Machine learning approach to support ticket forecasting from software logs

Matti Haukilintu

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.
!!!Espoo !!!chk.dt.2021

Supervisor

Prof. Arto Visala

Advisor

MSc Petri Pyöriä

Aalto University
School of Electrical
Engineering

Aalto University
School of Electrical Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Matti Haukilintu

Title Machine learning approach to support ticket forecasting from software logs

Degree programme Automation and electrical engineering

Major Control, Robotics and Autonomous Systems **Code of major** ELEC3025

Supervisor Prof. Arto Visala

Advisor MSc Petri Pyöriä

Date !!!chk.dt.2021 **Number of pages** 16+2 **Language** English

Abstract

!!!Your abstract in English. Keep the abstract short. The abstract explains your research topic, the methods you have used, and the results you obtained. In the PDF/A format of this thesis, in addition to the abstract page, the abstract text is written into the pdf file's metadata. Write here the text that goes into the metadata. The metadata cannot contain special characters, linebreak or paragraph break characters, so these must not be used here. If your abstract does not contain special characters and it does not require paragraphs, you may take advantage of the abstracttext macro (see the comment below). Otherwise, the metadata abstract text must be identical to the text on the abstract page.

Keywords Machine learning algorithms, Robotic process automation, log parsing, random delay, Microsoft Azure ML Studio

Aalto University
School of Electrical Engineering

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Matti Haukilintu

Työn nimi Sovelluslokien ja vikatikettien yhteyden löytäminen koneoppimista hyödyntäen

Koulutusohjelma Automaatio- ja sähkötekniikka

Pääaine Ohjaus, robotiikka ja autonomiset järjestelmät **Pääaineen koodi** ELEC3025

Työn valvoja Prof. Arto Visala

Työn ohjaaja FM Petri Pyöriä

Päivämäärä !!!chk.dt.2021 **Sivumäärä** 16+2 **Kieli** Englanti

Tiivistelmä

Tiivistelmässä on lyhyt selvitys kirjoituksen tärkeimmästä sisällöstä: mitä ja miten on tutkittu, sekä mitä tuloksia on saatu.

Avainsanat Koneoppiminen, koneoppimisalgoritmit, ohjelmistorobotiikka, loki, lokin parsiminen, satunnainen viive, satunnaisviive, Microsoft Azure ML Studio

Preface

Pitkäksi venynyt työ
Kiitokset

Otaniemi, March 16, 2022

Matti Haukilintu

Contents

Symbols and abbreviations

Symbols

$\mathbf{B}$	!!!magnetic flux density
$\mathbf{B}$	!!!magnetic flux density

Operators

$\nabla \times \mathbf{A}$	!!! curl of vectorin $\mathbf{A}$
$+$	!!! yep, a plus

Abbreviations

RPA	Robotic Process Automation
ML	Machine Learning
SQL	Structured Query Language
JSON	JavaScript Object Notation
CSV	Comma-Separated Values

1 Introduction

Artificial intelligence has found its way to more and more fields of business. [!!S] In banking business it is already used in [!!!...]

<What does Samlink do?>

<Samlink, small banking business and machine learning potential>

<RPA in use, but not intelligent processes>

<ML and Azure>

Samlink Oy (Samlink from now on) was founded in 1994 [1] some text [2] <some basic info about Samlink>
<founded year and such>
<what kind of development>
<RPA and such>

1.1 Background and motivation

In the field of information technology logging is one of the most important methods in problem solving, be it software or operating system related. [**?**] <importance of logs>

<current state of samlink AI processes>

<Log format>

<Intelligent automation>

1.2 Research objectives

This research aims to kickstart the machine learning application usage in Samlink operations. Multiple obstacles need to be tackled as most of the phases in this study has not yet been encountered inside the company.

First and foremost, it is crucial to construct some basic form for log data so it is usable by machine learning algorithms. Log data forming is on of the key elements in automatic log parsing applications as it is not for just machines but also for people to read.

