Comparison of Different Methods for Testing the Significance of Classification Efficiency

Edgard Nyssen

Department of Electronics, Brussels University (VUB),
Pleinlaan 2, B-1050 Brussels, BELGIUM
E-mail: ehnyssen@etro.vub.ac.be

Abstract. This paper discusses three methods to test the significance of classification efficiency (i.e. the fraction of correctly classified patterns of a test set), which can be applied to multiple class problems: the exact probability test, the Monte-Carlo test and the χ^2 test. First, a short theoretical description of the three methods is given. The methods have been applied to different classification problems. A comparison is made in terms of the following criteria: required assumptions, power and time behaviour. To conclude, the paper describes a set of criteria for the selection of the appropriate classification efficiency testing method.

1 Introduction

Consider a classification experiment involving k pattern classes and applied to a test set (referred to, from now on, by superscript T), independent of the set which has been used to design the classifier under study. The result of this experiment is a frequency distribution $N^{\mathrm{T}} = (n_{11}^{\mathrm{T}}, n_{12}^{\mathrm{T}}, \ldots, , n_{kk}^{\mathrm{T}})$ of the patterns, where n_{ij}^{T} represents the number of elements, belonging to class Ω_i, which have been assigned by the classifier to class Ω_j $(i, j \in \{1, \ldots, k\})$. The classification efficiency e for a set of patterns is defined as the correctly classified fraction:

$$e = \mathrm{e}(N) = \frac{\sum_{i=1}^{k} n_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij}} (\times 100\%).$$

In the case of the test set members, we have $e^{\mathrm{T}} = \mathrm{e}(N^{\mathrm{T}})$, which can be considered as an unbiased point estimation of the classification efficiency $\mathrm{E}\{e^{\mathrm{T}}\} = \epsilon$ at population level. Demonstrating the significance of e^{T}, can be approached as a statistical hypothesis testing problem, where H_0 —the null hypothesis— expresses that N^{T} is generated by a random classifier. H_0 is tested against the alternative hypothesis H_1: $\epsilon > \epsilon_0$, where ϵ_0 is the expected efficiency under H_0. We consider different methods which can be used to perform the significance test for $k \geq 2$: the exact probability test [2, 3], the Monte-Carlo test [4], and the χ^2 test (e.g. [8]).

In order to simplify the mathematical expressions, we introduce the following notations:

$$n_{[s]j} \triangleq \sum_{i=1}^{s} n_{ij} \ , \quad n_{i[t]} \triangleq \sum_{j=1}^{t} n_{ij} \ , \quad n_{[s][t]} \triangleq \sum_{i=1}^{s} \sum_{j=1}^{t} n_{ij} .$$

2 Methods

The Exact Probability Test. Testing the significance of the efficiency of a multiclass classifier can be performed by applying the following procedure[2]:

1. search for the distributions N, belonging to the set $\mathcal{N}_{>}$, defined as follows:

$$\mathcal{N} = \{N \mid N = (n_{11}, \ldots, n_{kk}) \in \mathbb{N}^{k^2}, n_{ij} \text{ satisfy } (1),(2)\},$$
$$\mathcal{N}_{>} = \{N \mid N \in \mathcal{N}, n_{ij} \text{ satisfy } (3)\},$$

referring to the constraints:

$$\forall i \in \{1, \ldots, k\} \ : \ n_{i[k]} = n^{\mathrm{T}}_{i[k]}, \tag{1}$$
$$\forall j \in \{1, \ldots, k\} \ : \ n_{[k]j} = n^{\mathrm{T}}_{[k]j}, \tag{2}$$
$$\sum_{i=1}^{k} n_{ii} \geq \sum_{i=1}^{k} n^{\mathrm{T}}_{ii}; \tag{3}$$

2. calculate

$$p_{>} = \sum_{N \in \mathcal{N}_{>}} \prod_{s=2}^{k} \prod_{t=2}^{k} \frac{\mathrm{C}_{n_{[s-1][t]}}^{n_{[s-1]t}} \mathrm{C}_{n_{s[t]}}^{n_{st}}}{\mathrm{C}_{n_{[s][t]}}^{n_{[s]t}}}, \tag{4}$$

where $\mathrm{C}_n^{n'} = n! / \left((n - n')! n'!\right)$ denotes the number of combinations in which n' objects can be selected from a set of n distinct objects;

3. if $p_{>} < \alpha$ (α is a chosen significance level), the null hypothesis H_0 that the distribution N^{T} was obtained through random classification, is rejected in favour of the alternative hypothesis H_1 that the efficiency of the classifier under study is higher than expected under H_0.

