Nachweis von Feature Freezes durch Clustering

Steffen Herbold

Institut für Informatik
Universität Göttingen

18.11.2008 - Metrikon 2008

Überblick

Grundidee

- Analyse von Metrikdaten
  - Verschiedene Messzeitpunkte
  - Metrikdaten als Eingabe für Lernalgorithmen
- Zwei generelle Möglichkeiten:
  - Retrospektiv
  - Während der Laufzeit
- Ziel: Zeigen, dass eine derartige Analyse möglich ist.

Messung

Definition (Messung nach Fenton/Pfleeger)

Messen ist ein Prozess, bei welchem Zahlen oder Symbole zu Attributen von Objekten aus der realen Welt nach klar definierten Regeln zugewiesen werden.

- Attribut: Körpergröße
- Regel zum Messen: Vermesse die Länge des Körpers in Zentimetern

Softwaremetrik nach IEEE 610.12

Definition

Eine Metrik ist eine quantitative Messung, zu welchem Grad ein System, eine Komponente oder ein Produkt ein gegebenes Attribut besitzt. Eine Qualitätsmetrik ist

1. eine quantitative Messung des Grades zu welchem ein Gegenstand ein gegebenes Qualitätsattribut besitzt.
2. eine Funktion, die als Eingabe Softwaredaten und als Ausgabe einen einzelnen numerischen Wert, welcher als der Grad, zu dem die Software ein gegebenes Qualitätsattribut besitzt interpretiert werden kann.

Prozessmetriken

- Messen den Entwicklungsprozess
- Beispiele:
  - Anzahl der Entwickler
  - Benötigte Zeit
  - Ausgegebenes Geld
  - Anzahl der Fehler (BUG)

Produktmetriken

- Messen ein Produkt direkt
- Komplexitätsmetriken:
  - Zyklomatische Zahl
  - Verschachtelungstiefe
- Größenmetriken:
  - Anzahl der Zeilen (LOC)
  - Anzahl der Statements

Was ist maschinelles Lernen?

Definition (Maschinelles Lernen nach Mitchell)

Man sagt das ein Computerprogramm aus einer Erfahrung E, unter Berücksichtigung einer Klasse von Aufgaben T und einem Maß für sein Leistungsverhalten P, lernt, falls das Leistungsverhalten von Aufgaben T, gemessen von P sich mit der Erfahrung E verbessert.

- In diesem Kontext bedeutet dies:
  - Die Erfahrung E sind Metrikdaten
  - Die Klasse von Aufgaben T sind die Lernalgorithmen, welche zur Analyse der Daten eingesetzt werden.
  - Das Maß für das Leistungsverhalten P ist, wie gut ein Algorithmus eine Eigenschaft bestimmen kann.

Überwacht und Unüberwacht

- Lernalgorithmen können überwacht oder unüberwacht sein.
  - Überwachte Algorithmen kennen die *Klassifikation* der Daten.
  - Man nennt die Daten ebenfalls überwacht oder unüberwacht.
- Die Qualität des Resultats eines Lernalgorithmus hängt von der Qualität der Daten ab.
  - Schlechte Daten führen zu schlechten Resultaten.

Der *k*-means Clusteralgorithmus

- Ein Cluster ist eine Menge von Punkten
  - Punkte stehen in Beziehung zueinander
- Clusteralgorithmen probieren in unüberwachten Daten Cluster zu finden
- Der *k*-means Algorithmus bestimmt *k* Cluster
- Cluster werden durch ihre Mittelpunkte repräsentiert
- Durch Iterieren der folgenden zwei Schritte, werden die Cluster bestimmt:
  1. Weise jeden Punkt dem Cluster zu, dessen Mittelpunkt ihm am nächsten ist.
  2. Ersetze für jedes Cluster den Mittelpunkt durch den Mittelwert (*mean value*) aller Punkte, die zu dem jeweiligen Cluster gehören.

