15
Comparing Classifiers

15.1 Introduction

In Chapter 12 we considered how to choose between different classifiers applied to the same dataset. For those with real datasets to analyse this is obviously the principal issue.

However there is an entirely different category of data miner: those who develop new algorithms or what they hope are improvements to existing algorithms designed to give superior performance on not just one dataset but a wide range of possible datasets most of which are not known or do not even exist at the time the new methods are developed. Into this category fall both academic researchers and commercial software developers.

Whatever new methods are developed in the future, we can be certain of this: no one is going to develop a new algorithm that out-performs all established methods of classification (such as those described in this book) for all possible datasets. Data mining packages intended for use in a wide variety of possible application areas will continue to need to include a choice of classification algorithms to use. The aim of further development is to establish new techniques that are generally preferable to well-established ones. To do this it is necessary to compare their performance against at least one established algorithm on a range of datasets.

There are many published papers giving descriptions of interesting new classification algorithms accompanied by a performance table such as Figure 15.1. Each column gives the predictive accuracy, expressed as a percentage, of one of

M. Bramer, *Principles of Data Mining*, Undergraduate Topics
in Computer Science, DOI 10.1007/978-1-4471-7307-6_15

Dataset	Established Classifier A	New Classifier B
dataset 1	80	85
dataset 2	73	70
dataset 3	85	85
dataset 4	68	74
dataset 5	82	71
dataset 6	75	65
dataset 7	73	77
dataset 8	64	73
dataset 9	75	75
dataset 10	69	76
Total	744	751
Average	74.4	75.1

Figure 15.1 Performance of classifiers A and B on 10 datasets

the classifiers on a range of datasets. (Note that for the method of comparison we describe below multiplying all the values in both columns by a constant has no effect on the outcome. Thus it makes no difference whether we represent predictive accuracy by percentages as here or by proportions between 0 and 1, such as 0.8 and 0.85.)

The production of tables of comparative values such as Figure 15.1 is a considerable improvement over the position with some of the older Data Mining literature where new algorithms are either not evaluated at all (leaving the brilliance of the author's ideas to speak for itself, one assumes) or are evaluated on datasets that are only available to the author and/or are not named. As time has gone by collections of 'standard' datasets have been assembled that make it possible for developers to compare their results with those obtained by other methods on the same datasets. In many cases the latter results are only available in the published literature, since with a few honourable exceptions authors do not generally make software implementing their algorithms accessible to other developers and researchers, except in the case of commercial packages.

A very widely-used collection of datasets is the 'UCI Repository' [1] which was introduced in Section 2.6. Being able to compare performance on the same datasets as those used by previous authors clearly makes it far easier to evaluate new algorithms. However the widespread use of such repositories is not an unmixed blessing as will be explained later.

Figure 15.1 shows the predictive accuracy of algorithms A and B on 10 datasets. We can see that in three cases A out-performed B, in two cases the performance was equal and in five cases B out-performed A. The average accuracy of A was 74.4% and the average accuracy of B was 75.1%. What can we conclude from all this?

15.2 The Paired t-Test

A commonly used method of comparing classification algorithms is the *paired t-test*. We will start by illustrating the method and then discuss a number of issues relating to it.

First we add to Figure 15.1 a column of the differences between the A and B values, i.e. B-A, which is traditionally denoted by the letter z. We also construct a column showing the square of the differences, i.e. z^2.

Dataset	Established Classifier A	New Classifier B	Difference z	Square of Difference z^2
dataset 1	80	85	5	25
dataset 2	73	70	−3	9
dataset 3	85	85	0	0
dataset 4	68	74	6	36
dataset 5	82	71	−11	121
dataset 6	75	65	−10	100
dataset 7	73	77	4	16
dataset 8	64	73	9	81
dataset 9	75	75	0	0
dataset 10	69	76	7	49
Total	744	751	7	437
Average	74.4	75.1	0.7	43.7

Figure 15.2 Performance of classifiers A and B on 10 datasets (with z and z^2 values)

We can see that the average difference between A and B is 0.7, i.e. 0.7% in favour of classifier B. This does not seem very much. Is it sufficient to reject the *null hypothesis* that the performance of classifiers A and B is effectively the same? We will address this question using a paired t-test. The word 'paired' in the name refers to the fact that the results fall into natural pairs, i.e. it is

sensible to compare the results for dataset 1 for classifiers A and B but these are separate from those for dataset 2 etc.

