Adv Data Anal Classif (2009) 3:25–38
DOI 10.1007/s11634-009-0037-8

REGULAR ARTICLE

Where are the large and difficult datasets?

Adrien Jamain · David J. Hand

Received: 21 August 2007 / Revised: 14 January 2009 / Accepted: 20 February 2009 /
Published online: 14 March 2009
© Springer-Verlag 2009

Abstract A great many comparative performance assessments of classification rules have been undertaken, ranging from small ones involving just one or two methods, to large ones involving many tens of methods. We are undertaking a meta-analytic study of these studies, attempting to distil some overall conclusions. This paper describes just one of our observations. The dataset analysed in this paper contains 5,203 error rates taken from 45 articles and describing 146 datasets. One curious general relationship which was persistent in our data, despite the fact that we were looking at results mixed between distributions rather than conditional on distributions, was that error rate decreased with increasing dataset size. We believe this to be an artefact of the way datasets are collected by the research community.

Keywords Error rate · Meta-analysis · Comparative studies · Repositories

Mathematics Subject Classification (2000) 6207 · 68T10

1 Introduction

Supervised classification is a central component in data mining and, in recent decades, a great deal of research effort has been put into the development of new methods. This

A. Jamain (✉)
BNP-Paribas, 10 Harewood Avenue, London NW1 6AA, UK
e-mail: adrien.jamain@uk.bnpparibas.com

D. J. Hand
Department of Mathematics, Institute for Mathematical Sciences,
Imperial College, London SW7 2AZ, UK
e-mail: d.j.hand@imperial.ac.uk

has been carried out in several different research communities, including statistics, pattern recognition, machine learning, neural networks, and, most recently, data mining. This work has necessarily been accompanied by a variety of comparative studies assessing the relative merits of the various classification methods. Many of these studies are small, occurring in the context of the development of a new classification method. But others are large, being attempts to develop some broad understanding of the relative merits of different methods. However, even the largest studies bring limited insights into the circumstances under which methods perform well or poorly, for the simple reason that they are arduous and time-consuming, even for relatively modest numbers of classification methods and design sets. Because of this, calls for posterior analyses—or meta-analyses—of these studies have been made (see for example Hand 1999; Quinlan 1994).

An important example of such a study is the MetaL project (Brazdil et al. 2003; METAL Consortium 2002), whose key achievement is an automatic advisor which allows the user to submit online the characteristics of his or her dataset, and then produces a predicted ranking of classification methods based on their known performances on a set of benchmark datasets. The former Delve database of results (Rasmussen et al. 1996), which may be seen as a predecessor of the MetaL advisor, illustrates the difficulties that this kind of approach may encounter, namely that the online databases need to grow and be updated, and hence require the active support of the research community.

Several articles re-analysing results from comparative studies have also been published; for example Sohn (1999) on the Statlog study Michie et al. (1994), and Sargent (2001) on the use of neural networks in the medical area. However, to the best of our knowledge there has not been a large-scale study of the results of classification methods in the literature, and this is what we carried out (Jamain 2004; Jamain and Hand 2008). Perhaps it is worth emphasizing that this implies a completely different approach from that of the usual comparative study: we did not apply a set of classification methods to existing datasets, as this has been done many times, but we rather compiled and analysed existing results from a large set of published studies. In the course of our exploration of the resulting data we made a few interesting observations, one of which concerning the provenance of the Naive Bayes method in the Statlog study (Jamain and Hand 2005). In the present article, we describe another of these peculiar observations, our attempts to explain it, and what it implies for the practical assessment of classification methods.

2 Collecting data from comparative studies

First we describe very briefly how we collected the data, and a few problems we had to tackle. The resulting database and the whole data collection process are detailed in Jamain (2004), and Jamain and Hand (2008).

The range of different intellectual communities which have been involved in developing classification methods presents a challenge in tracking down comparative studies. Furthermore, such comparative studies have also appeared in the literature of application domains, such as medicine and speech recognition. Moreover, the studies vary in their aims and ambition. Many studies merely display a few results for a handful of methods; for example, two of the most famous ones are those of Atlas et al. (1991), and Shavlik et al. (1991). In contrast, a few studies are very large and include several tens of datasets and methods; for example, the four largest that we know of are those of Eklund and Hoang (2002), Lim et al. (2000), Michie et al. (1994), and Zarndt (1995). One can note by the way that none of these studies share a common keyword in their titles, and this illustrates the difficulty of identifying relevant studies in an automatic fashion.

