A fast randomisation test for rule significance

Günther Gediga
FB Psychologie / Methodenlehre
Universität Osnabrück
49069 Osnabrück, Germany
gg@Luce.Psycho.Uni-Osnabrueck.DE

Ivo Düntsch
School of Information and Software Engineering
University of Ulster
Newtownabbey, BT 37 0QB, N.Ireland
I.Duentsch@ulst.ac.uk

Abstract

Randomisation is a method to test the statistical significance of a symbolic rule; it is, however, very expensive. In this paper we present a sequential randomisation test which dramatically reduces the number of steps needed for a conclusion.

1 Introduction

One problem of rule based data analysis is that the validity of a rule may be given, while its statistical significance is not: If rules are based on a few observations only, the granularity of the system is too high, and the rule may be due to chance. In order to test the significance of rules, one can use randomisation methods [4] to compute the conditional probability of the rule, assuming that the null hypothesis

"Objects are randomly assigned to decision classes"

is true. These procedures seem to be particularly suitable to non-invasive techniques of data mining such as rough set data analysis, since randomisation tests do not assume that the available data is a representative sample. This assumption is a general problem of statistical data mining techniques; the reason for this is the huge state complexity of the space of possible rules, even when there is only a small number of features. However, a drawback of randomisation is its costliness, and it would be of great value to have a less expensive procedure which has similar (few) model assumptions, and still gives us a reliable significance test.

In [2] we have developed two procedures, both based on randomisation techniques, which evaluate the significance of prediction rules obtained in rough set dependency analysis. In the present paper, we continue this work and present a sequential randomisation test which is cheap and reliably determines the statistical significance of a rule system.

2 Rule systems

We use the terminology of rough set data analysis [5], and briefly explain the basic concepts. For more information on rough set data analysis we invite the reader to consult the forthcoming [3].

An *information system* is a tuple $\mathcal{I} = \langle U, \Omega, V_a \rangle_{a \in \Omega}$, where

1. U is a finite set of objects.
2. Ω is a finite set of mappings $a : U \to V_a$. Each $a \in \Omega$ is called an *attribute* or *feature*.

If $x \in U$, we denote by $Q(x)$ the feature vector of x determined by the attributes in Q. Each non-empty subset Q of Ω induces an equivalence relation θ_Q on U by

$$x \equiv_{\theta_Q} y \text{ iff } a(x) = a(y) \text{ for all } a \in Q,$$

i.e.

$$x \equiv_{\theta_Q} y \text{ iff } Q(x) = Q(y).$$

Objects which are in the same equivalence class cannot be distinguished with the knowledge of Q.

Equivalence relations θ_Q, θ_P are used to obtain rules in the following way: Let $Q \to P$ be the relation

$$\langle X, Y \rangle \in Q \to P \text{ iff } X \text{ is a class of } \theta_Q,\ Y \text{ is a class of } \theta_P, \text{ and } X \cap Y \neq \emptyset.$$

A pair $\langle X, Y \rangle \in Q \to P$ is called a ***Q,P – rule*** (or just a rule, if Q and P are understood) and usually written as $X \to Y$. By some abuse of language we shall also call $Q \to P$ a rule when there is no danger of confusion.

Each equivalence class X of θ_Q corresponds to a vector $\vec{X}$ of Q-features, and analogously for P. Thus, if the class X of θ_Q intersects exactly the classes $Y_1, \ldots, Y_n$ of θP, then we obtain the rule

$$\text{If } Q(y) = \vec{X}, \text{ then } P(y) = \vec{Y_1} \text{ or } \ldots \text{ or } P(y) = \vec{Y_n}. \tag{2.1}$$

A class X of θ_Q is called P – deterministic, if $n = 1$ in (2.1), i.e. if there is exactly one class Y of P which intersects, and thus contains, X. We define the *quality of an approximation* of a an attribute set Q with respect to an attribute set P by

$$\gamma(Q \to P) = \frac{|\bigcup\{X : X \text{ is a } P\text{–deterministic class of } \theta_Q\}|}{|U|}. \tag{2.2}$$

The statistic $\gamma(Q \to P)$ measures the relative frequence of correctly P – classified objects with the data provided by Q.

3 Randomisation

Suppose that $\emptyset \neq Q, P \subseteq \Omega$, and that we want to evaluate the statistical significance of the rule $Q \to P$. Let Σ be the set of all permutations of U, and $\sigma \in \Sigma$. We define new attribute functions a^σ by

$$a^\sigma(x) \stackrel{\text{def}}{=} \begin{cases} a(\sigma(x)), & \text{if } a \in Q, \\ a(x), & \text{otherwise.} \end{cases}$$

The resulting information system $\mathcal{I}_\sigma$ permutes the Q–columns according to σ, while leaving the P–columns constant; we let Q^σ be the result of the permutation in the Q–columns, and $\gamma(Q^\sigma \to P)$ be the approximation quality of the prediction of P by Q^σ in $\mathcal{I}_\sigma$.

The value

$$p(\gamma(Q \to P)|H_0) := \frac{|\{\gamma(Q^\sigma \to P) \geq \gamma(Q \to P) : \sigma \in \Sigma\}|}{|U|!} \tag{3.1}$$

now measures the significance of the observed approximation quality. If $p(\gamma(Q \to P)|H_0)$ is low, traditionally below 5%, then the rule $Q \to P$ is deemed significant, and the (statistical) hypothesis "$Q \to P$ is due to chance" can be rejected.

