

Inhaltsverzeichnis

I. Ausgangssituation	1
1. Einleitung	3
1.1. Softwareklone	3
1.2. Problem	5
1.3. Lösungsansatz	5
1.3.1. Reduktion der Falsch-Positiven mit einem Feedback- Filter	6
1.3.2. Analyse der Bewertungen	7
1.3.3. Weiteres Vorgehen	9
1.4. Eigener Betrag	9
1.5. Konventionen	10
1.6. Gliederung der Arbeit	11
2. Stand der Forschung	13
2.1. Grundlegende Begriffe von Softwareklonen	13
2.2. Softwareklone	15
2.2.1. Klonbegriff	16
2.2.2. Falsch-Positive und Quellen von Redundanz	27
2.2.3. Auswirkungen von Softwareklonen	31
2.2.4. Anwendungen der Klonerkennung	32
2.2.5. Beurteilung von Klonen	41
2.2.6. Einpassung der Klonerkennung in einem Prozess	45
2.3. Klonerkennungsverfahren	47
2.3.1. Verfahren für Listen	49
2.3.2. Verfahren für Bäume	50
2.3.3. Verfahren für Graphen	51
2.3.4. Verfahren für Mengen	52
2.4. Lernverfahren	53
2.4.1. Anforderungen an Lernverfahren	54
2.4.2. Entscheidungsbäume	54
2.4.3. Random-Forest	56

2.4.4.	Support Vector Machine	57
2.4.5.	Weitere Verfahren	57
2.4.6.	Lernverfahren in der Klonerkennung	57
2.5.	Zusammenfassung	58

II. Erkennung mittels token-basierender Metriken 59

3. Der Feedback-Filter 61

3.1.	Bisheriger Umgang mit dem variablen Klonbegriff	61
3.2.	Der Feedback-Filter	62
3.2.1.	Arten des Lernens vom Anwender	63
3.2.2.	Vorgehen mit dem Feedback-Filter	65
3.3.	Modi des Feedback-Filters	65
3.4.	Vorgehen für die Bewertung des Feedback-Filters	66
3.5.	Datensätze	68
3.5.1.	Datensatz von Bellon	69
3.5.2.	Datensatz von Frenzel et al.	69
3.5.3.	Datensatz von Mende et al.	70
3.5.4.	Datensatz von Tiarks et al.	70
3.5.5.	Datensatz von Deissenboeck et al.	71
3.5.6.	Datensatz von Frenzel	71
3.5.7.	Weiterer Datensatz	71
3.5.8.	Datengruppen	74
3.5.9.	Zusammenfassung	74

4. Metriken 77

4.1.	Typen von Metriken	77
4.2.	Benennung von Metriken	79
4.3.	Ähnlichkeitsmaße	79
4.3.1.	Betrachtung als Mengen	81
4.3.2.	Vergleich der Histogramme	82
4.3.3.	Struktur mit N-Grammen	83
4.3.4.	Ähnlichkeit mit Hilfe von Editierkosten	84
4.3.5.	Ähnlichkeit aufgrund des gemeinsamen Informationsgehaltes	89
4.4.	Basismetriken auf dem Tokenstrom	96
4.4.1.	Längenmetriken	96
4.4.2.	Häufigkeiten	96
4.4.3.	Erstes und letztes Tokenvorkommen	98
4.4.4.	N-Gramme	99