In choosing which study to include in our meta-analysis, we have decided to be as broad as possible and include all sorts of studies, large and small. Our search procedure essentially consisted of an exploration of references, starting from those of the famous Statlog study (Michie et al. 1994), and a complementary online inquiry on the CiteSeer Digital Library[1] for the articles which are not published in standard journals (for example Eklund and Hoang 2002; Zarndt 1995). We looked for studies which had titles and keyword lists containing the terms *comparison*, *case study*, *benchmark*, or *experimental study*, and either *machine learning*, *pattern recognition*, or *classification*. One criterion which we had to apply was one of 'dataset popularity', that is, we discarded datasets on which less than three different method types were tested. This was to avoid obscure datasets, and to provide a relatively meaningful basis for comparisons. Other, less important, inclusion criteria are described in Jamain (2004), and Jamain and Hand (2008).

As for what type of results we would collect, we again decided to be as broad as possible and take into account all kinds of accuracy measures (error rate, weighted error rate, logarithmic score, quadratic score,…). However it turned out that the vast majority of published studies report the straightforward error rate, and hence we focused our analysis on this simple measure of performance. The data in the present article hence only consists of error rates, with equal misclassification costs.

In contrast with traditional meta-analysis where one can find—at least in theory—all the relevant results which have ever been produced, here there is potentially such a large number of results that we can only collect a relatively small sample, and hope that is reasonably representative. In the extreme, there are probably millions of daily applications of classification methods in the world, and one can adopt the view that all of them belong to the sample space. Our search for results is thus necessarily incomplete, and for this particular study we have used the data present in the final version of our database.[2]

While browsing the literature, we quickly realised that relatively little is usually reported about datasets. Sometimes, when the datasets are from the UCI repository (Blake and Merz 1998), only the name is even given, as in Viswanathan and Webb (1998), and the reader is then forced to guess from acronyms. Usually however, at least three characteristics are reported, broadly speaking: dataset 'size' (number of examples), dataset 'dimensionality' (number of variables), and dataset 'complexity' (number of classes). In some cases, as in the Statlog study, more elaborate dataset

[1] http://citeseer.nj.nec.com/.

[2] Available online: http://stats.ma.ic.ac.uk/n/nadams/public_html/classificationgroup/metaclass/.

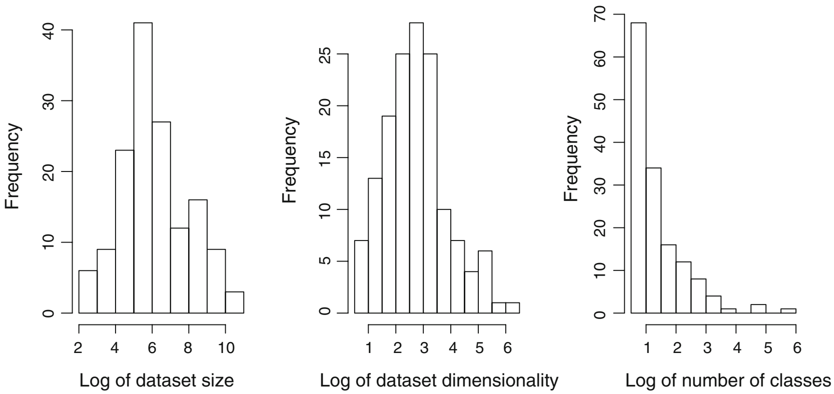

Fig. 1 Histograms of dataset characteristics (size, dimensionality, number of classes)

characteristics are available, but it is quite rare. We hence became interested in those three dataset characteristics.

However simple the three characteristics are, it is not completely trivial to decide how to define them when comparing them across datasets and studies. In the case of dataset size, we took the number of training examples when a division between training and test set was carried out, and the whole number of examples when cross-validation was used (another possibility would be to take the proportion of examples used during one fold). In the case of dimensionality, we took the final number of variables that the authors of each study reported, whether the variables were continuous, categorical, or binary. A more complicated scheme would be to transform all categorical variables to a common binary equivalent, but this was not always possible since the number of categories of each categorical variable is not systematically reported. Only the number of classes was straightforward to define. In the few cases where those characteristics were not explicit in the original studies, but where the datasets were identified in the UCI repository, we took the corresponding characteristics as those which were reported there. Finally, we log-transformed these characteristics as needed, since their distributions are heavily right-skewed as can be seen on Fig. 1. To make Fig. 1 more informative we averaged size and dimensionality within each dataset (as a dataset can be present in our data with several sizes and dimensionalities).

One thing which can be easily looked at is the relationship between these characteristics and classification 'performance'. Concerning this performance, the vast majority of studies report only error rate (with equal misclassification costs). Our database containing a few other kinds of results (e.g. quadratic scores), for the present study we restricted the data to error rate only, for the sake of simplicity. This meant 5,203 results, taken from 45 studies and related to 146 datasets. Of these 146, 101 were directly available in the UCI repository, and hence our data can be considered as representative of the set of datasets commonly used in the literature.

