BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 11 2014, pages 1635–1636
doi:10.1093/bioinformatics/btu086

Data and text mining

Advance Access publication February 10, 2014

HUM calculator and HUM package for R: easy-to-use software tools for multicategory receiver operating characteristic analysis

Natalia Novoselova[1,2,*], Cristina Della Beffa[2], Junxi Wang[2], Jialiang Li[3], Frank Pessler[4] and Frank Klawonn[2,5]

[1]Laboratory of Bioinformatics, United Institute of Informatics Problems, National Academy of Sciences of Belarus, Surganova 6, 220012 Minsk, Belarus, [2]Bioinformatics and Statistics, Helmholtz Centre for Infection Research, Braunschweig, Germany, [3]Department of Statistics and Applied Probability, National University of Singapore, Singapore, [4]TWINCORE Center for Experimental and Clinical Infection Research, Hannover and [5]Department of Computer Science, Ostfalia University of Applied Sciences, Wolfenbüttel, Germany

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Receiver operating characteristic (ROC) analysis is usually applied in bioinformatics to evaluate the abilities of biological markers to differentiate between the presence or absence of a disease. It includes the derivation of the useful scalar performance measure area under the ROC curve for binary classification tasks. As real applications often deal with more than two classes, multicategory ROC analysis and the corresponding hypervolume under the manifold (HUM) measure have become a topic of growing interest. To support researchers in carrying out multicategory ROC analysis, we have developed two tools in different programming environments which feature user-friendly, object-oriented and flexible interfaces and enable the user to compute HUM values and plot 2D- and 3D-ROC curves.

Availability: The software is freely available from our Web site http://public.ostfalia.de/~klawonn/HUM.htm

Contact: novos65@mail.ru

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

Received on November 26, 2013; revised on January 23, 2014; accepted on February 5, 2014

1 INTRODUCTION

Receiver operating characteristic (ROC) analysis as an alternative to classification accuracy has been used extensively to estimate the discriminative power of classifiers in two-class classification problems (Bradley, 1997; Fawcett, 2006; Sonego *et al*., 2008). During the last decade several theoretical investigations in the field of multicategory ROC analysis have provided different approaches to calculate the hypervolume under the manifold (HUM), the equivalent to the area under the ROC curve (AUC) of binary ROC analysis in a multidimensional setting. The special case of 3-class ROC analysis is described in (Sonego *et al*., 2008), where for each class the two possible misclassifications are treated equally (a so-called one-versus-rest scenario). In (Hand and Till, 2001) the average of all one-versus-rest AUCs is proposed as an approximation of the AUC. We recently reported a new extension of ROC analysis which uses a simple scheme to compare two or more classifiers in their abilities to differentiate among multiple categories (Li and Fine, 2008; Li and Zhou, 2009). This approach to multicategory ROC analysis expresses the discriminatory ability of the classifier as a HUM measure, with perfect discrimination corresponding to a HUM of 1 and the HUM of the null hypothesis to $1/n!$, where n corresponds to the number of categories to be separated. Several R packages for two-class ROC analysis exist (Robin *et al*., 2011), but they do not allow to take into account multiple categories. We here present two software tools that should make multicategory ROC analysis available to a broader scientific community.

2 HUM CALCULATOR PROGRAM

The HUM Calculator program is realized as a stand-alone software tool in the Visual Studio 2010 environment using the C# programming language. It uses the Microsoft.NET Framework class library (http://msdn.microsoft.com/en-us/library/zw4w595w.aspx) to provide the interface and functions of the HUM Calculator. The software tool allows to perform multicategory ROC analysis, including the possibilities to load the datasets in text format, to select the classifiers (termed 'variables' in the application) and categories ('diagnoses' in the application) and to calculate the corresponding AUC values for two-class problems and HUM values for more than two categories. It is also possible to construct and visualize 2D- and 3D-ROC curves. An advanced feature of the HUM Calculator is the function 'Exhaustive search along diagnoses', that makes it possible to estimate AUC or HUM values for several predictors and for all the combinations of the defined number of categories. The software tool requires a standard Windows operating system (Windows XP, Windows 7, etc.). It was developed and tested on a PC Intel Pentium 4 CPU, 3.00 GHz, 2.00 GB RAM. The HUM Calculator includes an extensive help file, which is accessible from the main menu and describes how the program works. See Supplementary Material for details of installation and use of the HUM Calculator to analyze a simulated dataset. The test datasets are included in the installation package.

As bioinformaticians and statisticians use the R programming language extensively, we have also developed a HUM R package

*To whom correspondence should be addressed.

