Anomalieerkennung und vorausschauende Warnung für z/OS-Systeme anhand von Log-Daten der IBM System Automation

Pavel Kravetskiy
Hochschule Darmstadt, Studiengang Data Science

Problemstellung

Mit der Entwicklung moderner Softwaresysteme werden ihre Überwachung, Verwaltung und Pflege zunehmend schwieriger. Aktuelle Ansätze zur Systemverwaltung beruhen vorwiegend auf Regeln und Richtlinien, die von Experten in diesem Bereich geprägt werden. Aufgrund der steigenden Komplexität und stets ändernden Umfelds verlangt dieser Prozess immer mehr Zeit und ist fehleranfällig. Dementsprechend besteht ein Bedarf an automatischer und effizienter Systemverwaltung, die auf der Analyse von Log-Daten basiert.

IBM System Automation (SA) ist ein Automatisierungsprodukt, das die hohe Verfügbarkeit von z/OS durch Starten, Stoppen und Wiederherstellen der z/OS-Anwendungen bzw. Anwendungsgruppen gewährleistet. Ein z/OS-System ist allerdings hochkomplex, da die Ressourcen nicht als unabhängige Einheiten betrachtet werden können. Eine manuelle Erkennung von Problemen, welche bei dermaßen komplexem System auftreten können, sowie deren mögliche Ursachen stellt einen großen Aufwand dar.

Daten

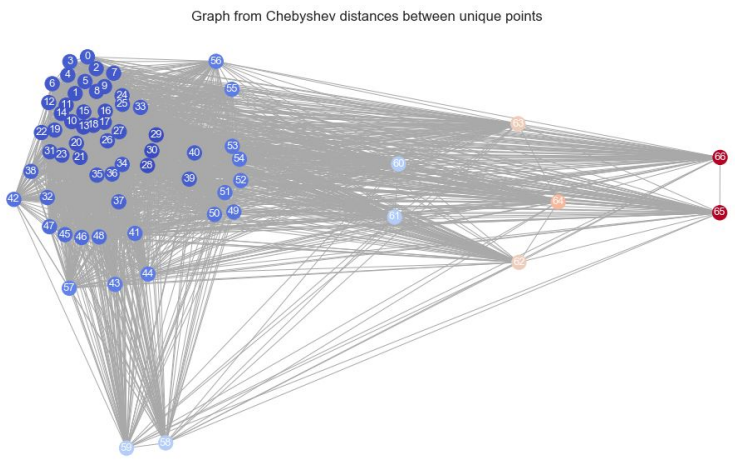
```
2019-06-05 09:59:29 Variable attribute/user/INGWHY_REASON manipulated by operator AUTRGT1
    HSAL6181I Variable attribute/user/INGWHY_REASON set to 00000000; ASIS                    (R#A12LA/APL/AOCA)

2019-06-05 09:59:29 Variable attribute/user/INGWHY_SITUATION manipulated by operator AUTRGT1
    HSAL6181I Variable attribute/user/INGWHY_SITUATION set to 00000000; SOFTDOWN             (R#A12LA/APL/AOCA)

2019-06-05 09:59:28 Termination processing for R#ABTST/APL/AOCC completed - ABENDED, CRITICAL THRESHOLD EXCEEDED
    HSAL6269I Status/Automation is Idle                                                      (R#ABTST/APL/AOCC)
    HSAL6348I Group Observer Update Requested                                                (R#ABTST/APL/AOCC)
    HSAL6427I Group requires evaluation                                                        (AOCC/SYG/AOCC)
    HSAL6427I Group requires evaluation                                                        (SYSPLEX/GRP)
    HSAL6276I Status/Compound is Problem                                                     (R#ABTST/APL/AOCC)
    HSAL6477I Resource Seen Active cleared                                                   (R#ABTST/APL/AOCC)
    HSAL6172I Group Observer update sent                                                     (R#ABTST/APL/AOCC)
    HSAL6172I Group Observer update sent                                                       (AOCC/SYG/AOCC)

2019-06-05 09:59:27 Agent status for R#ABTST/APL/AOCC = BROKEN
    HSAL6259I Status/Observed is HardDown                                                    (R#ABTST/APL/AOCC)
    HSAL6271I Status/Automation is Busy                                                      (R#ABTST/APL/AOCC)
    HSAL6348I Group Observer Update Requested                                                (R#ABTST/APL/AOCC)
    HSAL6119I Resource is not startable                                                      (R#ABTST/APL/AOCC)
    HSAL6337I Resource is not Viable                                                         (R#ABTST/APL/AOCC)
    HSAL6253I Status/Startable is No                                                         (R#ABTST/APL/AOCC)
    HSAL6135I Resource prestart cannot be run                                                (R#ABTST/APL/AOCC)
    HSAL6127I Resource cannot be started                                                     (R#ABTST/APL/AOCC)
    HSAL6427I Group requires evaluation                                                        (AOCC/SYG/AOCC)
    HSAL6427I Group requires evaluation                                                        (SYSPLEX/GRP)
    HSAL6172I Group Observer update sent                                                     (R#ABTST/APL/AOCC)
    HSAL6172I Group Observer update sent                                                       (AOCC/SYG/AOCC)
    HSAL6264I Status/Observed is Problem                                                       (AOCC/SYG/AOCC)
    HSAL6348I Group Observer Update Requested                                                  (AOCC/SYG/AOCC)
    HSAL6252I Status/Startable is Yes                                                          (AOCC/SYG/AOCC)
```