To give an overview of which classification methods this dataset contains, we categorized methods into different types. As there are no hard-and-fast rules to do so, our classification had to take into account various subjective factors, such as the general approach of the method (e.g. density estimation is traditionally linked with the area of

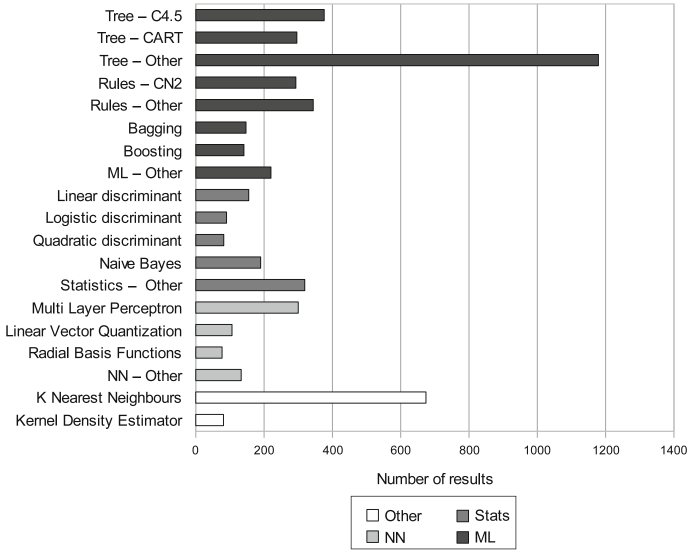

Fig. 2 Number of results per method type

Statistics), or the historical background of the authors who first published a description of the method. In our sample we have identified 64 different method types. There is a maximum of 43 different method types tested on a dataset (maximum attained for both UCI datasets *Pima Indians Diabetes* and *Image Segmentation*), and the median is 11 method types per dataset.

Figure 2 shows the number of results per method type, in which we have merged several of our initial 64 types into 'Other' categories to minimize the number of bars. Even if this is by no means a definitive description of this complex piece of data, it clearly demonstrates that the vast majority of results relates to methods associated with the Machine Learning area, and especially with its sub-family of classification trees (in total trees represent 1,851 results, more than a third of all results). Of course this may raise questions about the generality of our dataset—and the analyses which we have carried on, but as we tried not to apply any selective bias in our data collection we think that this bias actually comes from the literature itself. This reflects perhaps the fact that the Machine Learning area has traditionally been more open to empirical testing on benchmark datasets than other areas.

3 A strange observation

Figures 3 and 4 describe our rather puzzling observation. Figures 3a, b show error rate respectively plotted against (log) dataset size and (log) dataset dimensionality, with added density contours estimated via standard Gaussian kernels. Looking at the scatter of points in Fig. 3a, one may observe a near-triangular shape, with a clear lack of points in the upper right corner, where the large and difficult datasets should be. The pattern is also observable in Fig. 3b, although perhaps less marked. Both figures

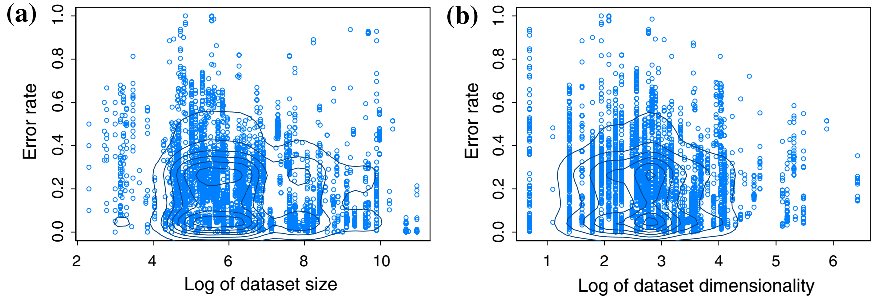

Fig. 3 Plots of error rate against (log) dataset size and dimensionality, with estimated density contours. **a** Error rate versus log of dataset size. **b** Error rate versus log of dataset dimensionality

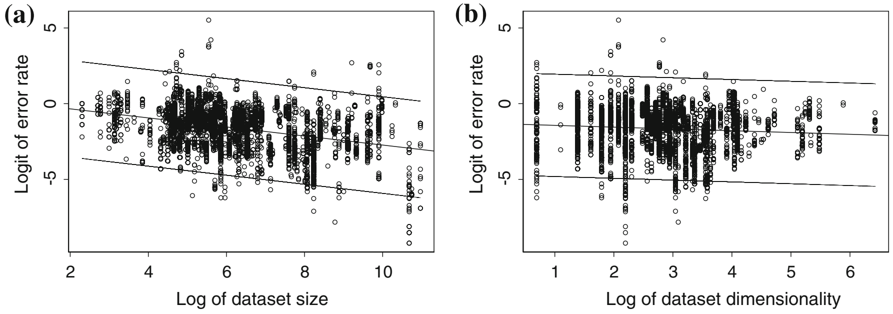

Fig. 4 Plots of logit error rate against (log) dataset size and dimensionality, with linear regression line and 95% predictive confidence intervals. **a** Logit of error rate versus log of dataset size (slope -0.30, $R^2 = 0.12$). **b** Logit of error rate versus log of dataset dimensionality (slope -0.12, $R^2 = 0.01$)

also show bimodality of the conditional distribution of error rate given the horizontal axis.