To perform a paired t-test we need only three values: the total of the values of z, the total of the z^2 values and the number of datasets. We well denote these by $\sum z$, $\sum z^2$, and n, respectively, so $\sum z = 7$, $\sum z^2 = 437$ and $n = 10$.[1]

From these three values we can calculate the value of a statistic which is traditionally represented by the variable t. The t-statistic was introduced in the early 20^{th} century by an English statistician named William Gosset, who is best known by his pen name of 'Student', and so this test is also often known as *Student's t-test*.

The calculation of the value of t can be broken down into the following steps.

Step 1. Calculate the average value of z: $\sum z/n = 7/10 = 0.7$.

Step 2. Calculate the value of $(\sum z)^{2/}n$. Here this gives $7^2/10 = 4.9$.

Step 3. Subtract the result of step 2 from $\sum z^2$. Here this gives $437 - 4.9 = 432.1$.

Step 4. Divide this value by $(n - 1)$ to give the *sample variance*, which is traditionally denoted by s^2. Here s^2 is $432.1/9 = 48.01$.

Step 5. Take the square root of s^2 to give s, known as the *sample standard deviation*. Here the value of s is $\sqrt{48.01} = 6.93$.

Step 6. Divide s by $\sqrt{n}$ to give the *standard error*. Here the value is $6.93/\sqrt{10} = 2.19$.

Step 7. Finally we divide the average value of z by the standard error to give the value of the t statistic. Here $t = 0.7/2.19 = 0.32$.

The word 'sample' in both 'sample variance' and 'sample standard deviation' refers to the fact that the 10 datasets given in the table are not all the possible datasets that exist to which the two classifiers may be applied. They are just a very small sample of all the possible datasets that exist or may exist in the future. We are using them as 'representatives' of this much larger collection of datasets. We will return to the question of how far this is reasonable.

The terms standard deviation and variance are commonly used in statistics. Standard deviation measures the fluctuation of the values of z about the mean

[1] For those not familiar with this notation, which uses the Greek letter $\sum$ (pronounced 'sigma') to denote summation, it is explained in Appendix A.1.1. The simplified variant used here leaves out the subscripts, as the values to be added are obvious. $\sum z$ (read as 'sigma z') denotes the sum of all values of z, which here is 7, $\sum z^2$ (read as 'sigma z squared') represents the sum of all the values of z^2, which is 437. The latter is not to be confused with $(\sum z)^2$, which is the square of $\sum z$, i.e. 49.

value, which here is 0.7. In Figure 15.2 the fluctuation is considerable: the differences between the values of z and the average value (0.7) vary from -11.7 to $+8.3$ and this is reflected in a sample standard deviation, s, value of 6.93, almost 10 times larger than the average itself. The calculation of the standard error value adjusts s to allow for the number of datasets in the sample. Because t is the average value of z divided by the standard error, it follows that the smaller the value of s (i.e. the fluctuation of z values about the average), the larger will be the value of t. (Readers interested in a full explanation of and justification for the t-test are referred to the many statistics textbooks that are available.)

Now we have calculated t, the next step is to use it to determine whether or not to accept the null hypothesis that the performance of classifiers A and B is effectively the same. We ask this question in an equivalent form: is the value of t sufficiently far away from zero to justify rejecting the null hypothesis? We say 'sufficiently far away from zero' rather than 'sufficiently large' because t can have either a positive or a negative value. (The average value of z can be positive or negative; standard error is always positive.)

We can now reformulate our question as: 'how likely is a value of t outside the range from -0.32 to $+0.32$ to occur by chance'? The answer to this depends on the number of datasets n, but statisticians refer instead to the number of *degrees of freedom*, which for our purposes is always one less than the number of datasets, i.e. $n - 1$.