4.4.5.	Balanciertheit	99
4.4.6.	Klontyp	100
4.4.7.	Überlappung	102
4.4.8.	Komplexität von Halstead	103
4.5.	Interne Wiederholungen	104
4.5.1.	Berechnung mit dem Suffixbaum	104
4.5.2.	Berechnung mithilfe von Kompression	109
4.5.3.	Anmerkung	110
4.6.	Metriken zur Umgebung des Kandidaten	110
4.6.1.	Vorkommensmetrik	111
4.6.2.	Funktionsgrenzen	111
4.6.3.	Positionsabstandsmaß	113
4.7.	Weitere Datengrundlagen	114
4.7.1.	Layout	114
4.7.2.	Kommentare	116
4.7.3.	Bezeichner	117
4.8.	Nicht umgesetzte Metriken	118
4.9.	Metrikgruppen	119
4.10.	Verarbeitung	120
4.11.	Zusammenführung der Mischdaten und Reduktion	121
4.12.	Zusammenfassung und Übersicht	124
5.	Analyse der individuellen Eignung von Metriken	133
5.1.	Konstante Metriken	133
5.2.	Unterteilungsgüte	135
5.2.1.	Auswertung	138
5.2.2.	Zusammenfassung	145
5.3.	Beziehungen	146
5.3.1.	Korrelation zwischen den Top-5-Metriken	147
5.3.2.	Korrelation über (fast) alle Metriken	149
5.4.	Vergleich mit Grundmenge	151
5.5.	Einfluss verschiedener Längenmaße	153
5.6.	Zusammenfassung	154
6.	Klassifikation von Kandidaten mit Lernverfahren	157
6.1.	Evaluierung von Lernverfahren	157
6.1.1.	Bewertung	158
6.1.2.	Kreuzvalidierung	161
6.1.3.	Späte Reduktion der Metrikmengen	163
6.1.4.	Lernen und Testen in unterschiedlichen Datengruppen	164

6.2.	Random-Forest als Basisverhalten	167
6.2.1.	Vorgehen	167
6.2.2.	Ergebnisse	168
6.3.	Einfluss von verschiedenen Lernverfahren	175
6.3.1.	Vorgehen	175
6.3.2.	Ergebnisse	176
6.4.	Einfluss des Lernmengenumfangs	178
6.4.1.	Vorgehen	178
6.4.2.	Ergebnisse	178
6.5.	Interdatengruppenübertragbarkeit	186
6.5.1.	Vorstellung	186
6.5.2.	Ergebnisse	186
6.6.	Asymmetrische Kosten	192
6.6.1.	Vorstellung	193
6.6.2.	Ergebnisse	194
6.7.	Ableitung konkreter Regeln	195
6.7.1.	Vorgehen	195
6.7.2.	Ergebnisse	196
6.8.	Überprüfendes Experiment	201
6.8.1.	Vorgehen	201
6.8.2.	Ergebnisse	202
6.9.	Zusammenfassung	205

III. Weitere Erkennungsverfahren 207

7. Lernen von Tokensequenzen 209

7.1.	Kandidaten als Sprache	209
7.1.1.	Verwandte Arbeiten	210
7.1.2.	Grammatikinduktion	211
7.2.	Vorgehen	215
7.3.	Evaluierung	216
7.4.	Vergleich mit den bisher untersuchten Lernverfahren	221
7.5.	Zusammenfassung	223

IV. Finale 225

8. Zusammenfassung und Ausblick 227

8.1.	Zusammenfassung	227
------	---------------------------	-----

8.2. Gefahren für Validität	229
8.2.1. Gefahren im Zusammenhang mit den Datensätzen	229
8.2.2. Gefahren bei der Durchführung der Experimente	231
8.2.3. Gefahren bei der Interpretation der Ergebnisse	232
8.3. Ausblick	233
8.3.1. Einsatz von Klonerkennung im Softwareentwick- lungsprozess	233
8.3.2. Blick über den Copy-Paste-Tellerrand	234
8.3.3. Auswirkungen der Klonentfernung modellieren	234
8.3.4. Häufiges Vorkommen als Indiz für Falsch-Positive	235
8.3.5. Bewertung von Klonerkennungswerkzeugen	235
8.3.6. Feedback-Filter und Lernverfahren	237
8.4. Schlusswort	238
A. Tabellen	239
Literaturverzeichnis	273
Glossar	291
Index	295