KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 443 - 446

INTERPRETATION OF PATTERN CLASSIFICATION RESULTS, OBTAINED FROM A TEST SET

EDGARD NYSSEN[1]

The present paper presents and discusses a methodology for interpreting the results, obtained from the application of a pattern classifier to an independent test set. It addresses the problem of testing the random classification null hypothesis in the multiclass case, by introducing an exact probability technique. The discussion of this technique includes the presentation of an interval estimation technique for the probability of correct classification, which is slightly more accurate than the ones described in some statistics textbooks.

1. INTRODUCTION

Consider an experiment where a k-class classifier under study is applied to a test set of patterns which is independent of the learning set. The result of this experiment — referred to by superscript T — 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 $(i, j \in \{1, \ldots, k\})$.

A simplified representation of the mathematical expressions will be obtained by introducing the following notations:

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

The classification efficiency e is defined as the correctly classified fraction of the observations.

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

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.

An important problem — especially when only a small test set is studied — is the demonstration that this efficiency is statistically significant.

[1]Supported by the Grant for Program No. G.0042.96 of the Belgian National Research Fund (NFWO).

2. TESTING THE SIGNIFICANCE OF THE EFFICIENCY

In the case that only two classes are involved, Fisher's exact probability test is the appropriate method to test the significance of efficiency [4]. The methodology, discussed in this paper, involves an arbitrary number of classes.

Let us consider the population $\mathcal{N}_{H_0}$ of frequency distributions N, generated through random classification of the patterns in the test set and satisfying the same marginal totals as N^T, i.e.:

$$N = (n_{11}, \ldots, n_{kk}) \ , \quad \forall i,j \in \{1,\ldots,k\} : n_{i[k]} = n^T_{i[k]} \ , \quad n_{[k]j} = n^T_{[k]j}. \tag{2.1}$$

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 taken from the population $\mathcal{N}_{H_0}$ (and hence, generated by a random classifier). H_0 is tested against the alternative hypothesis H_1: $E\{e^T\} > \epsilon_0$; here, ϵ_0 is the expected efficiency under H_0 [3].

To implement this approach, it is sufficient that we are able to calculate for any distribution N the probability $P(N \mid H_0, \text{Eq. 2.1})$ of its occurrence under H_0 and conditioned by the constraints given by Eq. 2.1. From this probability, one can derive the probability distribution p_e —under H_0— of the test statistic e as well as the minimal significance level $p_>$ at which H_0 can be rejected on the basis of N^T:

$$p_> = \sum_{e=e^T}^{1} p_e \quad \text{with} \quad p_e = \sum_{N:e(N)=e} P(N \mid H_0, \text{Eq. 2.1}). \tag{2.2}$$

Under H_0, all patterns have the same probability of being assigned by the classifier to a given class, independently of their real class membership. This means that all possible distinct assignments of the $n_{[k][k]}$ patterns to the k classes, satisfying Eq. 2.1, are equiprobable. This leads us to the following expression for $P(N \mid H_0, \text{Eq. 2.1})$:

$$\begin{aligned} P(N \mid H_0, \text{Eq. 2.1}) &= \underbrace{\left(\frac{n_{[k][k]}!}{\prod_{j=1}^{k} n_{[k]j}!} \right)^{-1}}_{\text{(A)}} \underbrace{\left(\prod_{i=1}^{k} \frac{n_{i[k]}!}{\prod_{j=1}^{k} n_{ij}!} \right)}_{\text{(B)}} \\ &= \prod_{s=2}^{k} \prod_{t=2}^{k} \underbrace{\frac{C^{n_{[s-1]t}}_{n_{[s-1][t]}} C^{n_{st}}_{n_{s[t]}}}{C^{n_{[s]t}}_{n_{[s][t]}}}}_{\text{(C)}} \end{aligned} \tag{2.3}$$

where expression (A) contains the number of the distinct assignments, mentioned before, and expression (B) contains the number of distinct assignments, satisfying distribution N. In expression (C), one recognises the hypergeometric distribution probability. Note that a hypothesis test based on Eq. 2.2 and 2.3, reduces exactly to Fisher's exact probability test, when $k = 2$.

3. DISCUSSION

The hypothesis testing technique, presented in this paper, is based on the calculation of an exact probability. Like Fisher's test, the technique can therefore be applied to classification results, obtained from arbitrarily small test sets. The technique doesn't evaluate the importance of the efficiency: even when a low efficiency e^{T} is observed for the test set, this efficiency can still be significantly higher than the efficiency ϵ_0, expected under H_0. To illustrate this, consider as example the case of a test set with the same subset size for each class, i.e.: $\forall i \in \{1,\ldots,k\} : n^{\mathrm{T}}_{i[k]} = n^{\mathrm{T}}_{[k][k]}/k$. In that case, ϵ_0 equals $1/k$. If we consider specifically an experiment on a test set, involving $k = 4$ classes, where for all i and j values ($i \neq j$), $n_{ii} = 3$ and $n_{ij} = 1$, we have $\epsilon_0 = 0.25$ and $e = 0.5$; athough e is low, we have $p_> = 0.0085$, which means that the observed efficiency e is significantly higher than ϵ_0 at a significance level of e. g. 1 %.