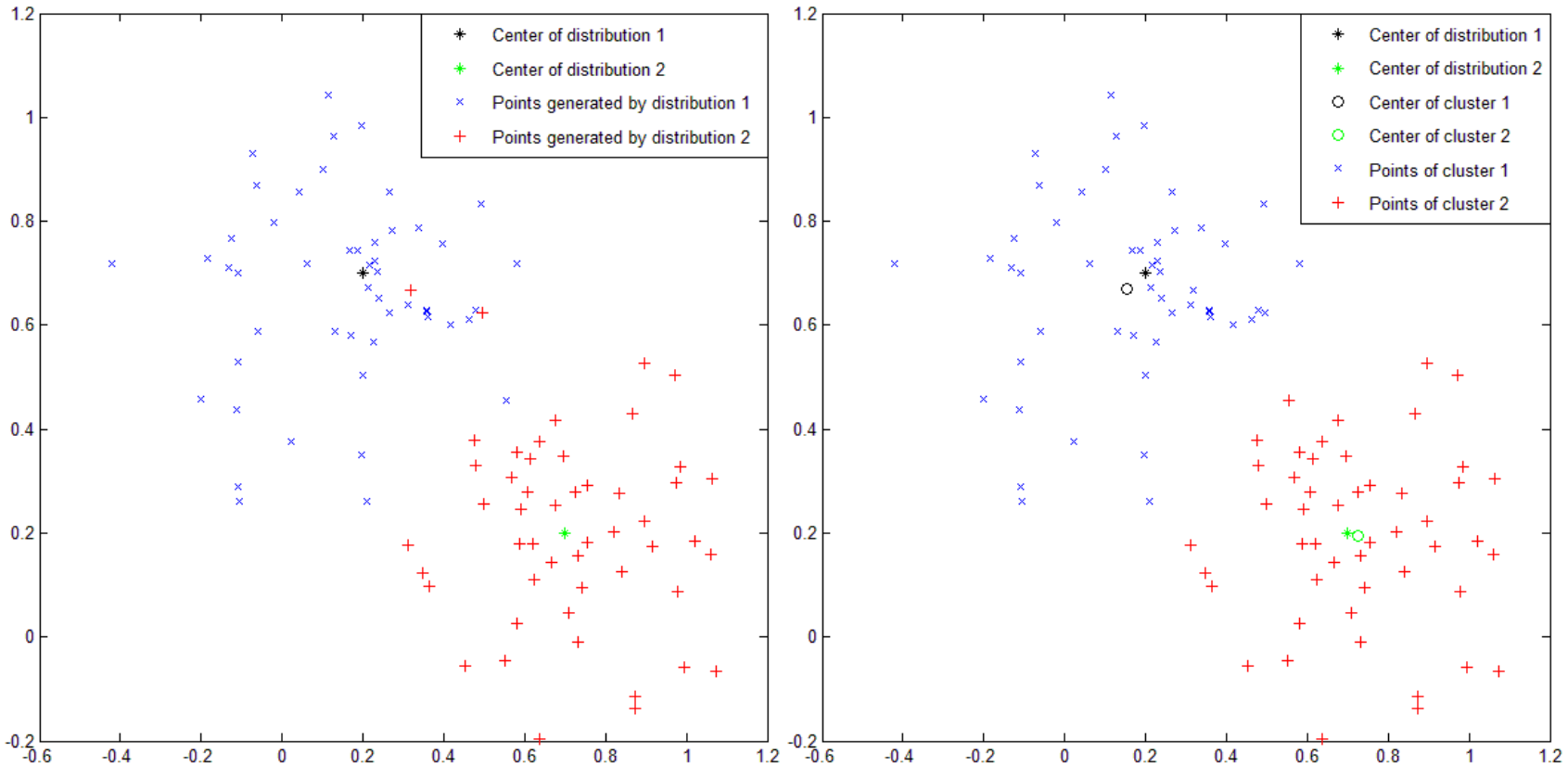


Figure: Links: Simulierte Daten, die durch zwei Normalverteilungen generiert wurden. Rechts: Die Cluster, die der k-means Clusteralgorithmus gefunden hat.

Art der Daten

- Metrikdaten zur automatisierten Analyse benötigt
- Verschiedene Zeitpunkte während der Prozessdurchführung
  - Zeitpunkt nicht zufällig wählen
  - Meilensteine sind mögliche Messpunkte
  - Fortschritt der Metrikdaten nutzen, um den Fortschritt eines Prozesses/Projektes vorherzusagen

Messen in Softwarearchiven

- Direkt am Quelltext messbare Metriken
  - Benötigt ein Quelltext Versionierungsystem
    - Zum Beispiel CVS oder SVN
  - Quelltext kann aus dem Repository bezogen werden
  - Kann durch Tools automatisch gemessen werden
- Auf Fehlerzahlen bezogene Metriken
  - Analyse von Bugtracking Systemen
    - Durch Zugriff auf das Webinterface
    - Durch direkten Zugriff auf die zugrunde liegende Datenbank

Bisher gemessene Daten

- 14 Versionen von 2 Projekten
  - Eclipse Java Development Tools 3.2
  - Eclipse Platform Project 3.2
- Die zugrunde liegende Version 3.1
- Die Meilensteine 1 bis 6
- Die Release Candidates 1 bis 6
- Die Endversion 3.2