Die SA-Log-Daten eines z/OS-Systems weisen eine deutlich hierarchische Struktur auf, als gewöhnliche Log-Daten. Ein Eintrag, der mit einem Zeitstempel beginnt, stellt ein sogenanntes Workitem dar. Jedes Workitem bezieht sich auf eine bestimmte Aktion, die durch das Workitem-Header definiert ist und dem Workitemtyp entspricht. Workitems beinhalten zeitlich geordnete Sequenzen von verschiedenen Nachrichten, beginnend mit HSAL. Jede HSAL-Nachricht informiert über eine Aktivität oder Statusänderung einer Anwendung bzw. Anwendungsgruppe. Es handelt sich hierbei um ungelabelten Daten.

Anomalieerkennungsmodelle

Im Rahmen dieser Arbeit wurden drei verschiedene Modelle für die Anomalieerkennung eingesetzt und miteinander verglichen:

1. Metrisches Modell



Metrisches Modell basiert auf der Annahme, dass die Anomalien entfernt von der Gesamtstreuung der Daten liegen. Je weiter ein Datenpunkt entfernt liegt, umso anomaler wird er von dem Modell gekennzeichnet. Für dieses Modell wurde das Feature Engineering auf den Log-Daten durchgeführt, damit jeder Workitem durch einen Feature-Vektor fester Länge repräsentiert wird.
Die ausgerechneten Entfernungen wurden mithilfe der gaußschen Skalierung auf ein [0,1]-Intervall normiert und danach als Pseudo-Wahrscheinlichkeit benutzt.