An interval estimation of ϵ, the efficiency expected for a set of patterns, satisfying $\forall i \in \{1,\ldots,k\} : n_{i[k]} = n^{\mathrm{T}}_{i[k]}$, can be calculated on the basis of any existing appropriate technique for the interval estimation of proportions. When $n_{[k][k]}$ is large, e is approximately normally distributed. In classical handbooks on statistics, on can find the necessary fomulæ, yielding for our problem a $\gamma(\times 100\,\%)$ confidence interval of $e \pm c\sqrt{e(1-e)/n_{[k][k]}}$, where c is half the interval width holding for a standard normally distributed statistic z which satisfies: $\mathrm{P}(-c \leq z \leq c) = \gamma$ (e. g. [1, 2]). This estimation is an approximation, not only because of the approximating normality assumption for e, but because it is derived from the interval estimate $e \pm c\sqrt{\epsilon(1-\epsilon)/n_{[k][k]}}$, where ϵ, which is evidently unknown, is replaced by its point estimate, e. One undesirable consequence is that an interval limit may become negative or exceed 1, which occurs when e or $(1-e)$ has a value less than $c^2/(c^2+n_{[k][k]})$.

A slightly more accurate result is obtained through the following reasoning. Assume that $n_{[k][k]}$ is sufficiently large for e to be normally distributed in good approximation. In that case, e and ϵ satisfy:

$$\mathrm{P}\left(\epsilon - c\sqrt{\frac{\epsilon(1-\epsilon)}{n_{[k][k]}}} \leq e \leq \epsilon + c\sqrt{\frac{\epsilon(1-\epsilon)}{n_{[k][k]}}}\right) \approx \gamma.$$

Note that if this equation holds approximatively, this is only due to the deviation from normality of the distribution of e. The left hand side can be rewritten as:

$$\mathrm{P}\left(|e-\epsilon| \leq c\sqrt{\frac{\epsilon(1-\epsilon)}{n_{[k][k]}}}\right) = \mathrm{P}\left((e-\epsilon)^2 - c^2\,\frac{\epsilon(1-\epsilon)}{n_{[k][k]}} \leq 0\right).$$

After putting $d = c^2/n_{[k][k]}$, the values of ϵ, which satisfy this inequallity, are situated between the roots of the quadratic equation $(e-\epsilon)^2 - d(\epsilon(1-\epsilon)) = 0$ and therefore, the interval estimate for ϵ —at (approximately) a confidence level $\gamma(\times 100\,\%)$— is given by:

$$e_{\mathrm{max\,/\,min}} = \frac{2e + d \pm \sqrt{(2e+d)^2 - 4(1+d)e^2}}{2(1+d)}.$$

This interval $[e_{\mathrm{min}}, e_{\mathrm{max}}]$ has the following properties:

- $e \in [e_{\min}, e_{\max}] \subset [0,1]$; e —the unbiased point estimate of ϵ— is however not the centre of $[e_{\min}, e_{\max}]$ (except when $e = 0.5$);
- $P(\epsilon < e_{\min}) = P(\epsilon > e_{\max}) \approx (1-\gamma)/2$, which is consistent with the previous property in case e is close to 0 or 1.

Let us consider as example, a test set of size 100, where 99 patterns were classified correctly. The estimation method described here, yields for ϵ a 95% confidence interval [0.946,0.998]; the classical method yields [0.970,1.01], which has an upper limit exceeding 1.

(Received December 18, 1997.)

REFERENCES

[1] H. M. Blalock: Social Statistics. International Student Edition, 1979.
[2] D. C. Montgomery and G. C. Runger: Applied Statistics and Probability for Engineers. Wiley, New York 1994.
[3] E. Nyssen: Evaluation of pattern classifiers – testing the significance of classification efficiency using an exact probability technique. Pattern Recognition Lett. *17* (1996), 11, 1125–1129.
[4] S. Siegel: Nonparametric Statistics for the Behavioural Sciences. McGraw Hill, New York 1956.

Prof. Edgard Nyssen, PhD., VUB – Faculteit T.W., Dienst Elektronica (Brussels Free University – Faculty of Applied Sciences, Electronics Department), Pleinlaan 2, B1050 Brussel. Belgium.
e-mail:ehnyssen@etro.vub.ac.be