INSTITUTIONEN FÖR REGLERTEKNIK | LUNDS TEKNISKA HÖGSKOLA | PRESENTERAD 2022-06-20

EXAMENSARBETE Log Anomaly Detection of Structured Logs in a Distributed Cloud System
STUDENTER David Nilsson, Albin Olsson
HANDLEDARE Johan Eker (LTH)
EXAMINATOR Karl-Erik Årzén (LTH)

Anomalidetektion av loggmeddelanden med hjälp av maskininlärning för ett molnbaserat system

POPULÄRVETENSKAPLIG SAMMANFATTNING **David Nilsson, Albin Olsson**

Anomalidetektion är något som blir allt vanligare för att utvärdera ett systems hälsa samt skydda mot angrepp. I detta arbete jämför vi olika metoder för anomalidetektion gentemot varandra samt mot ett existerande regelbaserat felhanteringssystem.

I takt med att datorsystem växer och blir mer komplexa, så ökar även svårigheten att manuellt analysera och hitta möjliga säkerhetsbrister hos systemet. Genom att analysera loggmeddelanden med hjälp av maskininlärningsbaserad anomalidetektion kan hårdkodade säkerhetssystem nu göras både mer flexibla och omfattande. Detta för att man inte längre behöver specificera alla möjliga fel som ska detekteras. Loggmeddelanden är generellt textstycken som beskriver vad som händer i ett datorsystem.

I detta examensarbete undersöker vi hur anomalidetektion kan implementeras och vilken anomalidetektionsmetod som presterar bäst för ett specifikt molnbaserat system. Vi jämför även dessa metoder mot ett existerande felhanteringssystem som vi försöker förbättra. Det existerande felhanteringssystemet är regelbaserat och kräver därför regelbundna uppdateringar. Det täcker även få av de fel som kan uppstå. De tre metoderna som jämförs är *Clustering*, *Principal Component Analysis (PCA)*, samt *Autoencoder*. Metoderna jämförs genom att vi injicerar olika typer av anomalier i träningsdatan för att se hur bra de olika metoderna är på att hitta dessa.

I våra experiment ser vi en markant förbättring över det existerande systemet för alla tre anomalidetektionsmetoder. De bästa uppmätta resultaten för varje metod kan ses i tabellen nedan. Anomalidetektionsmetoderna presterar särskilt bra då data från samma tidsperiod används för både träning och testning. Clustering presterar även bra då data från olika tidsperioder används. Både PCA och Autoencoder tycker att många vanliga loggsekvenser är anomalier då olika tidsperioder analyseras, vilket inte clustering-metoden har lika stora bekymmer med. Vi har även sett att Clustering är känslig för anomalier i träningsdatan. Med dessa resultat i åtanke kan vi dra slutsatsen att anomalidetektion är ett bättre alternativ än det nuvarande feldetektionssystemet.

Metod	Prestanda (F1-Score)
Regelbaserat system	0.450
Clustering	0.972
PCA	0.990
Autoencoder	0.953