A simulation study done in [2] indicates that the randomisation procedure has a reasonable power if the rule structure of the attributes is known.

We see from the denominator $|U|!$ of $p(\gamma(Q \to P)|H_0)$ that the computational cost of obtaining the significance is feasible only for small values of $|U|$. A fairly simple tool to shorten the processing time of the randomisation test is the adaptation of a sequential testing scheme to the given situation. Because this sequential testing scheme can be used as a general tool in randomisation analysis, we present the approach in a more general way.

Suppose that θ is a a statistic with realizations θ_i, and a fixed realization θ_c. We can think of θ_c as $\gamma(Q \to P)$ and θ_i as $\gamma(Q^\sigma \to P)$. Recall that the statistic θ is called α – *significant*, if the true value $p(\theta \geq \theta_c|H_0)$ is smaller than α. Traditionally, $\alpha = 0.05$, and in this case, one speaks just of *significance*.

An evaluation of the hypothesis $\theta \geq \theta_c$ given the hypothesis H_0 can be done by using a sample of size n from the θ distribution, and counting the number k of θ_i for which $\theta_i \geq \theta_c$. The evaluation of $p(\theta \geq \theta_c|H_0)$ can now be done by the estimator $\hat{p}_n(\theta \geq \theta_c|H_0) = \frac{k}{n}$, and the comparison $\hat{p}_n(\theta \geq \theta_c|H_0) < \alpha$ will be performed to test the significance of the statistic. For this to work we have to assume that the simulation is asymptotically correct, i.e. that

$$lim_{n\to\infty}\hat{p}_n(\theta \geq \theta_c|H_0) = p(\theta \geq \theta_c|H_0). \tag{3.2}$$

In order to find a quicker evaluation scheme of the significance, it should be noted that the results of the simulation k out of n can be described by a binomial distribution with parameter $p(\theta \geq \theta_c|H_0)$.

The fit of the approximation of $\hat{p}_n(\theta \geq \theta_c|H_0)$ can be determined by the confidence interval of the binomial distribution.

In order to control the fit of the approximation more explicitly, we introduce another procedure within our significance testing scheme. Let

$$H_b : p(\theta \geq \theta_c|H_0)) \in [0, \alpha) \tag{3.3}$$

$$H_a : p(\theta \geq \theta_c|H_0)) \in [\alpha, 1] \tag{3.4}$$

be another pair of statistical hypotheses, which are strongly connected to the original ones: If H_b holds, we can conclude that the test is α–significant, if H_a holds, we conclude that it is not.

Because we want to do a finite approximation of the test procedure, we need to control the precision of the approximation; to this end, we define two additional error components:

1. $r =$ probability that H_a is true, but H_b is the outcome of the approximative test.
2. $s =$ probability that H_b is true, but H_a is the outcome of the approximative test.

The pair (r, s) is called the *precision* of the approximative test. To result in a good approximation, the values r, s should be small (e.g. $r = s = 0.05$); at any rate, we assume that $r + s \lneq 1$, so that $\frac{s}{1-r} \lneq \frac{1-s}{r}$, which will be needed below.

Using the Wald-procedure [6], we define the likelihood ratio

$$LQ(n) = \frac{\sup_{p\in[0,\alpha)} p^k(1-p)^{n-k}}{\sup_{p\in[\alpha,1]} p^k(1-p)^{n-k}}, \tag{3.5}$$

and we obtain the following approximative sequential testing scheme:

1. If
$$LQ(n) \lneq \frac{s}{1-r},$$
then H_a is true with probability at most s.

2. If
$$LQ(n) \gneq \frac{1-s}{r},$$
then H_b is true with probability at most r.

3. Otherwise
$$\frac{s}{1-r} \leq LQ(n) \leq \frac{1-s}{r},$$
and no decision with precision (r, s) is possible. Hence, the simulation must continue.

With this procedure, which is implemented in our rough set engine GROBIAN[1] [1], the computational effort for the significance test in most cases breaks down dramatically, and a majority of the tests need less than 100 simulations.

[1] http://www.infj.ulst.ac.uk/~cccz23/grobian/grobian.html

References

[1] Düntsch, I. & Gediga, G. (1997a). The rough set engine GROBIAN. In A. Sydow (Ed.), *Proc. 15th IMACS World Congress, Berlin*, vol. 4, 613–618, Berlin. Wissenschaft und Technik Verlag.

[2] Düntsch, I. & Gediga, G. (1997b). Statistical evaluation of rough set dependency analysis. *International Journal of Human–Computer Studies*, **46**, 589–604.

[3] Düntsch, I. & Gediga, G. (2000). Rough set data analysis. In *Encyclopedia of Computer Science and Technology*. Marcel Dekker. To appear.

[4] Manly, B. F. J. (1991). Randomization and Monte Carlo Methods in Biology. London: Chapman and Hall.

[5] Pawlak, Z. (1982). Rough sets. *Internat. J. Comput. Inform. Sci.*, **11**, 341–356.

[6] Wald, A. (1947). Sequential Analysis. New York: Wiley.