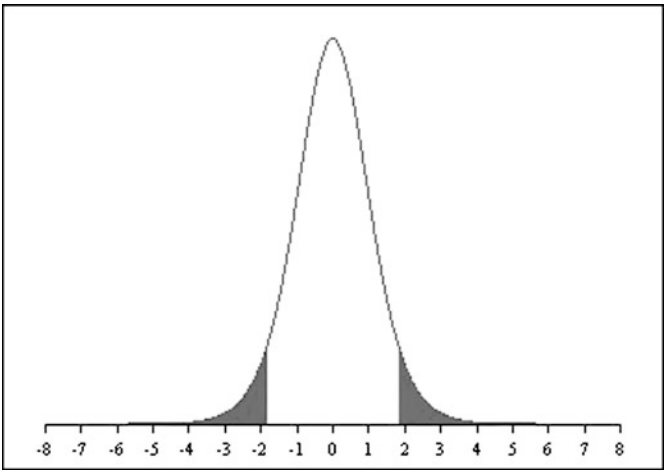

Figure 15.3 t-distribution for 9 degrees of freedom

Figure 15.3 shows the distribution of the t-statistic for 9 degrees of freedom (chosen because there are 10 datasets in the tables shown so far).

The left- and right-hand ends of the curve (called its 'tails') go on infinitely in both directions. The area between the entire curve and the horizontal axis, i.e. the t-axis, gives the probability that t will take one of its possible values, which of course is one.

The figure has the values $t = -1.83$ and $t = +1.83$ marked with vertical lines. The area between the parts of the curve that are to the left of $t = -1.83$ or to the right of $t = +1.83$ and the horizontal axis is the probability of the t value being ≤ -1.83 or $\geq +1.83$, i.e. at least as far away from zero as 1.83. We need to look at both tails in this way as a negative value of -1.83 is just as much evidence that the null hypothesis (that the two classifiers are equivalent) is false as the positive value $+1.83$. When we compare two classifiers there is no reason to believe that if A and B are significantly different then B must be better than A; it might also be that B is worse than A.

The area shaded in Figure 15.3, i.e. the probability that t is at least 1.83 either side of zero can be calculated to be 0.1005.

Looking at the probability that $t \leq -1.83$ or $t \geq +1.83$, or in general that $t \leq -a$ or $t \geq +a$, for any positive value a, gives us what is known as a *two-tailed test of significance*.

The value of the area under the two tails $t \leq -a$ and $t \geq +a$ have been calculated for different degrees of freedom and values of a corresponding to probabilities of particular interest. Some of these are summarised in Figure 15.4. Figure 15.4 shows some key values for the t statistic for degrees of freedom from 1 to 19, i.e. for comparisons based on anything from 2 to 20 datasets. (Note that because we are using a two-tailed test, probabilities 0.10, 0.05 and 0.01 in the table correspond to $a = 0.05$, 0.025 and 0.005 respectively in the previous discussion.)

Looking at the values for 9 degrees of freedom (i.e. for $n = 10$) the value of 1.833 in the 'Probability 0.10' column indicates that a value of $t \geq 1.833$ (or ≤ -1.833) would only be expected to happen by chance with probability 0.10 or less, i.e. no more than 1 time out of 10. If we had a t value of 2.1, say, we could reject the null hypothesis 'at the 10% level', implying that such an extreme value of t would only be expected to occur by chance fewer than one time in 10. This is a commonly used criterion for rejecting a null hypothesis and on this basis we could confidently say that classifier B is significantly better than classifier A.

A value of $t \geq 2.262$ (or ≤ -2.262) would enable us to reject the null hypothesis at the 5% level, and a value of $t \geq 3.250$ (or ≤ -3.250) would enable us to reject the null hypothesis at the 1% level, as such values would only be expected to occur by chance one time in 20 and 1 time in 100 respectively.