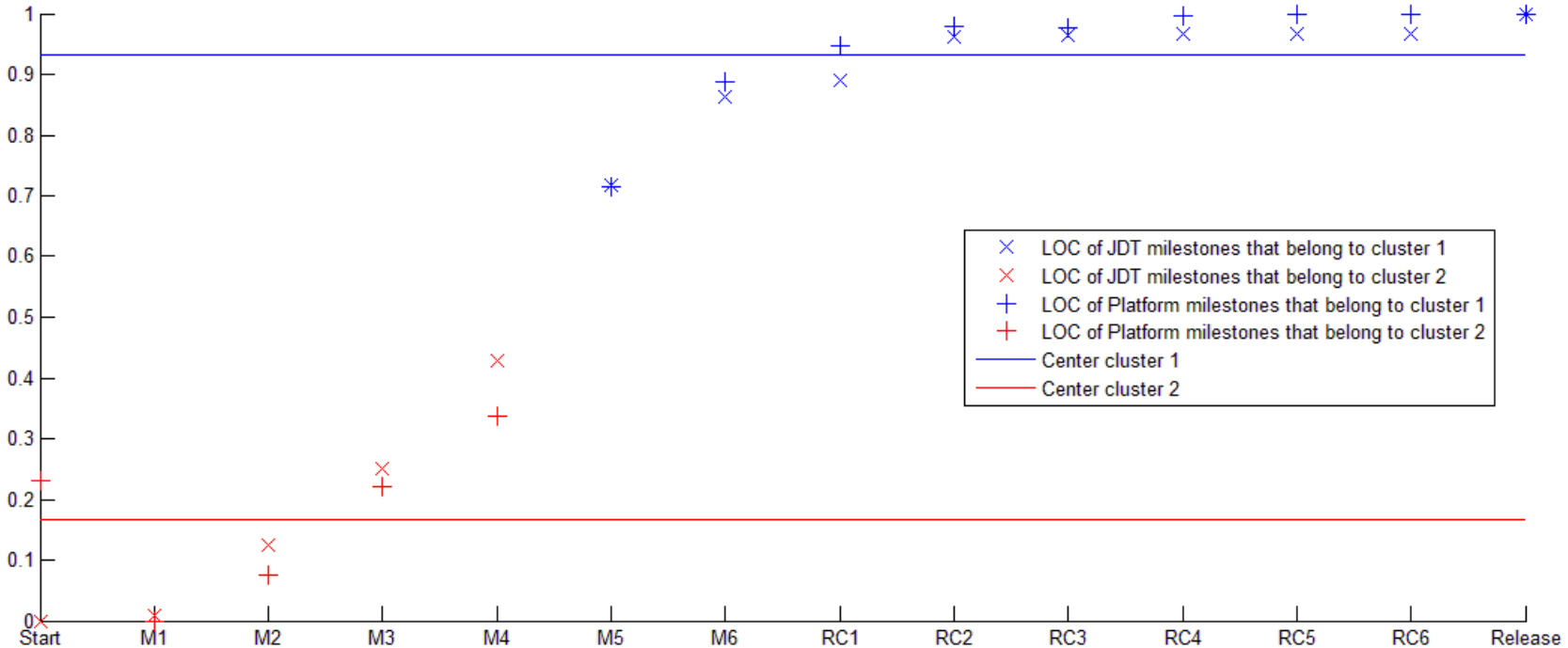
Feature Freezes

- Spezieller Zeitpunkt während eines Projektes
  - Danach werden keine neuen Funktionen mehr hinzugefügt
  - Der Fokus des Projektes verändert sich zu stabilisierenden Aufgaben
    - Testen
    - Fehlerbehebung
    - Dokumentation
- Fokusänderung durch zielgerichtet ausgesuchte Metriken feststellen

Auswahl von Metriken

- *Lines of Code* (LOC) und *Number of Bugs* (BUG) wurden ausgewählt
  - LOC um den Fortschritt des Projektes zu messen
  - BUG um zu Messen, ob sich das Projekt stabilisiert
- Die Werte beider Metriken müssen normalisiert werden

- Normalisierte Metrikwerte als Eingabe für k-means
- Zwei Cluster gesucht
  1. Vor dem Feature Freeze
  2. Nach dem Feature Freeze
- Feature Freeze hat an Meilenstein 5 stattgefunden


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Start
M1
M2
M3
M4
M5
M6
RC1
RC2
RC3
RC4
RC5
RC6
Release
LOC of JDT milestones that belong to cluster 1
LOC of JDT milestones that belong to cluster 2
LOC of Platform milestones that belong to cluster 1
LOC of Platform milestones that belong to cluster 2
Center cluster 1
Center cluster 2

Analyse der Ergebnisse

- Die erfolgreiche Erkennung von Feature Freezes zeigt, dass automatisierte Analyse aufgrund von Metrikdaten möglich ist.
- Der Ansatz ist ist nicht ohne Probleme
  - Es werden immer zwei Cluster gefunden
  - Es nicht nicht immer klar was gefunden wird
  - Nur Anhand von 2 Projekten verifiziert
- Probleme des Ansatzes, nicht ob automatisierte Analyse möglich ist

Ausblick

- Clusterbasierte Ansätze
  - Andere Algorithmen, unterschiedliche Clusteranzahl
- Überwachte Ansätze
  - Klassifikationen oder Zustände lernen
  - Modell Extrapolation aus vorhandenen Daten
    - Kostenschätzung: Nutze aktuelle Metrikwerte um Gesamtkosten Vorauszusagen
    - Kontrollmaßnahmen: Abweichung von vorhergesagten Werten als Indikator für Probleme

Vielen Dank für Ihre Aufmerksamkeit!