This procedure is based on the idea that distribution N^{T} is a member of the population of distributions N with the same marginal totals in the rows and columns of the $k \times k$ contingency table (i.e., a square table containing the n_{ij} values, where i is the row index and j, the column index) —which is expressed by constraints (1) and (2). The terms in (4) are the exact probabilities, under H_0, of those distributions showing the same or a higher classification efficiency than N^{T} —which is expressed by constraint (3).

The Monte-Carlo Test. This method is based on a comparison of the distribution N^{T} with the members of a sample of distributions $\mathcal{N}^{\mathrm{R}}$, which have the same marginal distributions as N^{T} and which are generated by a truly random classifier. The generation of distribution sample $\mathcal{N}^{\mathrm{R}}$ can easily be performed through stochastic simulation: in order to obtain one member of it, it sufficient

- to start with a sample of patterns containing for each class Ω_i, $n^{\mathrm{T}}_{i[k]}$ class members,
- to rank these patterns randomly, yielding a random series,
- to assign the first $n^{\mathrm{T}}_{[k]1}$ pattern of this series to class Ω_1, the subsequent $n^{\mathrm{T}}_{[k]2}$ patterns to class Ω_2 ... and so on.

The size f of $\mathcal{N}^{\mathrm{R}}$ (i.e., $f = \#\{N \mid N \in \mathcal{N}^{\mathrm{R}}\}$) should be chosen carefully. The author suggests to use an f value, exceeding the values f_{min} given by Table 1 [4].

α	f_{min}
0.05	5024
0.01	26074
0.005	52386
0.001	262880

Table 1. Suggested values for f_{min} for different values of the significance level α.

In order to perform the Monte-Carlo test[6], we define the frequency distribution f_l of the efficiencies e(N), calculated for the elements of sample $\mathcal{N}^{\mathrm{R}}$, i.e.:

$$\forall l \in 1,\ldots,n^{\mathrm{T}}_{[k][k]} \;:\; f_l = \#\{N \mid N \in \mathcal{N}^{\mathrm{R}}, \sum_{i=1}^{k} n_{ii} = l\}.$$

The test procedure consists of calculating

$$p'_{>} = \frac{\left(\sum_{l=l^{\mathrm{T}}}^{n^{\mathrm{T}}_{[k][k]}} f_l\right) + 1}{f+1} \quad \text{with} \quad l^{\mathrm{T}} = \sum_{i=1}^{k} n^{\mathrm{T}}_{ii},$$

and rejecting the null hypothesis H_0 if $p'_{>} < \alpha$ (α is a chosen significance level).

The χ^2 Test. Here again, distribution N^{T}, obtained through the classification experiment on a test set, is regarded as an instance of the population of distributions, having the same marginal distributions. Under the null hypothesis H_0 that the population was obtained through random classification, the probability that a pattern, belonging to class Ω_i, is assigned by the classifier to Ω_j, is independent of the former class; this probability is obviously $n^{\mathrm{T}}_{[k]j}/n^{\mathrm{T}}_{[k][k]}$. Therefore, since the subsample size of the patterns, belonging to class Ω_i, is $n^{\mathrm{T}}_{i[k]}$, the expected value n^{E}_{ij} of n_{ij} under H_0, is:

$$n^{\mathrm{E}}_{ij} = \frac{n^{\mathrm{T}}_{i[k]} n^{\mathrm{T}}_{[k]j}}{n^{\mathrm{T}}_{[k][k]}}.$$

In order to obtain a χ^2 testing procedure for the appropriate alternative hypothesis as specified in Sect. 1, we define n_{c} and n_{w}:

$$n_{\mathrm{c}} \triangleq \sum_{i=1}^{k} n_{ii} \quad \text{and} \quad n_{\mathrm{w}} \triangleq n_{[k][k]} - n_{\mathrm{c}},$$

symbolising resp. the number of correctly and wrongly classified patterns in a sample. The test starts with the calculation of

$$h = \frac{(n^{\mathrm{T}}_{\mathrm{c}} - n^{\mathrm{E}}_{\mathrm{c}})^2}{n^{\mathrm{E}}_{\mathrm{c}}} + \frac{(n^{\mathrm{T}}_{\mathrm{w}} - n^{\mathrm{E}}_{\mathrm{w}})^2}{n^{\mathrm{E}}_{\mathrm{w}}}.$$

Tables of critical values $\chi^2_{\mathrm{crit}}(df = 1)$ or probabilities p_{crit} associated to critical values for the χ^2 test, can be found in any standard text book on statistics (e.g. [8]). The null hypothesis H_0 should however only be rejected, when *two* conditions are satisfied *simultaneously*:

1. $n^{\mathrm{T}}_{\mathrm{c}} > n^{\mathrm{E}}_{\mathrm{c}}$;
2. $h > \chi^2_{\mathrm{crit}}(df = 1)$.