Degrees of Freedom	Probability 0.10	Probability 0.05	Probability 0.01
1	6.314	12.71	63.66
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861

Figure 15.4 t values for 1 to 19 degrees of freedom (two-tailed test)

Naturally we could use other threshold values and work out a value of t that would only be exceeded by chance one time in six on average, say, but conventionally we use one of the thresholds shown in Figure 15.4. The least restrictive condition generally imposed is that to reject a null hypothesis we require a value of t that would occur no more than 1 time in 10 by chance.

Returning to our example, the value of t calculated was only 0.32, which with 9 degrees of freedom is nowhere near the 10% value of 1.833. We can safely accept the null hypothesis. On the basis of the evidence presented it would be unsafe to say that the performance of classifier B was significantly different from that of classifier A.

It is important to appreciate that the reason for this disappointing result (certainly disappointing to the creator of classifier B) is not the relatively low average value of z (0.7). It is the relatively high value of the standard error (2.19) relative to the average value of z.

To illustrate this we will introduce a new classifier C, which will turn out to be much more successful as a challenger to classifier A.

Dataset	Established Classifier A	New Classifier C	Difference z	Square of Difference z^2
dataset 1	80	81	1	1
dataset 2	73	74	1	1
dataset 3	85	86	1	1
dataset 4	68	69	1	1
dataset 5	82	83	1	1
dataset 6	75	75	0	0
dataset 7	73	75	2	4
dataset 8	64	63	−1	1
dataset 9	75	75	0	0
dataset 10	69	70	1	1
Total	744	751	7	11
Average	74.4	75.1	0.7	1.1

Figure 15.5 Performance of classifiers A and C on 10 datasets

Figure 15.5 shows the percentage accuracy of each classifier on the 10 datasets. Once again the average value of z is 0.7 but this time there is far less spread of z values around the average. The differences between the values of z and the average value (0.7) vary from −1.7 to +1.3.

This time the significant values are $\sum z = 7$, $\sum z^2 = 11$ and $n = 10$. Only the second of these has changed but the effect is considerable. The seven step calculation of t now goes as follows.

Step 1. Calculate the average value of z: $\sum z/n = 7/10 = 0.7$ [as before].

Step 2. Calculate the value of $(\sum z)^2/n$. Here this gives $7^2/10 = 4.9$ [as before].

Step 3. Subtract the result of step 2 from $\sum z^2$. Here this gives $11 - 4.9 = 6.1$.

Step 4. Divide this value by $(n - 1)$ to give the sample variance s^2. Here s^2 is $6.1/9 = 0.68$.

Step 5. Take the square root of s^2 to give the sample standard deviation. Here the value of s is $\sqrt{0.68} = 0.82$.

Step 6. Divide s by $\sqrt{n}$ to give the *standard error*. Here the value is $0.82/\sqrt{10} = 0.26$, which is considerably less than the standard error calculated from Figure 15.2 (i.e. 2.19).

Step 7. Finally we divide the average value of z by the standard error to give the value of the t statistic. Here $t = 0.7/0.26 = 2.69$.

This value of t is greater than the 5% value for 9 degrees of freedom in Figure 15.4. We can say that classifier C is significantly better than classifier A at the 5% level.

The decisive difference between this example and the earlier one using Figure 15.2 was not the average value of z (they were the same) but the much smaller standard error.

15.3 Choosing Datasets for Comparative Evaluation

We will now return to the original problem of whether classifier B is better than (or perhaps worse than) classifier A.

Suppose now that for whatever reason datasets 5 and 6, both of which give results very favourable to classifier A, had been omitted from the sample investigated. We would then have a revised version of Figure 15.2, with only 8 datasets, as shown in Figure 15.6.

Dataset	Established Classifier A	New Classifier B	Difference z	Square of Difference z^2
dataset 1	80	85	5	25
dataset 2	73	70	−3	9
dataset 3	85	85	0	0
dataset 4	68	74	6	36
dataset 7	73	77	4	16
dataset 8	64	73	9	81
dataset 9	75	75	0	0
dataset 10	69	76	7	49
Total	587	615	28	216
Average	73.375	76.875	3.5	27

Figure 15.6 Performance of classifiers A and B with datasets 5 and 6 removed.