2. Analytisches Modell

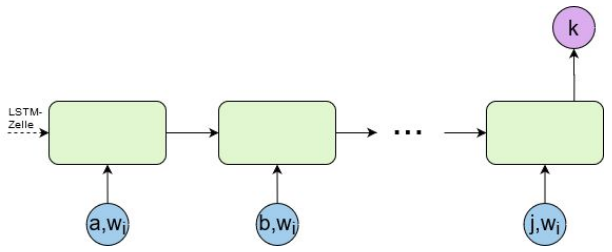
Analytisches Modell nutzt die HSAL-Sequenzen und die Bayes'sche Statistik, um die Wahrscheinlichkeit einer Nachricht anhand ihrer Vorgänger auszurechnen. Das Ausrechnen aller Wahrscheinlichkeitsverteilungen für die bestehende Datenmenge ist rechenintensiv und benötigt viel Speicherplatz. Um das handhabbar zu machen, wurde eine Näherung getroffen, konditionale Sequenzen ab einer bestimmten Länge abzuschneiden. Das Problem wird in Form eines Nachrichtenübergangs $a \rightarrow b$ oder $a \rightarrow \overline{b}$ betrachtet, so dass man zur Binomialverteilung kommt. Um die Wahrscheinlichkeitsverteilungen auszurechnen, wird die Beta-Verteilung $B(\alpha, \beta)$ verwendet, da sie die konjugierte A-priori-Verteilung (conjugate prior) der Binomialverteilung ist. Der bedingte Erwartungswert der Posteriori-Verteilung (Übergangswahrscheinlichkeit) mit Berücksichtigung des Workitemtyps w_i für eine Sequenz $a \rightarrow b \rightarrow c$ sieht wie folgt aus:

$$E(c \mid [a,\ b],\ w_i) = \frac{\alpha + k}{\alpha + \beta + N}$$

wobei k und N die Anzahl der jeweiligen Nachrichten-Sequenzen $a \rightarrow b \rightarrow c$ und $a \rightarrow b \rightarrow X$ sind. α und β sind Parameter der Beta-Verteilung.

3. LSTM-Modell

Da die HSAL-Nachrichten sequentiell aufgebaut sind und gewisse Abhängigkeiten von ihren Vorgängern aufweisen, eignet sich für das Problem ein LSTM-Modell, welches auch zeitlich rückwirkende Beziehungen berücksichtigt.



Mithilfe dieses Modells wird die Wahrscheinlichkeitsverteilung der nächsten Nachricht in Abhängigkeit von ihren Vorgängern und mit Berücksichtigung des Workitemtyps durch eine numerische Funktion approximiert.

Ergebnisse

Im Rahmen dieser Arbeit wurde eine Softwarelösung für unüberwachte Anomalieerkennung anhand von Log-Daten der IBM System Automation entwickelt. Als erster Schritt wurden die charakteristischen Merkmale aus den Log-Daten in Form von Feature-Vektoren und Nachrichten-Sequenzen dargestellt. Diese Darstellungen wurden im zweiten Schritt mithilfe der existierenden Methoden des Data-Minings analysiert, um Indikatoren verschiedener Anomalien im Log aufzuzeigen. Abschließend wurden die verwendeten Methoden miteinander verglichen.

Metrisches Modell konnte zwar nicht alle Anomalietypen erkennen, wies allerdings eine gute Interpretierbarkeit auf, welche für einen Experten eine wesentliche Rolle spielen kann. Außerdem muss die Anomalität eines Workitems nicht extra berechnet werden.

Falls ausreichend Daten zur Verfügung stehen und keinen großen Wert auf die Interpretierbarkeit gelegt werden würde, sollte das LSTM-Modell für die Anomalieerkennung in Betrieb genommen werden, weil es sich als vielversprechend und mächtig erwiesen hat und weil das Auftreten vollkommen fremder Sequenzen in diesem Anwendungsszenario sehr unwahrscheinlich ist.

Ausblick

In einem nächsten Schritt sollte eine für die Validierung ausreichende Menge echter gelabelten Daten, die auch tatsächliche Anomalien enthalten, durch Zusammenarbeit mit Kunden erhoben werden. Der gelabelte Datensatz sollte nicht nur das Finetuning der in dieser Arbeit umgesetzten Modelle erleichtern, sondern auch die Umsetzung anderer (beobachteten) Anomalieerkennungsmodelle ermöglichen. Falls man einen Zugriff zum laufenden z/OS-System hätte, könnte man beim Feature Engineering Kontextinformationen wie Anwendungsgruppentype und ihre Abhängigkeiten im Ressourcenbaum für die Verbesserung der Anomalieerkennung verwenden.