CCSU

DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Thursday, April 25
12:15 – 1:15 PM
Maria Sanford Hall, Room 101

APPLYING MISCLASSIFICATION COSTS TO AMELIORATE THE FALSE POSITIVE RATE IN BIOASSAY SCREENING

KAY BATTA

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: This thesis evaluates and extends work done in the Journal of Cheminformatics article "Virtual screening of bioassay data" by Amanda Schierz (2009) and the datasets it provides. Schierz was able to locate several primary datasets in PubChem, a repository of bioassay data, and their corresponding secondary screening results. Bioassay data is very imbalanced in the sense that only a few substances will show an effect on the target protein and, hence, can possibly be developed into medicines. Schierz experimented with several data mining tools in the software package WEKA using misclassification costs to improve results and found that the optimum misclassification cost was dependent on which data mining tool was used. The purpose of this study is to see if there are other factors that also affect the optimum misclassification cost. The C5.0 data mining tool in IBM SPSS Modeler was used and is similar to the algorithm called MetaCost J48 in WEKA. Using Modeler, C5.0, gave much better results than MetaCost J48 and in three to six minutes instead of the 60 minutes taken by the WEKA algorithm. The optimum misclassification cost was very different. For Modeler it occurred at cost 469 with a true positive rate of 92% and a false positive rate of 4.76%. For MetaCost J48 it occurred at cost 1500 with a true positive rate of 54%. The false positive rate was not given for J48 but the best WEKA algorithm (CSC SMO) had a true positive rate of 76.92% and a false positive rate of 11.91%. Experiments within C5.0 showed that how the parameters were set affected the results. For example, changing the pruning severity from the default of 75% to 25% changed the optimum misclassification cost from 42 with a true positive rate of 85% and a false positive rate of 1.11% to an optimum misclassification cost of 469 with a true positive rate of 92% and a false positive rate of 4.76%. Changing the number of predictors also affected the optimum misclassification cost.

For further information:

gotchevi@ccsu.edu 860-832-2839
http://www.math.ccsu.edu/gotchev/colloquium/