Now $\sum z = 28$, $\sum z^2 = 216$ and $n = 8$.

The average value of z is 3.5. The standard error is 1.45 and the value of t is 2.41. This is large enough for classifier B to be declared significantly better than classifier A at the 5% level. (With 7 degrees of freedom the threshold value

for probability 0.05 is 2.365.) The developer of classifier B is clearly fortunate that datasets 5 and 6 were left out of the analysis.

Suppose now that datasets 5 and 6 were omitted, but two further datasets, 11 and 12, both of which are favourable to classifier B, were included in the analysis, giving the results shown in Figure 15.7.

Dataset	Established Classifier A	New Classifier B	Difference z	Square of Difference z^2
dataset 1	80	85	5	25
dataset 2	73	70	−3	9
dataset 3	85	85	0	0
dataset 4	68	74	6	36
dataset 7	73	77	4	16
dataset 8	64	73	9	81
dataset 9	75	75	0	0
dataset 10	69	76	7	49
dataset 11	75	80	5	25
dataset 12	82	88	6	36
Total	704	783	39	277
Average	70.4	78.3	3.9	27.7

Figure 15.7 Performance of classifiers A and B with datasets 11 and 12 replacing 5 and 6.

Now $\sum z = 39$, $\sum z^2 = 277$ and $n = 10$.

The average value of z is 3.9. The standard error is 1.18 and the value of t is 3.31. This is large enough to be significant at the 1% level.

Paradoxically if the results for classifier B with datasets 11 and 12 had been much better, say 95% and 99% respectively, the value of t would have been lower at 2.81. Intuitively, we may say that by increasing the fluctuation around the average value of z we make it more likely that the difference between the classifiers has occurred by chance. To obtain a significant value of t, it is generally far more important that the values of z have low variability than that the average value of z is large.

It is clear that the choice of datasets to include in a performance table such as Figure 15.1 is of critical importance. A comparison of the t values calculated from Figures 15.2, 15.6 and 15.7 shows that leaving out (or including) datasets on which the new algorithm B performs badly (or well) can make the difference between a 'no significant difference' result and a significant improvement (or

vice versa). Paradoxically, omitting particularly favourable results, by lowering the standard error, can also increase the t value.

Is it too indelicate to raise here the issue of cheating? It would be very easy to leave out a few unfavourable results to make the t-value come out as significant. Naturally no reader of this book would ever be tempted to leave out poor results just to gain public recognition, a higher degree, a pay bonus or promotion, but it is possible that others are not always so scrupulous. Although this is always a possibility, a much bigger problem may be that of 'cheating oneself'. Having obtained good results for a new method, how much incentive is there to hunt for other datasets for which the results may be far worse?

15.3.1 Confidence Intervals

Having established that for the results given in Figure 15.6 classifier B is statistically significantly better than classifier A at the 5% level, and the average improvement for the eight datasets listed is 3.5%, it would be helpful to establish a *confidence interval* for the average improvement to indicate within what limits the true improvement for datasets not included in the table is likely to lie.

For this example the average value of z is 3.5 and the standard error is 1.45. As the t value in the 'Probability 0.05' column of Figure 15.4 for 7 degrees of freedom is 2.365, we can say that the 95% confidence interval for the true average difference is $3.5 \pm (2.365 * 1.45) = 3.5 \pm 3.429$. We can be 95% certain that the true average improvement lies between 0.071% and 6.929%.

For the performance figures given in Figure 15.7 classifier B is significantly better than classifier A at the 1% level. Here the average value of z is 3.9 and the standard error is 1.18. There are 9 degrees of freedom and the value of t in the 'Probability 0.01' column for that number of degrees of freedom is 3.250. We can say that the 99% confidence interval for the true average difference is $3.9 \pm (3.250 * 1.18) = 3.9 \pm 3.835$. We can be 99% certain that the true average improvement lies between 0.065% and 7.735%.

