BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 24 2011, pages 3439–3440
doi:10.1093/bioinformatics/btr574

Data and text mining

Advance Access publication October 13, 2011

BDTcomparator: a program for comparing binary classifiers

Kamil Fijorek[1,*], Damian Fijorek, Barbara Wisniowska[2] and Sebastian Polak[2]

[1]Department of Statistics, Cracow University of Economics, 31-510 Cracow and [2]Unit of Pharmacoepidemiology and Pharmacoeconomics, Jagiellonian University, Medical College, 30-688 Cracow, Poland

Associate Editor: Jonathan Wren

ABSTRACT

Summary: The BDTcomparator facilitates the selection of the best performing binary classification model or binary diagnostic procedure from the many possible alternatives by comparing their predictions with a known output, measured with the use of a system recognized as the gold standard. The program calculates the estimates of accuracy, sensitivity, specificity, predictive values and diagnostic likelihood ratios along with appropriate confidence intervals. Furthermore, all pairwise comparisons with respect to the above-mentioned measures are calculated. The formatted results can be exported to a text-file.

Availability and Implementation: BDTcomparator is distributed under the GNU GPLv3 license and is freely available for download from http://www.tox-portal.net. We provide programs for both Linux and Windows operating systems. The source code of the program is provided in our companion website http://code.google.com/p/bdtcomparator/.

Contact: kamil.fijorek@uek.krakow.pl

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on August 13, 2011; revised on August 13, 2011; accepted on October 9, 2011

1 INTRODUCTION

Data analysis is one of the main driving forces of modern science. Among others, it results in the fast development of disease screening procedures and classification algorithms for data mining. In both areas, the ultimate goal is to select the best performing model or diagnostic procedure (DP) from the many possible alternatives by comparing their predictions with a known output, measured with the use of a system recognized as a gold standard. Furthermore, one may be interested in pairwise performance comparisons between the many competing models/DPs. In our report, we focus on the binary DPs/classification problems (both prediction and measured output can be either zero or one) and propose comprehensive, freely available software supporting this type of analyses. A real-life example of its functioning is presented based on the drug safety assessment problem.

2 STATISTICAL METHODS

In this section, we describe the statistical methods that are the basis of our program. The terminology used takes the clinical binary diagnostic testing point of view; however, the same methods can be applied to data that originated from other binary classification tasks. More detailed formulas are given in the Supplementary Materials.

Table 1. Cross-classification of DP results by GS status

DP_k\GS	GS+	GS−	Total
DP_k+	A	B	A+B
DP_k-	C	D	C+D
Total	A+C	B+D	N=A+B+C+D

Table 2. Performance measures of a DP

Performance measure	Formula
Diagnostic accuracy	ACC=(A+D)/N
Sensitivity	SE=P(DP_k+|GS+)=A/(A+C)
Specificity	SP=P(DP_k-| GS−)=D/(B+D)
Positive predictive value	PPV=P(GS+| DP_k+)=A/(A+B)
Negative predictive value	NPV=P(GS−| DP_k-)=D/(C+D)
Diagnostic likelihood ratio of a positive test	DLR(+)=SE/(1−SP)
Diagnostic likelihood ratio of a negative test	DLR(−)=(1−SE)/SP

Assume that N patients were tested for the presence/absence (positive/negative cases) of some characteristic using $K \geq 1$ ($k = 1, \ldots, K$) DPs. The true status of the patients is known as a result of applying the gold standard (GS). The outcome of this experiment for the k-th DP is summarized in Table 1.

In order to describe the ability of the k-th DP to discriminate between positive and negative cases, we use the performance measures given in Table 2.