Consequently, since under H_0 condition 1 occurs with a probability of 0.5 and is independent of condition 2, the tabulated probability p_{crit} must be halved. To enable a comparison between the different techniques, we set $p''_{>} = p_{\mathrm{crit}}/2$ and describe the testing procedure as rejecting H_0 in favour of H_1, when condition 1 is verified, and $p''_{>} < \alpha$ (α is a chosen significance level).

3 Experiments

3.1 Experiments Involving Ulcer Patient Data

The methods have been applied to the distribution, already described in [2], shown in Fig. 1 and taken from [1]. The exact probability test, applied to this example, yields $p_{>} = 5.85 \times 10^{-5}$. The Monte-Carlo test, used with f values of $10^3, 10^4, 10^5, 10^6, 10^7$, yields $p'_{>}$ values of resp. 10^{-3}, 10^{-4}, 6.0×10^{-5}, 6.6×10^{-5}, 5.87×10^{-5}. The χ^2 testing method yields: $p''_{>} = 4.08 \times 10^{-5}$.

We also reduced artificially the size of the test set, by replacing the entries n^{T}_{ij} of the contingency table by the rounded result of $n^{\mathrm{T}}_{ij}/3$, and we applied both the exact probability test and the χ^2 test. The methods yield resp. $p_{>} = 1.13\%$ and $p''_{>} = 0.68\%$.

9	7	3
15	17	13
3	7	28

Fig. 1. Contingency table, showing classification results for the data of 102 ulcer patients; the three classes correspond to three types of pathology evolution.

3.2 Evaluation of Time Behaviour

The exact probability test and the Monte-Carlo test have been implemented in the C programming language on a SPARCstation 1, under the Solaris UNIX operating system. Source code can be obtained from the author (via electronic mail only).

In order to illustrate the behaviour of the exact probability testing method, we applied it to a number of uniform N^{T} distributions, i.e. where all n^{T}_{ij} are equal to the same value n. Figure 2 shows for different values of k and n the size of $\mathcal{N}_>$ and consequently, the number of terms that have to be evaluated in the right hand side of (4). For $k = 5$ and $n^{\mathrm{T}}_{i[k]} = n^{\mathrm{T}}_{[k]j} = 5$ (for all values of i and j) — like in the last row of the table in this figure—, but with $e^{\mathrm{T}} = 40\%$, the number of terms in (4) is still 1254250. In order to calculate the exact probability value ($p_> = 1.95\%$), the mentioned computer program needs about 393s of computing time; the Monte-Carlo testing program needed about 3.4s, to obtain a $p'_>$ value of 2.04% with $f = 30000$; (the χ^2 test yielded $p''_> = 0.62\%$).

k	n	# of terms	$p_>$
3	1	30	0.61
	2	204	0.58
	3	748	0.57
4	1	5532	0.59
	2	501394	0.55
5	1	11809690	0.58

Fig. 2. Number of terms, to be calculated and summed to yield $p_>$ in the worst case of a contingency table showing a uniform distribution (all $n^{\mathrm{T}}_{ij} = n$) of the test set patterns over the k classes.

In order to show the applicability of the Monte-Carlo method for much larger problems, we applied it to a problem involving 10000 test patterns: $k = 10$, $n^{\mathrm{T}}_{i[k]} = 1000$ and $n^{\mathrm{T}}_{[k]j} = 1000$ (for all values of i and j). We applied the technique with $f = 30000$ (for different efficiency values); in all cases, the computing time was about 1440s. For a classification efficiency of 10.5%, we obtained $p'_> = 5.0\%$ (χ^2 test: $p''_> = 4.8\%$); for a classification efficiency of 11%, we obtained $p'_> = 7.7 \times 10^{-4}$ (χ^2 test: $p''_> = 4.3 \times 10^{-4}$).

4 Discussion and Conclusion

From a theoretical point of view, the most appropriate test method of the ones, proposed here, is the exact probability method:

- it maximally exploits all available information in the distribution of the test set patterns and in that sense it is power efficient;
- its application does not require the choice of other test procedure parameters than α, the significance level (unlike the Monte-Carlo method, which requires the choice of f);
- it does not require special assumptions, like the χ^2 test where the calculated h-variable is assumed to satisfy a χ^2 distribution.

As a first conclusion, we recommend the reader to use the exact probability test, whenever the size of the problem does not prevent its use due to combinatorial explosion (as described in Sect. 3.2 and illustrated by Fig. 2).