As today more and more concern is set on anonymization [!!!...] <Anonymization scope>

<Azure ML studio setup and prerequisites>

<Log data and timestamp combining>

<Connection for data, ML estimates>

1.3 Scope

In order to limit the study to feasible length and content it is necessary to define the scope for the thesis. As major part of the time used for the study was used for data acquiring and anonymizing yet the most value comes from content considering

machine learning, the length of the content per section does not reflect the amount of time used for each section.

Data anonymization

Anonymization refers to <!!! open up term>
In this study we aim to create a dataset usable in ML training. In this respect anonymization is not in the main focus of the study but only treated as a sub-phase of the data preprocessing in whole. Nevertheless, anonymization is from the security point of view the most important phase of data preprocessing.
In this study we <!!!how widely we consider anonymizing?>

Azure setup

The ML training and result scoring is done in Azure ML environment. As part of the study aims to create an initial guideline for ML process commissioning the Azure setup phase is documented in such detail reflecting the importance of this information for future developers starting Azure ML projects.
<!!! what more?>

Data requirements

In ML point of view the data purity in a sense of <!!! what is data purity?>

Machine learning methods

Several different machine learning algorithms exist <!!! source> that are aimed for different applications in mind. For example, to make an algorithm that can predict the price of an apartment listed[**?**] we could be using linear regression and in order to detect possible cyber threats from network traffic[**?**] a two-class support vector machine could be utilized. These two methods are very different in usage and has their pros and cons in different applications.
As different methods can be used in creative ways in very different applications depending on how the data is presented and how the ML problem is formed, this study focuses on just a few easily approachable training methods.

1.4 Structure

<how the thesis is organized>

2 Background

2.1 Cloud services

<Azure ML and other cloud services>

2.2 Data protection

<Anonymization and data sensitivity>

2.3 Machine learning methods

<Machine learning field and studies>
<Log data parsing with ML>
<random delay in timeseries(?)>

3 Research material and methods

Tässä osassa kuvataan käytetty tutkimusaineisto ja tutkimuksen metodologiset valinnat, sekä kerrotaan tutkimuksen toteutustapa ja käytetyt menetelmät.

3.1 Data

The data in the research consists mainly of two parts. The most important part is, obviously, the log data produced by the numerous RPA processes. The second part complementing the study is the support ticket data written by clerks of customer banks.

RPA log data

<what form was the data in?>
<how it was acquired?>

Support ticket data

Samlink

3.2 Data anonymization

Samlink handles highly sensitive banking customer data in its processes, such as personal identification numbers, home addresses, email addresses and bank account numbers. All possibly sensitive data must be removed before data can be transferred out from production environment. To improve security, it was decided to assume that personal customer details are not critical information while training the ML algorithm. This way it was not necessary to achieve adequate security by less secure and more effort consuming ways like pseudonymisation or k-anonymization, which would have also required strict inspections before data could have been approved for cloud processing.
As production environment is built on Microsoft Server based solution, and because it was highly unrecommended to install additional software to the production server, data acquiring and anonymization tools were chosen based on what was already usable in the RPA production environment. Microsoft Powershell offers sufficient tools for database SQL querying and stream editing. The amount of data was significant which made straight file editing impossible due to the memory limitations. Thus, stream editing was necessary for finding and replacing sensitive information from the data <!!! add some references>

3.3 Azure ML Studio

Machine Learning methods and components

Several different ML methods are usable with Azure ML Studio. <???>
Usually <with usual ml methods> the estimates created using ML algorithms are

formed based on the certain features presented on a one element of the data, or on one row. This means that in typical case, there is one column in the data given to the ML algorithm that is removed from the training data and this column value is what algorithm aims to predict.
In this study case, however, data does not contain clear values that are being estimated and that can be used as comparison.