Given the empirically observed values of A, B, C and D for the k-th DP, we calculate the estimates of performance measure and also the appropriate confidence intervals (CIs) to quantify the uncertainty in the estimates. CIs for ACC, SE, SP, PPV and negative predictive value (NPV) are constructed using the method of Clopper and Pearson (Agresti, 2002). CIs for diagnostic likelihood ratios (DLRs) are calculated by the method of Simel *et al.* (1991). A description of the performance of each individual DP is followed by a comparison between them. To test the null hypothesis of the equality of the diagnostic accuracies (ACC) of multiple DPs, we use Cochran's Q test. The alternative hypothesis states that at least two diagnostic accuracies are different. To test the equality of multiple sensitivities (specificities), we use Cochran's Q applied only to GS+ patients (GS- patients). The rejection of the Cochran's Q null

*To whom correspondence should be addressed.

hypothesis is followed by pairwise comparisons using McNemar's test accompanied by the appropriate CIs for the difference in the two proportions (Fleiss *et al.*, 2003). We do not use any corrections for multiple testing; however, our program allows the user to freely modify the confidence level which may be used to obtain Bonferroni-corrected CIs. The pairwise comparisons of predictive values and DLRs are based on the methods of Moskowitz and Pepe (2006) and Nofuentes and Castillo (2007). Consequently, the CIs are built for the ratios of two predictive values or DLRs.

To the best of our knowledge, the BDTcomparator is the first available program that integrates the described above statistical results.

3 IMPLEMENTATION DETAILS

The BDTcomparator (Binary Diagnostic Tests Comparator) is implemented in an object-oriented fashion using C++ programming language. The Graphical User Interface was made with a Qt cross-platform application framework (http://qt.nokia.com).

The input data has to be given in a tab-delimited text file (an example data-file is available with the program). The first line has to contain variables names, any subsequent line represents a row of data (valid values are either ones or zeros). The output is saved as a tab-delimited text file.

Typical usage is as follows:

- open an input data file—the first icon in the top menu,
- set options on the right side of the program window,
- calculate the output—the second icon in the top menu,
- in the output tab use the combo box to browse through results and
- save the output to a text file—the third icon in the top menu.

4 EXAMPLE OF USE

Drug safety is a rising concern for patients, health-care providers, pharmaceutical companies and insurers. According to the available data, cardiotoxicity is one of the major issues. The mechanism is mainly based on the drug—hERG channels interaction and respective ionic current inhibition. The GS laboratory technique is based on the cell membrane electric potential measurement and done with the use of the PatchClamp technique (Polak *et al.*, 2009). To avoid this costly procedure, safety screening mathematical models are introduced at the early stage of the drug development process. Their task is defined as a binary classification problem where chemicals are classified as potential blockers and non-blockers. In our case, various modeling techniques and algorithms were used to develop a predictive model based on the previously collected literature data. Their predictions were compared with the known output obtained from the laboratory PatchClamp measurements. The BDTcomparator was used to calculate relevant statistics for nine developed models. The program's output supported the choice of the neuro-fuzzy system with one hidden layer (20 neurons) and the linear activation function. The estimated diagnostic accuracy and NPV (the model classifies potentially dangerous drugs) of this model was 76.1% (95% CI: 72.1–79.8%) and 84.0% (95% CI: 77.6–89.2%), respectively. The NPV of this model, however, was not significantly higher than the second best NPV ($P = 0.186$, 95% CI: 0.971–1.161). The model was further utilized as a part of the drug cardiac safety assessment platform (www.tox-portal.net). The Supplementary Materials contain the detailed problem description and calculation results.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (2002) *Categorical Data Analysis*. John Wiley and Sons, Hoboken, New Jersey.

Fleiss,J.L. *et al.* (2003) Statistical Methods for Rates and Proportions. John Wiley and Sons, Hoboken, New Jersey.

Moskowitz,C.S. and Pepe,M.S. (2006) Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clinical Trials*, **3**, 272–279.

Nofuentes,J.A.R. and Castillo, J.D.D.L. (2007) Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. *Stat. Med.*, **26**, 4179–4201.

Polak,S. *et al.* (2009) Collation, assessment and analysis of literature in vitro data on hERG receptor blocking potency for subsequent modeling of drugs cardiotoxic properties. *J. Appl. Toxicol.*, **29**, 183–206.

Simel,D.L. *et al.* (1991) Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J. Clin. Epidemiol.*, **44**, 763–770.