in the AI and Machine Learning literature.

We do not claim that the methods presented here are the best for the given purpose: they are *examples* of tests that can be used for the comparison of the prediction accuracy of different functions. We hope that in future research more attention will be given to this subject, and perhaps more appropriate methods will be found.

References

[CH90] M.J. Crowder and D.J. Hand. *Analysis of repeated measures.* Chapman & Hall, London, 1990.

[CO94] P. Cheeseman and R.W. Oldford, editors. *Selecting models from data: AI and statistics IV*. Lecture notes in statistics nr. 89. Springer-Verlag, New York, 1994.

[DM94] F. Diebold and R. Mariano. Comparing predictive accuracy. Technical report, University of Pennsylvania, 1994.

[Dun61] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64, 1961.

[FG93] D. Fletcher and E. Goss. Forecasting with neural networks: an application using bankruptcy data. *Information & Management*, 24:159–167, 1993.

[Han93] D. J. Hand, editor. *Artificial intelligence frontiers in statistics: AI and statistics III*. Chapman & Hall, London, 1993.

[Hay88] W. Hays. *Statistics*. Holt, Rinehart and Winston, Inc, Fort Worth, 1988.

[HT87] Y. Hochberg and A. Tamhane. *Multiple comparison procedures*. Wiley & Sons, New York, 1987.

[KWR93] J. Kim, H. Weistroffer, and R. Redmond. Expert systems for bond rating: a comparative analysis of statistical, rule-based and neural network systems. *Expert Systems*, 10:167–171, 1993.

[MM77] L. Marascuilo and M. McSweeney. *Nonparametric and distribution-free methods for the social sciences*. Brooks/Cole Publishing Company, Monterey, CA, 1977.

[MST94] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors. *Machine learning, neural and statistical classification*. Ellis Horwood, New York, 1994.

[RABCK93] A. Refenes, M. Azema-Barac, L. Chen, and S. Karoussos. Currency exchange rate prediction and neural network design strategies. *Neural Computing & Applications*, 1:46–58, 1993.

[Rip93] B.D. Ripley. Flexible non-linear approaches to classification. In V. Cherkassky, J.H. Friedman, and H. Wechsler, editors, *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*. Springer-Verlag, 1993.

[TAF91] Z. Tang, C. de Almeida, and P. Fishwick. Time series forecasing using neural networks vs. box-jenkins methodology. *Simulation*, 57:303–310, 1991.

[TK92] Kar Yan Tam and Melody Y. Kiang. Managerial applications of neural networks: the case of bank failure predictions. *Management science*, 38(7):926–947, 1992.

[WBM91] B. Winer, D. Brown, and K. Michels. *Statistical principles in experimental design*. McGraw-Hill, New York, 1991.

[WG94] A. Weigend and N. Gershenfield. *Time series prediction: forecasting the future and understanding the past*. Addison-Wesley, Reading, 1994.

[WHR90] A. Weigend, A. Huberman, and D. Rumelhart. Predicting the future: a connectionist approach. *International Journal of Neural Systems*, 1(3):193–209, 1990.

[WY93] P.H. Westfall and S.S. Young. *Resampling-Based Multiple Testing*. John Wiley & Sons, New York, 1993.