To investigate this curious pattern more closely, we looked at the logit error rate in function of the two characteristics (Fig. 4a, b). Both show a decrease of the error rate, although the overall trend of Fig. 4a seems clearer than that of Fig. 4b. Simple regression lines (and their R^2 statistic, shown below each plot) confirm that impression. Note the high variability about the fitted line that the predictive confidence intervals show, which reflects the messiness of the data. When considering these predictive intervals one also has to keep in mind that our data is a representative sample of the available literature, which may have little to do with how classification methods are really applied, as we mentioned in Sect. 2.

At this point, perhaps the reader will think that the trend of Fig. 4a is quite negligible. In fact, we have reasons to believe it is not, as we will see in Sect. 4.1. For the moment however, let us only say that if we consider the mixed character of the data, it is perhaps surprising to observe anything at all. For example, plots for number of classes (not shown) showed nothing but random variation, but of course with little variation on the x-axes any trend is more difficult to spot.

Simple explanations spring to mind when looking at Fig. 4. It is a known fact (see, for example, Perlich et al. 2003) that, for given underlying distributions, increasing training set size decreases the expected true error rate (that is, the expected error rate of a given classification method over the distribution of possible training sets). One could also say that given an underlying classification problem adding variables increases the separability of the classes, provided these variables are 'reasonably' chosen—and one can reasonably assume they are for non-pathological datasets. However, our plots and the associated regression slopes are across distributions, not within distributions, so it seems that those cannot be straightforward explanations. We come to this point in the following section.

4 Within versus between distributions

The fact that increasing training set size for samples drawn from given distributions leads to reduction in error rate is well known (Perlich et al. 2003). This, of course, relates not to the Bayes error rate, which remains constant if the underlying distributions are unchanged, but to the expected true error rate of a classification rule (that is, the expected error rate of a given classification method over the distribution of possible training sets). There are two causes for this reduction. On the one hand, for fixed rule complexity, increasing training set size will reduce the variance in the estimate of the classification rule: this will apply, for example, to a rule with a fixed number of parameters, such as logistic regression. On the other hand, certain types of rule automatically increase their complexity with increasing training set size. Tree methods, such as `C4.5`, are examples of such rules. Beyond a certain value, of course, the reduction in variance of the parameter estimates will be negligible, while the increase in model complexity (for those models which increase complexity in this way) will continue.

One should also expect that, for a given problem, increasing the number of variables will be associated with an improvement in accuracy, for any method. Adding variables to a dataset will tend to increase the separability between the classes, provided the added variables are properly chosen, and we might expect them to be so chosen since datasets are usually created using some prior knowledge of which variables are relevant. This is of course not taking into account the voluntary addition of irrelevant variables, such as in the famous artificial datasets `Led` (7 variables + sometimes 17 noisy variables) and `Waveform` (21 variables + sometimes 19 noisy variables).

However, our situation is slightly different, in that we are concerned not with how different sample sizes or varying dimensionality influence the error rate for a given underlying distribution, but rather with how different sample sizes and number of variables chosen from different distributions influence error rate. The above two explanations, which are valid *within distributions*, do not necessarily explain why we observe a similar phenomenon *across distributions*. In the rest of this article we will focus on sample size, given that the trend was more marked in our data (and for another reason which will appear in Sect. 5), but the same line of argument is valid for dimensionality.

