HOW RELIABLE IS A MEASURE OF MODEL RELIABILITY? BOOTSTRAP CONFIDENCE INTERVALS OVER VALIDATION RESULTS

Marcin Kozniewski[1,3], Mario A. Cypko[2], Marek J. Druzdzel[1,3]

[1] School of Information Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

[2] The Innovation Center for Computer Assisted Surgery, University of Leipzig, Leipzig, Germany

[3] Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland

Abstract: A researcher testing a model will frequently question the reliability of the test results, understanding well the intuition that verification performed on a handful of cases is less reliable than verification based on very large numbers of cases. Because a limited number of verification cases happens pretty often in very specific domains, a question of practical importance is, thus, how reliable is a reported reliability measure.
We propose a methodology based on deriving confidence intervals over various measures of accuracy of Bayesian network models by means of bootstrap confidence intervals. We evaluate our approach on ROC and calibration curves derived for a model derived from an UC Irvine Machine Learning Repository data set and a sizeable (over 300 variables) practical model constructed using expert knowledge and evaluated on merely 66 accumulated real patient cases. We show how increasing the number of test cases impacts the width of confidence intervals and how this can aid in estimating a reasonable number of verification cases that will increase the confidence in model reliability.

Keywords: Bayesian networks, bootstrap confidence intervals, validation

Advances in Computer Science Research, vol. 13, pp. 27-41, 2016.

JAK WIARYGODNA JEST MIARA OCENY MODELU? BOOTSTRAPOWE PRZEDZIAŁY UFNOŚCI DLA MIAR DOKŁADNOŚCI MODELU

Streszczenie Przy testowaniu modelu należy zdawć sobie z tego sprawę, że weryfikacja modelu przy pomocy małego zbioru danych jest mniej przekonywująca niż weryfikacja bazująca na dużym zbiorze danych. Często napotyka się sytuację, w której do analizy modelu dysponujemy nieznaczną ilością rekordów. Nasuwa się pytanie o wiarygodność oceny modelu.
Proponujemy w takiej sytuacji przyjrzeć się bootrstrapowym przedziałom ufności różnych miar dokładności modelu. W tej pracy określamy bootstrapowe przedziały ufności dla krzywych ROC i krzywych kalibracji modeli uzyskanych z danych z repozytorium UC Irvine. Czynność powtarzamy dla modelu skonstruwanego na podstawie wiedzy ekspertów (ponad 300 zmiennych) i testowanego na 66 zebranych rekordach pacjentów. Pokazujemy jak wzrost liczby rekordów wpływa na szerokość bootstrapowych przedziałów ufności oraz jak taka analiza może pomóc w określeniu liczby rekordów, która może podwyższyć rzetelność weryfikacji modelu.

Słowa kluczowe: sieci bayesowskie, bootstrapowe przedziały ufności, walidacja