LOG_DATA				EFECTE_DATA
a=date	b=msg	c=etc.		A YYYY.MM.DD hh:mm:ss
a	b	c	n1	B YYYY.MM.DD hh:mm:ss
a	b	c	n2	C YYYY.MM.DD hh:mm:ss
a	b	c	n3	D YYYY.MM.DD hh:mm:ss
a	b	c	n4	E YYYY.MM.DD hh:mm:ss

n1 = SUM(AB)
n2 = SUM(C)
n3 = SUM(DE)
=>
We could try to predict nx but usually this is done by making estimate based on a, b and c. Instead, we aim to estimate the sum of events in timeframe. We should also skip event instances that are close to each other to avoid counting multiple values linked to same error as different possible ticket creators.

**** ????
Is it possible to use full row as a feature?
5000 rows / week
count events / week
=> estimate from data

preformatting data!

1. remove fingerprint etc unique values from raw message
2. calculate anomaly probability per line

   ! CAN WE COMBINE THIS PER JOB-ID?
3. create new table consisting:
   (a) amount of rows per timeframe
   (b) anomaly probability value (median, mean etc)
   (c) efecte tickets received in said timeframe

???? ****

4 Results

4.1 Anonymization

Anonymization took good proportion of the time in workdays as processes were slow, amount of data was big and multiple re-runs were needed before the results was seemed adequate.

[!!! appendix of the script used]

4.2 Data formatting

At the beginning of the research, the log data from RPA was in SQL database. However, the database used was not »pure» in a way that typical relational databases are, but some columns included JSON-formatted data in them. For ML algorithms to be able to read the given data with ease this sort of »impurities» needed to be cleared from the data.

4.3 Azure and Azure ML Studio

<general about ml inside azure>

Azure resources

<what resources was needed inside Azure?>

<virtual machines etc.>

Azure ML Studio components

<clusters and data>

<Memory problems>

During the initial pipeline runs the execution came to an abrupt stop and Azure notified about memory issues. These problems were linked to the data amount which had to be reduced to 600 megabytes before any pipeline could be finished using the data. This reduction was against the initial goal where preferably all the data could have been used.

Considerable amount of time was used to fix or avoid this issue but nothing clear was found that would explain the error received. While working with with the issue it was also noted that data needed more cleaning in order to ease the preprocessing phase as described with more detail in section <!!!4.2> Thus, the data had to be imported from log archive and anonymized once more.

Two choices was possible to take:

1. Continue working with full data and attempting to fix the memory issue by consulting Azure experts

2. Trim the data to reduce the data size by declaring info-type log messages as unnecessary and working with vastly diminished data until the memory issue would be solved one way or another

To advance the study more efficiently it was decided to trim info-type log messages from data hence reducing the data amount considerably. Meanwhile, <fixing the memory issue>

4.4 Machine learning methods and pipelines

<anomaly detection>
<N-Gram Feature extracting>
??? <Poisson regression (predicts event counts)>
<two-class classification> <support vector machine etc>

<Integrating with timestamps>

```
<todo:

***A***
count sum of incidents in timeframe x
set x to each row in data by timestamp
predict amount of incidents based on data

***B***
Use efecte data as reference values
(regression, predict amount in timeframe -> compare)
(classification, count TRUE in timeframe -> compare amount)

>
```

5 Summary

<Sum up here what we did and why>

5.1 Discussion

<Here some thinking what should have been improved>

Data formatting

The most time consuming tasks in the study was without a doubt the anonymization and preformatting of the data. Although sensitive information may sometimes be crucial in error fixing as problems may consider just one client, it is necessary that the data sanitation is possible to do in order to use the data in less secure environment. By preformatting the data in such way that all different personal information types do not differ between use cases. <!!! hetu in weird form>

References

[1] Winky K.O. Ho, Bo-Sin Tang and Siu Wai Wong *Predicting property prices with machine learning algorithms* Journal of Property Research, 38:1, 48-70, DOI: 10.1080/09599916.2020.1832558

[2] K. Ghanem, F. J. Aparicio-Navarro, K. G. Kyriakopoulos, S. Lambotharan and J. A. Chambers *Support Vector Machine for Network Intrusion and Cyber-Attack Detection* 2017 Sensor Signal Processing for Defence Conference (SSPD), 2017, pp. 1-5, doi: 10.1109/SSPD.2017.8233268.

A Esimerkki liitteestä