Thilo Mende: **On the Evaluation of Defect Prediction Models**

Promotion: Universität Bremen, Fachbereich 3/Informatik
Erstgutachter: Prof. Dr. Rainer Koschke, Universität Bremen
Zweitgutachter: Prof. Dr. Jan Peleska, Universität Bremen
Datum der Prüfung: 4. November 2011
Veröffentlichung: Verlags-Haus Monsenstein und Vannerdat, Münster, 2012 (ISBN: 978-3869915005).
Eine elektronische Fassung ist auf Anfrage verfügbar: tmende@informatik.uni-bremen.de

Kurzfassung:

Fehlervorhersagemodelle zielen darauf ab, fehleranfällige Module eines Software-Systems automatisch zu identifizieren. Dabei werden Verfahren des statistischen Lernens oder des Data Mining verwendet, um einen Zusammenhang zwischen Software-Metriken und Fehlern auf Basis historischer Daten zu identifizieren. Die daraus resultierenden Modelle können dann auf neue Daten angewendet werden, um Vorhersagen über die Fehleranfälligkeit einzelner Module zu generieren. Diese Informationen sollen dann dazu genutzt werden, qualitätssichernde Maßnahmen, wie zum Beispiel Tests oder Code-Inspektionen, zu steuern und zu optimieren. So könnten die Module, für die eine sehr hohe Fehleranfälligkeit prognostiziert wird, besonders intensiv getestet werden, um bei insgesamt gleichem Aufwand mehr Fehler zu identifizieren.

Aufgrund ihres Potentials für Kosteneinsparungen werden solche Modelle zur Vorhersage von Fehlern seit über 10 Jahren intensiv erforscht, was zu weit mehr als 100 veröffentlichten Forschungsartikeln führte. Darüber hinaus werden Fehlervorhersagemodelle oftmals verwendet, um Software-Metriken und Hypothesen aus dem Bereich der Softwaretechnik empirisch zu validieren. Eine sorgfältige, realistische und reproduzierbare Evaluation ist hier von besonderer Bedeutung.

Trotz der großen Anzahl an veröffentlichen Arbeiten wurde der Evaluation von Fehlervorhersagemodellen bislang nur vergleichsweise wenig Aufmerksamkeit gewidmet. Dies wird durch die große Zahl an nur leicht unterschiedlichen Ansätzen zur Evaluation deutlich, die in der Literatur verwendet werden. Dies betrifft insbesondere die Evaluationsmaße, die zur Bewertung der Vorhersagegüte verwendet werden. Hier werden oftmals Maße für binäre Klassifikationsmodelle, wie zum Beispiel *Recall* und *Precision*, verwendet. Aufgrund einiger besonderer Charakteristika von Software-Metriken und der Verteilung von Fehlern in einem Software-System stellt sich jedoch die Frage, ob diese Maße wirklich geeignet sind, die praktische Anwendbarkeit von Fehlervorhersagemodellen zu bewerten. Die vorliegende Arbeit untersucht daher Vor- und Nachteile verschiedener Evaluationsansätze und beschreibt Richtlinien für eine angemessene und praxisnahe Evaluation von Fehlervorhersagemodellen.

Zunächst werden verschiedene Evaluationsansätze in einer Literaturübersicht über 107 Artikel gesammelt und zusammengefasst. Die am häufigsten verwendeten Verfahren werden dann im Detail untersucht. Dazu werden öffentlich verfügbare Fehlerdatensätze genutzt, und es wird gezeigt, dass die verschiedenen Ansätze Vor- und Nachteile haben und zu unterschiedlichen Ergebnissen führen können. Dies führt zu Problemen in der Vergleichbarkeit verschiedener Arbeiten und der Replizierbarkeit von Experimenten, selbst wenn die eigentlichen Fehlerdaten öffentlich verfügbar sind. Die in der Arbeit vorgestellten Ergebnisse und Richtlinien zur Evaluation können diesen Problemen entgegenwirken.

Darüber hinaus wird gezeigt, dass sehr einfache Modelle, die ausschließlich auf der Größe von Modulen basieren, zu überraschend guter Vorhersagegüte in Bezug auf die klassischen Evaluationsansätze führen können. Der Grund dafür liegt in einer impliziten Annahme fast aller Ansätze, nämlich dass die Aufwände für zusätzliche qualitätssichernde Maßnahmen gleichmäßig über alle Module eines Systems verteilt sind. Da diese Annahme für viele Anwendungsfälle nicht realistisch ist, wird das Konzept der Aufwandssensitivität definiert, welches ungleich verteilte Aufwände betrachtet. Die meisten Fehlervorhersagemodelle, denen ursprünglich eine gute Vorhersagegüte attestiert wurde, sind nicht effektiv, wenn aufwandssensitiv evaluiert wird. Die Vorhersagegüte kann statistisch und praktisch signifikant gesteigert werden, wenn der Aufwand schon während der Erzeugung der Modelle einbezogen wird. Ob die daraus resultierenden Modelle in der Praxis kosteneffektiv eingesetzt werden können, bleibt aber zweifelhaft, wie eine Fallstudie deutlich macht.

Zusammenfassend zeigen die in dieser Arbeit vorgestellten Experimente, dass die meistens verwendeten Evaluationsansätze die praktische Vorhersagegüte überschätzen und nicht zu kosteneffektiven Fehlervorhersagemodellen für viele Anwendungsszenarien führen. Die aus den Experimenten abgeleiteten Richtlinien, und insbesondere das Konzept der aufwandssensitiven Vorhersage und Evaluation, können helfen, Fehlervorhersagemodelle zu generieren, die in der Praxis einsetzbar sind.