In Fig. 5, we show a, very simplified, example of what could occur in our case. For each of the distributions along the lines 1, 2 and 3 a decrease in error rate takes place when dataset size increases (for ease of drawing we have supposed that these decreases are linear and that their slopes are identical, but the argument is still valid

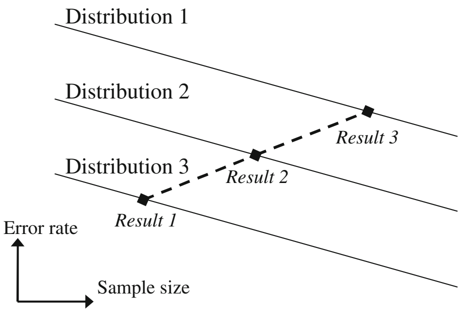

Fig. 5 Within versus between distributions

if these assumptions do not hold). However if we observe a sample consisting of results taken at different points from the different distributions (the points labelled as Results 1, 2 and 3), then *any* sort of trend could result. In the figure, an increasing trend results. This is an example of the *ecological fallacy* phenomenon, where one can see trends within groups of data without necessarily observing a similar trend overall. Of course the observed overall trend will depend on the separation between the distributions; for example, if the lines formed a tight band in Fig. 5 we would tend to observe a decreasing trend across the points labelled Results 1, 2, and 3. This explanation only deepens the mystery: why would the learning curves be so tightly bunched? In other words why are there so few results that lie in the upper right corner of Fig. 3a? We will come back to this in our conclusion, but first need to make some more investigations.

Things are complicated by the fact that some of the points displayed in Fig. 4a in fact do correspond to different sample sizes drawn from the same distributions. More than one sample size is reported for 63 of the 146 datasets amongst those we have analysed in this study, with 3,542 of the 5,203 results being from these 63 data sets. However, within any of these multiple-sized datasets there is usually little variation in size: authors of different studies use different training sizes, but they do not change the dataset dramatically. A typical example of this situation is the `Letter-recognition` UCI dataset which has sometimes been used with 15,000 training examples and sometimes with 16,000. Nevertheless it is possible that the phenomenon explored in Perlich et al. (2003) is still inducing some effect. Given that the regression slope reported in Fig. 4a is a priori not very large, an obvious question is whether it could be entirely attributed to these influences.

To test this, we repeated the analyses for the results which were reported only for single sized training sets and separately for those which were reported for multiple sized sets. Plots are shown in Fig. 6. The resulting slopes are almost identical, between themselves and with that in Fig. 4a. It thus seems that the phenomenon described in Perlich et al. (2003) does not provide a direct explanation for the trend of Fig. 4a.

4.1 Simulation study

The phenomenon in the real data illustrated in Fig. 4a arises from two sources—the within and between data set reduction in error rate with increasing data set size. Teasing

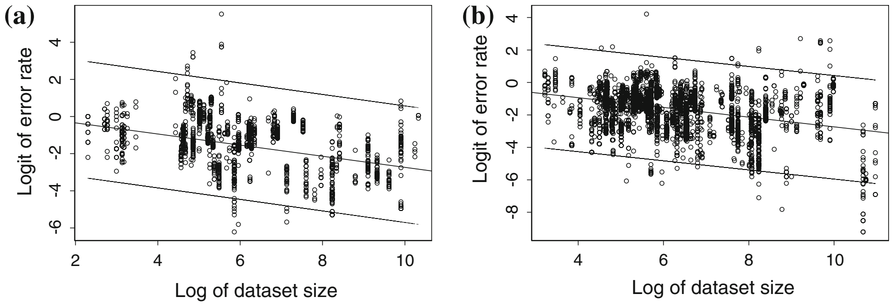

Fig. 6 (Logit error rate in function of log dataset size for datasets with multiple and unique sizes, with linear regression line and 95% confidence envelopes. **a** Datasets with unique size (slope −0.31, $R^2 = 0.15$). **b** Datasets with multiple sizes, (slope −0.28, $R^2 = 0.09$)

these sources apart to assess the magnitude of the between data set effect is difficult. An alternative approach is to attempt to gauge the size of the within data set reduction in error rate as sample size increases, and compare this with the combination effect manifest in Fig. 4a. To find such an estimate of the within data set reduction, we carried out a simulation. Of course, this simulation does not take any of the characteristics of the real datasets into account, but it does provide a baseline—and our results are consistent with the studies of Perlich et al. (2003).

This experiment was designed as follows:

- 2 normal classes in 10 dimensions, one located in the origin and the other at $(1, 1, \ldots, 1)$.
- 200 training sizes varying between 50 and 10,000, with a constant step on the logarithmic scale.
- Variable test sets, each of the same size as the corresponding training set.
- 100 runs (including sampling training and test sets) at each size.

We used identity covariance matrices for both classes, and the linear discriminant. This is thus the simplest setting one could think about, since the linear discriminant is the Bayes rule in this case and the classes are quite well separated (the Bayes error rate, given by $1 - \Phi(\sqrt{10}/2)$, is about 0.057). Results of the simulations are plotted in Fig. 7, and the observed trend is similar to those reported in Perlich et al. (2003) (the logit transform barely affects the shape of the curves, since in the region we are investigating this function is almost linear). The expectation of the logit error rate decreases with increasing training set size as a clear exponential curve, as the regression line of the type $y = a + b\exp(-cx)$ shown on the figure tends to demonstrate. The parameters (a,b,c) of these lines are shown inside the plot.