15.4 Sampling

So far we have shown how to test for the significance of a difference in performance between two classifiers on some specified datasets. However in most cases we do this not because we are particularly interested in those datasets but

because we would like our new method to be considered better on all possible datasets. This brings us to the issue of *sampling*.

Any collection of datasets can be considered to be a *sample* from the complete collection of all the world's datasets (which is not accessible to us of course), but is it a *representative sample*, i.e. one that accurately reflects the members of the entire population? If not, why should anyone imagine that a classifier's improved performance on datasets 1–10, say, should generalise to imply improved performance on all other (or indeed any other) datasets?

The situation is similar to the world of advertising, where it is common to see claims such as '8 out of 10 women prefer product B to product A'. (The laws of libel prevent us using more realistic examples in this section.)

Does this claim mean that the advertiser has asked exactly 10 women, perhaps all close friends, family members or employees? That would not be very convincing. Why should those 10 speak for all the women of the world? Even if we restrict ourselves to the aim of speaking for, say, all the women in Great Britain, it is obvious that just asking 10 people is hopelessly inadequate.

Some advertisements go further and say (e.g.) 'total number of women asked = 94'. This is better, but how were the 94 selected? If they were all questioned on the same Tuesday morning at the same shopping centre, or sports event say, the bias towards selecting people living in a small geographical area with particular interests and availability for answering surveys on Tuesday mornings is surely obvious.

To make any meaningful statement about the views of the female population of Great Britain we need to sub-divide the population into a number of mutually exclusive and homogeneous sub-groups, based on features such as geographical location, age group and socio-economic status and then ensure we interview a reasonably large group of women that is broken down in the same proportions for each sub-group as the overall population. This is known as *stratified sampling* and is the approach typically adopted by companies conducting opinion surveys.

Returning to data mining, a natural question to ask when faced with a table showing the comparative performance of different classifiers on a number of datasets is how were those datasets selected? It would be good to believe that they were a carefully selected representative sample of all the world's datasets, but that is hardly realistic. Let us suppose that all the datasets were chosen from a standard repository, such as the UCI one, which was established to facilitate comparison with the work of previous software developers. Is there any reason to suppose that they are a representative sample (rather than just a sample) of all the datasets in the UCI Repository?

It would be possible to attempt to achieve this, although unavoidably imprecisely, e.g. by choosing a number of datasets that are believed to include a

substantial proportion of noise, a number believed to be noise free, some with all attributes categorical, some with all attributes continuous, and so on.

In practice, most authors make no attempt to claim that their datasets are a representative sample of the UCI Repository. In many cases those chosen were almost certainly just those that were readily available to the developers. This is known as using an *opportunity sample* and is a reasonable way of proceeding in some circumstances, but such a sample is most unlikely to be representative.

When the aim is to make a comparison with results published, perhaps years earlier, by the celebrated Data Mining expert Professor X, there is really little choice but to use the same datasets as were used by X in his or her celebrated work. Developers of new methods can hardly be blamed for doing this, but again it begs the question: how did X select those datasets?

Even assuming that we could find a way of selecting a representative sample of the datasets in the UCI Repository would that guarantee that we had a representative sample of all the world's datasets? Unfortunately not. There is no reason to believe that datasets are entered into the Repository in a random fashion. We might hypothesise that in many cases they are datasets on which well-established methods give good predictive accuracy, placed in a Repository as a challenge for future workers to get even better results. Those who work on 'difficult' datasets and fail to make progress may be assumed to be much less likely to place the datasets in a Repository as a reminder of their failure.