The Monte-Carlo test is as reliable as the exact probability test, but its power is greatly influenced by the statistical experiment, which not only consists of observing the distribution N^{T} of test set patterns, but also includes the observation of a set $\mathcal{N}^{\mathrm{R}}$ of distributions, generated through the stochastic simulation of a random classifier. The null hypothesis H_0 is actually the hypothesis that N^{T} and the f members of $\mathcal{N}^{\mathrm{R}}$ are taken from the same population (i.e.: obtained through random classification). Choosing a small f-value weakens the testing procedure as demonstrated in Sect. 3.1; as a consequence, $p'_>$ values are higher than the exact probability value $p_>$; it can however been shown that $\lim_{f\to\infty} p'_> = p_>$ [4]. The time behaviour of the method for larger problems on the other hand, is more favourable than the exact probability method; time consumption is practically independent of k and α and is linearly dependent on the total number $n^{\mathrm{T}}_{[k][k]}$ of test set patterns. As demonstrated in Sect. 3.2, the method can be successfully applied for large test sets, provided that the f-value is chosen with care (e.g., using the table of Fig. 1). In [4], a slightly more complicated, but more refined methodology for choosing f, is described. As conclusion of this paragraph, we recommend the reader to use the Monte-Carlo test method as an alternative to the exact probability test; its application is also limited by the size of the problem and by the quality of the pseudo-random number generator, used during stochastic simulation (we use function `ran0()` (cf. [5])).

The χ^2 test method outperforms the other methods so far as computing time consumption is concerned: it essentially involves the calculation of the value of h (requiring mainly a sum of k^2 values to obtain $n^{\mathrm{T}}_{[k][k]}$, a sum of k values to obtain $n^{\mathrm{T}}_{\mathrm{c}}$ and a sum of k simple expressions to obtain $n^{\mathrm{E}}_{\mathrm{c}}$) and the calculation of $p''_>$ or $\chi^2_{\mathrm{crit}}(df = 1)$. When applying the χ^2 testing method, one should however bear in mind that it is based on an assumption which is never fully satisfied and in some cases may severely be violated: the calculated test variable h is assumed to satisfy a χ^2-distribution. This probably explains why the values of $p''_>$ are systematically lower than $p_>$. (Like in parametric hypothesis testing methods, the introduction of supplementary assumptions enhances the power of the test).

Also, the difference between $p''_>$ and $p_>$ decreases for increasing test set sizes and for increasing significance levels. Textbooks (e.g. [7]) mention that a valid χ^2 test, applied to the cell frequencies of a contingency table, can only be performed under very specific conditions: all expected frequencies n^{E}_{ij} should at least be 1 and at least 80% of the cells should have an expected frequency, exceeding 5. The example of Sect. 3.2, where $k = 5$, $n^{\mathrm{T}}_{i[k]} = n^{\mathrm{T}}_{[k]j} = 5$ and $e^{\mathrm{T}} = 40\%$ is an example where the method should not be applied (which is confirmed by the unacceptable difference between $p''_> = 0.62\%$ and $p_> = 1.95\%$). The last paragraphs of 3.1 and 3.2 illustrate that, even when the mentioned criteria are satisfied, one should be careful in using the χ^2 method. We conclude this paragraph by suggesting to use the χ^2 test only when the size of the problem prohibits the use two other methods; in that case, we recommend the choice of a sufficiently high significance level.

References

1. Robert Jennrich and Paul Sampson. Stepwise discriminant analysis. In Wilfrid Joseph Dixon, editor, *BMDP Statistical Software Manual.* University of California Press, 1988.
2. Edgard Nyssen. Evaluation of pattern classifiers — testing the significance of classification efficiency using an exact probability technique. *Pattern Recognition Letters*, 17(11):1125–1129, September 1996.
3. Edgard Nyssen. Interpretation of pattern classification results, obtained from a test set. In *Proceedings 1st IAPR TC1 Intl. Workshop on Statistical Techniques in Pattern Recognition, STIPR'97*, pages 103–105, Prague, June 1997.
4. Edgard Nyssen. Evaluation of pattern classifiers — applying a Monte-Carlo significance test to the classification efficiency. *Pattern Recognition Letters*, 1998. Accepted for publication.
5. William H. Press, Saul A. Teukolsky, and William T. Vetterling. *Numerical recipes in C: the art of scientific computing.* University Press : Cambridge, 1995.
6. Brian D. Ripley. *Stochastic Simulation.* John Wiley & Sons, 1987.
7. Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers and Scientists.* John Wiley & Sons, 1987.
8. Sidney Siegel. *Nonparametric Statistics for the Behavioural Sciences.* McGraw Hill, 1956.