Since the learning curve has a different shape from those of Fig. 4a, the comparison cannot be straightforward. We will hence compare the b of the exponential regression, with an ‘equivalent’ value derived from the slope of the line of Fig. 4a: we take this equivalent value as the opposite of the slope times the length of the horizontal axis in Fig. 7 (i.e. $\log(10000) - \log(50)$). This represents the overall decrease in (logit) error rate over this length, and is thus comparable with b.

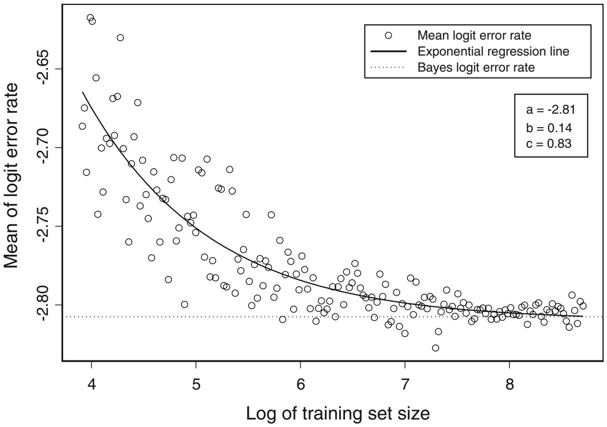

Fig. 7 Training curve (on the logit scale) of the linear discriminant for two homoscedastic, well-separated classes

The value we found for this decrease was 1.58, compared with a value of 0.14 for *b* (Fig. 7). We tried a few different settings for the simulations (but with roughly the same class separation), and the ratio between the overall decrease of Fig. 4a and *b* was between 5 and 10. Of course we acknowledge the limited character of this conclusion, due to the minimal nature of the present simulation study on one trivial dataset. In any case it is possible to design an example which will make the decrease of Fig. 4a appear small or large, but we did not design the study with any particular objective other than to put the observed phenomenon of Fig. 4a into context—and it does show that the magnitude of the trend is not negligible.

5 Other explorations

5.1 Correlation between size and dimensionality

One complication is that the (log) dataset size and the (log) dataset dimensionality are positively correlated (correlation coefficient: 0.30), as one may expect. This means that the observed decrease in error rate when one of the quantities increases could be due to changes in the other quantity. To take this into account, we fitted a linear model with both quantities as predictor variables. This regression model takes the following simple form:

$$logit(e_r) = \alpha + \beta \log(t) + \gamma \log(f) \tag{1}$$

where e_r is the error rate, t the size of the dataset and f its dimensionality. The values of the fitted coefficients are shown in Table 1.

Comparing the coefficients of this multiple regression with those of separate linear regressions (see Fig. 4), one may observe that the coefficient of dataset size has not changed much (−0.30 to −0.31), whereas that of dimensionality has, and quite

Table 1 Fitted values, standard errors, and p-values, for the coefficients of Eq. 1

Coefficient	Value	SE	p-value
α	1.07	0.11	0.03
β	−0.31	0.01	0
γ	0.04	0.02	0.05

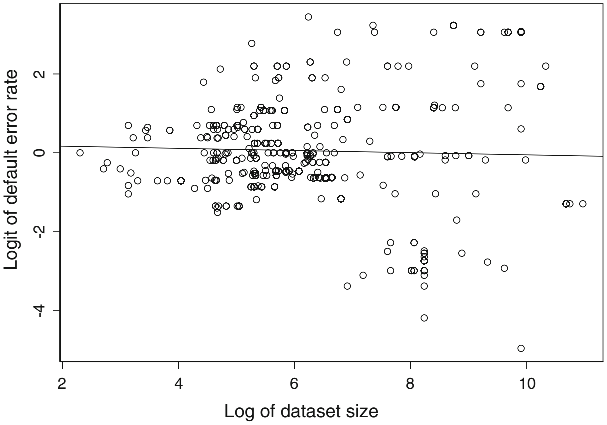

Fig. 8 Logit of default error rate against log of training size, with linear regression line (slope −0.03, standard error 0.01)

dramatically (−0.12 to +0.04). In any case, the relatively large standard error (though still quite small) and p-value for the latter coefficient seem to indicate that most of the trend observed in Fig. 3b was in fact due to the correlation between dataset size and dimensionality. The coefficient for sample size, β, is still markedly different from 0 in this model.

