Statistical Issues in Machine Learning – Towards Reliable Split Selection and Variable Importance Measures

Carolin Strobl

2. Juli 2008

Die Anwendung von Methoden des Rekursiven Partitionierens aus dem Maschinellen Lernen ist in vielen Forschungsgebieten, wie z.B. in der Genetik und Bioinformatik, weit verbreitet. Der Vortrag setzt sich aus statistischer Sicht mit den zwei Hauptproblemen des Rekursiven Partitionierens, Instabilität und verzerrter Variablenselektion, auseinander. Im Hinblick auf das erste Thema, die Instabilität, wird das gesamte Methodenspektrum von herkömmlichen Klassifikationebäumen über robustifizierte Klassifikationsbäume und Ensemble Methoden wie TWIX, Bagging und Random Forests abgedeckt. Ensemble Methoden erweisen sich als deutlich stabiler als einzelne Klassifikationebäume, verlieren aber auch grösstenteils ihre Interpretierbarkeit. Deshalb wird ein adaptives Bruchpunkt-Selektionskriterium vorgeschlagen, mit dem ein TWIX Ensemble auf einen einzelnen Klassifikationsbaum reduziert wird, falls die Partition stabil genug ist. Im Hinblick auf das zweite Thema, die verzerrte Variablenselektion, werden die statistischen Ursachen für dieses Artefakt in einzelnen Bäumen und eine neue Form von Verzerrung, die in Ensemble Methoden auftritt die auf Bootrap-Stichproben beruhen, untersucht. Aufgrund der Ergebnisse für einzelne Bäume und weiteren Untersuchungsergebnissen zu den Auswirkungen von Bootrap-Stichprobenverfahren auf Assoziationsmasse wird gezeigt dass, neben der Verwendung von unverzerrten Selektionskriterien, Teilstichprobenverfahren anstelle von Bootstrap-Stichprobenverfahren in Ensemble Methoden verwendet werden sollten, um die Variable Importance Werte von Prädiktorvariablen unterschiedlicher Art zuverlässig vergleichen zu können. Die statistischen Eigenschaften und die Nullhypothese eines Test für die Variable Importance von Random Forest werden kritisch untersucht. Abschliessend wird eine neue, bedingte Variable Importance vorgeschlagen, die im Fall von korrelierten Prädiktorvariablen einen fairen Vergleich erlaubt und die interessierende Nullhypothese besser widerspiegelt.