Unfortunately the problems relating to the widespread use of the UCI Repository go far beyond this. They were discussed in a paper by Salzberg [2] as far back as 1997, which refers to a 'community experiments' effect. He says: 'many people are sharing a small repository of datasets and repeatedly using those same datasets for experiments. Thus there is a substantial danger that published results, even when using strict significance criteria and the appropriate significance tests, will be mere accidents of chance.... Suppose that 100 different people are studying the effects of algorithms A and B, trying to determine which one is better. Suppose that in fact both have the same mean accuracy (on some very large population of datasets), although the algorithms vary randomly in their performance on specific datasets. Now, if 100 people are studying the effect of algorithms A and B, we would *expect* that five of them will get results that are statistically significant at the [0.05] level, and one will get significance at the 0.01 level! ... Clearly in this case these results are due to chance, but if the 100 people are working separately, the ones who get significant results will publish, while the others will simply move on to other experiments'.

The problem of the community experiments effect can only have become more severe since. In the short term, it can be countered by creating new repositories, used by fewer people. However, in the long run the large number

of people experimenting with classification algorithms and the desirability of producing results that can be compared with those obtained by others in the future mean that the community experiments effect will inevitably affect these new repositories too.

It is perhaps becoming clear why evaluation is the Achilles Heel of much of the published literature about new classification algorithms. At the very least those publishing comparison tables such as Figure 15.1 should explain how the datasets listed were selected – but remarkably few seem to do so.

Faced with these problems, all that can be asked is that developers do the best they can. Publishing results for more datasets is obviously desirable, not only for those trying to judge their work but as benchmarks for future work. Most importantly, developers should always explain how and why they chose the datasets they analysed – and of course that choice should always be made *before* running any new algorithm on them.

15.5 How Bad Is a 'No Significant Difference' Result?

Whilst it is certainly desirable to have a range of classification algorithms available, as no one algorithm can ever be guaranteed to give the best possible performance on all datasets, the comments about 'community experiments' quoted above reflect a situation where many experiments with new classifiers have been and continue to be carried out, most of them giving a very similar performance across a range of familiar datasets.

The world does not need an endless supply of classification algorithms that are not significantly different from well-established ones or give only slightly better performance on a small number of datasets. Nevertheless there are reasons why developing a new classification algorithm may be desirable even though its performance measured by predictive accuracy is not significantly different from that of well-known 'standard' classifiers.

Predictive accuracy is not the only way to judge the quality of a classifier. A new classifier B may be better than an existing classifier A for other reasons, for example:

- B may be better founded in theory than A
- B may be computationally more efficient than A
- B may produce a model that is more human-understandable than A does

– B may give better performance for certain types of dataset than A, for example where there are many missing values or where there is likely to be a high proportion of noise present.

Given a performance table such as Figure 15.1 the question that needs to be addressed is what distinguishes those datasets for which the B value is greater than the A value from those the other way round. Often there may be no discernible reason for the differences but, where there is, a valuable new algorithm for particular types of dataset may have been found.

15.6 Chapter Summary

This chapter considers how to compare the performance of alternative classifiers across a range of datasets. The commonly used paired t-test is described and illustrated with worked examples, leading to the use of confidence intervals when the predictive accuracies of two classifiers are found to be significantly different.

Pitfalls involved in comparing classifiers are discussed, leading to alternative ways of comparing their performance that do not rely on comparisons of predictive accuracy.

15.7 Self-assessment Exercises for Chapter 15

Given the following table showing the percentage accuracy of two classifiers A and B on 20 datasets

1. Calculate the average value of the difference $B - A$.
2. Calculate the value of the standard error and the t-statistic.
3. Determine whether classifier B is significantly better or worse than classifier A at the 5% level.
4. If the answer to question 3 is yes, calculate the 95% confidence interval for the true difference in percentage accuracy between classifiers A and B.

Dataset	Classifier *A*	Classifier *B*
1	74	86
2	69	75
3	80	86
4	67	69
5	84	83
6	87	95
7	69	65
8	74	81
9	78	74
10	72	80
11	75	73
12	72	82
13	70	68
14	75	78
15	80	78
16	84	85
17	79	79
18	79	78
19	63	76
20	75	71

References

[1] Blake, C. L., & Merz, C. J. (1998). *UCI repository of machine learning databases*. Irvine: University of California, Department of Information and Computer Science. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] Salzberg, S. L. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, *1*, 317–327. Kluwer.