5.2 Default error rate

Another potential explanation for the trend of Fig. 4a would be that, for some curious reason, large datasets are ‘simpler’. It is difficult to test this hypothesis with the data we have at hand (besides, defining the ‘difficulty’ of a dataset is non-trivial), but as a very crude answer we can look at the default error rate, i.e. the error rate of a classification method which would assign every example to the most represented class in the training set. Figure 8 shows the logit of this default error rate against the log of dataset size. The plot displays a slight decreasing trend, but certainly nothing like that of Fig. 4a. Hence there is no reason to believe that large datasets are particularly easier, at least using this very crude measure of default error rate.

5.3 Error rate estimation

One possible factor which may have an influence on the observed trend is the method of error rate estimation. The two methods which are widely used (indeed the only ones

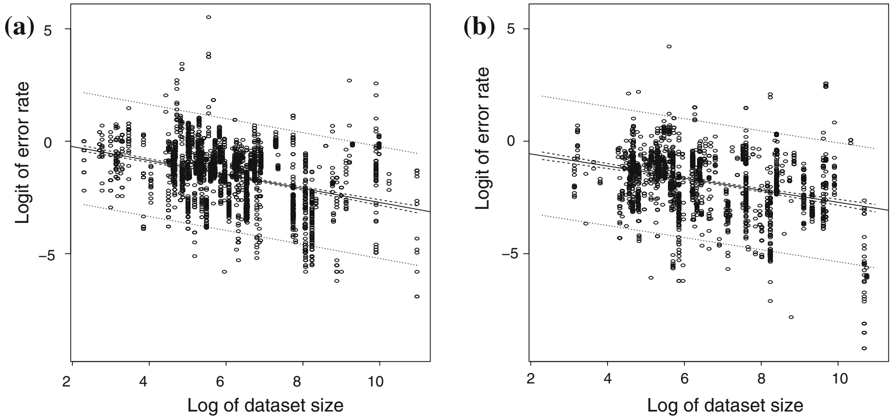

Fig. 9 Logit of error rate against (log) dataset size, according to error rate estimation procedure. **a** Cross-validation, 3,145 results (slope −0.31, $R^2 = 0.12$). **b** Single test set, 2,050 results (slope −0.27, $R^2 = 0.10$)

which we encountered) are cross-validation and single partition in training and test set. As the first one is being used mostly for small datasets and the second one mostly for large ones, one could wonder whether the trend we observed was partly due to this factor. We hence repeated the previous analysis after splitting our dataset in two subsets, one for each estimation method. As an aside, we considered 'leave-one-out' as a particular case of cross-validation and we removed the 8 observations where the error rate estimation procedure was unspecified. Results are shown on Fig. 9. Both slopes are still negative and the same magnitude, and hence we may conclude that the estimation procedure appears to have little to no influence on the observed trend.

5.4 Large datasets

Another objection to our finding would perhaps be that some datasets, especially the very large ones, are distorting the fit. To investigate this issue we removed the points with the 5% largest sizes in our data (i.e. sizes of 11,220 or more), and repeated the analysis. This represents the removal of 291 results related to 11 datasets (*Chrom, CredMan, Cut50, Digit, Krk, Letter, Netpho, Netstr, Nettyp, Protein, SatIm, Shuttle*). Again, Fig. 10 shows that this operation had almost no influence on the trend.

6 Conclusion

Despite our best efforts, the trend of Fig. 4a is still puzzling. The simulation study we carried out demonstrated that it is not negligible, compared with the negative slope of a classic learning curve. The results of a multiple regression, accounting for the correlation between datasets size and dimensionality, did not reduce the coefficient of dataset size. And finally, the default error rate does not decrease in the same way as the actual error rate.

We are thus left with the impression that there is simply a sizeable dearth of large and difficult datasets and, for lack of other convincing explanations, we suspect that

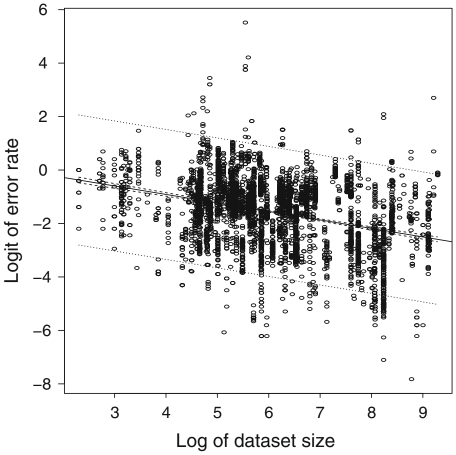

Fig. 10 Logit of default error rate against log of training size, with the 5% largest sizes removed (slope −0.32, $R^2 = 0.11$)

this has more to do with how datasets are collected and/or designed. The simplest explanation of this is perhaps what may be called the 'pessimism' of dataset building. Whether datasets are completely manufactured or collected from real physical processes, the scientists who build them would usually start incrementally, by looking at few examples, and if the results do not seem promising they would include more variables, or change the problem completely, or even give up. This perfectly sensible behaviour results logically in few large and difficult datasets being created, and perhaps even fewer stored in a repository for technical reasons. To put it in the language of publication bias, classification tasks where one has to collect thousands of examples to get a reasonable error rate are likely to stay in the 'file drawer' or never actually appear.