Table 1. Description of the main functions of the HUM R package

Function	Description
CalculateHUM_seq	Calculate the maximal HUM value and the corresponding permutation of categories
CalculateHUM_Ex	Calculate the HUM values with an exhaustive search of a specified number of categories
CalculateHUM_ROC	Construct and plot 2D- or 3D-ROC curves
CalcGene	Compute the HUM value for one feature
CalcROC	Compute the coordinates to plot 2D- or 3D-ROC curve, optimal thresholds and the classifier accuracies for them
CalculateHUM_Plot	Plot the 2D-ROC curve
Calculate3D	Plot the 3D-ROC curve

Note: The basic unit of the HUM package is the 'CalculateHUM_seq' function. It calculates the AUC in case of two class labels and the HUM if more than two class labels are available for the selected predictors. The function 'CalculateHUM_Ex' is an extension of the main function and enables calculating HUM values for all combinations of a defined number of categories from the whole set of categories. The function 'CalculateHUM_ROC' calculates the coordinates in order to plot 2D- and 3D-ROC curves. The functions 'CalcGene' and 'CalcROC' are the auxiliary functions to perform the calculation. The software package comes with extensive documentation.

and the corresponding Shiny web application (http://CRAN.R-project.org/package=shiny) to provide the possibility to deploy multicategory ROC analysis functions over the web.

3 REALIZATION IN THE R ENVIRONMENT

The software tool in the R programming language includes the HUM R package and the Shiny application, which provides a web interface for accessing the main functions of the R package.

The HUM R package provides a consistent set of functions for computing and visualizing HUM values, for building and plotting 2D- or 3D-ROC curves, calculating optimal threshold points and classifier accuracies for the optimal thresholds (Table 1).

The HUM R package was implemented in R, a free software environment for statistical computing and graphics (http://www.R-project.org/). The auxiliary functions 'CalcGene' and 'CalcROC' of the HUM R package are written in the C++ language and are integrated in R through the Rcpp package (Eddelbuettel and Francois, 2011). These functions improve computational efficiency and shorten the computational times of the main functions.

The web application was developed using the Shiny R package, which facilitates building interactive web applications (Fig. 1). It provides the possibility to load the datasets for the analysis and to access all functions of the HUM R package. The Shiny application features a directory called 'HUM', containing a user-interface definition in the 'ui.R' file and a server script in the 'server.R' file. See Supplementary Material for details on installation and use of HUM R package functions and the main procedures to perform multicategory ROC analysis with the Shiny application.

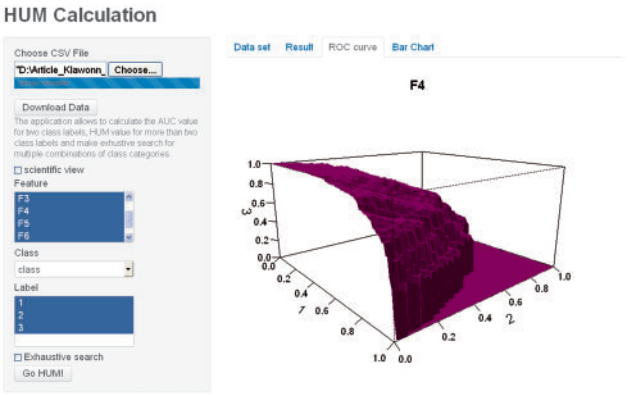

Fig. 1. 3D-ROC curve in Shiny application

We believe that the HUM Calculator program and the HUM R package with the Shiny application will provide researchers, especially in the bioinformatics and biostatistics communities, the necessary tools to better interpret their results in clinical and biomedical classification studies.

The tools are offered free of charge to all users on the website http://public.ostfalia.de/~klawonn/HUM.htm and at http://cran.r-project.org/web/packages/HUM/ (R package only).

ACKNOWLEDGEMENTS

We thank the colleagues of the Bioinformatics and Statistics Group of the Helmholtz Center for Infection Research for their participation in testing the software tools.

Funding: Innovative Medicines Initiative Joint Undertaking (grant agreement number 115523-2–COMBACTE); German Academic Exchange Service (grant number A/13/00004 to N.N.); Helmholtz Association's Cross Programme Initiative in Individualized Medicine (iMed).

Conflict of Interest: none declared.

REFERENCES

Bradley,A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recog.*, **30**, 1145–1159.

Eddelbuettel,D. and Francois,R. (2011) Rcpp: Seamless R and C++. *Integr. J. Stat. Soft.*, **40**, 1–18.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recog. Lett.*, **27**, 861–874.

Hand,D.J. and Till,R.J. (2001) A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.*, **45**, 171–186.

Li,J. and Fine,J.P. (2008) ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, **9**, 566–576.

Li,J. and Zhou,X.H. (2009) Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *J. Stat. Planning Int.*, **139**, 4133–4142.

Robin,X. *et al.* (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform*, **12**, 77.

Sonego,P. *et al.* (2008) ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief Bioinform.*, **9**, 198–209.