A COMBINED DATA MINING APPROACH FOR INFREQUENT EVENTS: ANALYZING HIV MUTATION CHANGES BASED ON TREATMENT HISTORY

Ray S. Lin[1], Soo-Yon Rhee[2], Robert W. Shafer[2], and Amar K. Das1
[1]Stanford Medical Informatics and [2]Division of Infectious Diseases
Department of Medicine, Stanford University
Stanford, CA 94305, United States

Many biological databases contain a large number of variables, among which events of interest may be very infrequent. Using a single data mining method to analyze such databases may not find adequate predictors. The HIV Drug Resistance Database at Stanford University stores sequential HIV-1 genotype-test results on patients taking antiretroviral drugs. We have analyzed the infrequent event of gene mutation changes by combining three data mining methods. We first use association rule analysis to scan through the database and identify potentially interesting mutation patterns with relatively high frequency. Next, we use logistic regression and classification trees to further investigate these patterns by analyzing the relationship between treatment history and mutation changes. Although the AUC measures of the overall prediction is not very high, our approach can effectively identify strong predictors of mutation change and thus focus the analytic efforts of researchers in verifying these results.