Perhaps the most important question is: what does this mean in practice? If nothing else, it seems to bring some concrete evidence to the apparent consensus that datasets taken from the UCI repository (like most of those of our study) are 'too easy'. Holte (1993) was already arguing along those lines more than a decade ago, but hopefully this will change with the inclusion of larger, harder datasets. The community has responded to this by a sort of 'healthy doubt' about the results of comparative studies including UCI datasets (Salzberg 1997; Soares 2002), but they still do appear very frequently in published work.

Because learning curves of different methods often cross, Perlich et al. (2003) concluded "*These results* (…) *call into question the practice of experimenting with smaller data sets* (*for efficiency reasons*) *to choose the best learning algorithm, and then 'scaling up' the learning with the chosen algorithm*". Our own observation hence additionally suggests that, if one wishes to bypass this 'small-dataset' curse, and hence consciously chooses large datasets from the set of commonly available datasets, then one *will* tend to get good results, whatever method is used.

Acknowledgments The work was sponsored by the MOD Corporate Research Programme, CISP. We would like to express our appreciation to Andrew Webb and Wojtek Krzanowski for their helpful comments and advice on this work.

References

Atlas L, Connor J, Dong P, Lippman A, Cole R, Muthusamy Y (1991) A performance comparison of trained multi-player perceptrons and trained classification trees. In: Systems, man and cybernetics: proceedings of the 1989 IEEE international conference, Cambridge, Hyatt Regency, pp 915–920

Blake CL, Merz CJ (1998) UCI repository of machine learning databases. http://www1.ics.uci.edu/~mlearn/MLRepository.html, University of California, Irvine, Dept. of Information and Computer Sciences

Brazdil PB, Soares C, Pinto da Costa J (2003) Ranking learning algorithms: using IBL and meta-learning on accuracy and time results. Mach Learn 50:251–277

Eklund PW, Hoang A (2002) A performance survey of public domain supervised machine learning algorithms. http://citeseer.nj.nec.com/551273.html

Hand DJ (1999) Intelligent data analysis: an introduction. In: Berthold M, Hand DJ (eds) Intelligent data analysis. Springer, Berlin

Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11:63–91

Jamain A (2004) Meta-analysis of classification methods. PhD thesis, Department of Mathematics, Imperial College, London (2004)

Jamain A, Hand DJ (2005) The Naive Bayes mystery: a classification detective story. Pattern Recognit Lett 26:1752–1760

Jamain A., Hand DJ (2008) Mining supervised classification performance studies: a meta-analytic investigation. J Classif 25(1):87–112

Lim T, Loh W, Shih Y (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Mach Learn 40:203–228

METAL Consortium . Esprit project METAL (#26.357). http://www.metal-kdd.org, 2002

Michie D, Spiegelhalter DJ, Taylor CC (1994) Machine learning, neural and statistical classification. Ellis Horwood, New York

Perlich C, Provost F, Simonoff JS (2003) Tree induction versus logistic regresion: a learning-curve analysis. J Mach Learn Res 4:211–255

Quinlan JR (1994) Comparing connectionist and symbolic learning methods, volume I: constraints and Prospects. MIT Press, Cambridge, pp 445–456. http://citeseer.nj.nec.com/quinlan94comparing.html

Rasmussen CE, Neal RM, Hinton GE, van Camp D, Revow M, Ghahramani Z, Kustra R, Tibshirani R (1996) DELVE, Data for evaluating learning in valid experiments. http://www.cs.toronto.edu/~delve/

Salzberg SL (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Discov 1:317–328

Sargent DJ (2001) Comparison of artificial neural networks with other statistical approaches. Cancer 91:1636–1642

Shavlik JW, Mooney RJ, Towell GG (1991) Symbolic and neural learning algorithms: an experimental comparison. Mach Learn 6:111–143

Soares C (2002) Is the UCI repository useful for data mining? In: Lavrac N, Motoda H, Fawcett T (eds) Proceedings of the ICML-2002 workshop on data mining lessons learned

Sohn SY (1999) Meta-analysis of classification algorithms for pattern recognition. IEEE Trans Pattern Recognit Mach Intell 21(11):1137–1144

Viswanathan M, Webb GI (1998) Classification learning using all rules. In: 11th European conference on machine learning. Springer, Berlin, pp 150–159

Zarndt F (1995) A comprehensive case study: an examination of machine learning and connectionnist algorithms. http://citeseer.nj.nec.com/481595.html