Anomaly Detection in Digital Signature

Name, university, student number, contact info *YunkunNiu, 244910, [yunkun.niu@student.tut.fi](mailto:yunkun.niu@student.tut.fi)*

*Bilal waris, 245004, [bilal.waris@student.tut.fi](mailto:bilal.waris@student.tut.fi)*

*Qi Tang, 239064, [qi.tang@student.tut.fi](mailto:qi.tang@student.tut.fi)*

*Yibin Qian, 232807, [yibinqian@hotmail.com](mailto:yibinqian@hotmail.com)*

Project partner *Intel*

[*kalle.koivisto@intel.com*](mailto:kalle.koivisto@intel.com)*,* [*rauno.tamminen@intel.com*](mailto:rauno.tamminen@intel.com)

Responsible teacher *Sari Peltonen,* [*sari.peltonen@tut.fi*](mailto:sari.peltonen@tut.fi)

*Tampere University of Technology*

Facilitator *AinoSiippainen, [aino@demola.net](mailto:aino@demola.net),*

Project report version *Final*

(draft/final/update1/update2/…)

Date *23/06/2015*

# Table of success

[1 Table of success 1](#_Toc422832986)

[2 Introduction 1](#_Toc422832987)

[3 Defining the problem 2](#_Toc422832988)

[4 Solution 2](#_Toc422832989)

[5 Usefulness of the solution, efficiency and usability 2](#_Toc422832990)

[6 Project management and Workflow 3](#_Toc422832991)

[7 Development ideas and suggestions for execution 4](#_Toc422832992)

[8 Learning experiences 4](#_Toc422832993)

[Appendix 5](#_Toc422832994)

# Introduction

**Who is the project partner?**

Intel: Rauno Tamminen, Kalle Koivisto

**Who are the team members?**

Yunkun Niu, Qi Tang, Bilal Waris, Yibin Qian

**Tell briefly about the background of the project.**

The logic of the detection system is quite analogue to different recognition methods in signal

processing (e.g. face and speech recognition). There are also some studies of anomaly detections

in software science e.g. URI screening using N‐grams (Mikhail Zolotukhin, Timo Hämäläinen, Antti Juvonen: Online anomaly detection by using N‐gram model and growing hierarchical self‐organizing maps. IWCMC 2012: 47--‐52).

Signing systems play big roles in companies’ core security. Therefore controlling and monitoring the input data, especially in more complex infrastructures, is one way to improve security.

# Defining the problem

**What was the need you were solving?**

The agenda is to build and demonstrate a system that can learn to detect anomalies in input

signing request data. The detection is done by data pattern recognition and preferably happens

in real-time. The system is taught by feeding it with big number of different sorts of valid signing

request data e.g. from existing logs

**What parties have an interest in this problem?**

* Open source community
* Software companies with high security

# Solution

**What was your approach to the problem?**

Use Machine Learning Algorithm to detect anomalies automatically

**What was your solution?**

At first, we examine different algorithms and choose the best one as the final solution. But later, we find we have to design a new customized algorithm just for the data of the project. And that is exact what we provide as the final solution in this project.

**List things that the solution consists of or that describe it:**

* SOM (Self Organizing Map) Algorithm
* GMM (Gaussian Mixture Model) Algorithm
* RP (Relation Probability) Algorithm

# Usefulness of the solution, efficiency and usability

**Analyze your solution. Have you reached your goals defined in the problem section or have they changed?**

The process of anomaly detection can be divided into two phases, training and detection.

Training phase’s time cost could be long and depended on the size of training data set. But this

phase is one-off. It is executed only once for the whole life‐span of the system.

Detection phase is real‐time. Because of the key‐to‐probability hash-table, it costs O(1) time to detect a new request. The real time cost for detecting a new request is less than 1ms.

**What is the bigger picture? What parts can you utilize from the solution later on?**

The algorithm designed by my team is suitable for not only the specific project but also similar data which contains low percent of anomalies. In the future, it may be applied in the detection of products from other companies with high security. Such as software licences.

# Project management and Workflow

**How was the communication between the project members, project partner, teachers and Demola?**

We always had meetings with project partner in Intel office after every two weeks and also we were very active on email communication for any kind of conversations

**How did you work, how would you improve your methods in a future project? What would you do differently?**

We worked independently and team as well, most of time we worked independently at home but we also used to meet in university whenever we faced any difficulties in working alone

**How many working hours have the individual team members spent on the project and how much is it altogether?**

|  |  |  |
| --- | --- | --- |
| NAME | NUMBER | HOURS |
| Yunkun Niu | 244910 | 320 |
| Qi Tang | 239064 | 260 |
| Yibin Qian | 232807 | 200 |
| Bilal Waris | 245004 | 200 |

In total, we spent 980 hours.

**What was your accomplished schedule of the project, what has been done (documents, pitches, workshops, etc.)?**

* Playbook
* detection system with SOM algorithm
* mid-pitching presentation
* detection system with GMM algorithm
* detection system with RP algorithm
* final-pitching presentation
* git repository
* technical report
* final report

**Risk management: did you have problems/issues during your work and how did you overcome them?**

This is a research project. The result maybe just analysis of failures rather than a successful solution. This kind of result is disappointing. So we have to learn new knowledge as fast as we can and reanalyze the project and the data again and again.

**Include here your personal work amounts and responsibilities. Be concrete.**

**Yunkun Niu – Manager and developer**

First, I am the manager of our team. I am responsible for the final result of the project. I organize every meeting or communication with Intel and within my team. I assign tasks to each member and encourage everyone to find new methods or think new ideas when the project stagnates. I join every part of the project, developing, communicating, pitching and reporting.

Second, I am also a developer of our team. I analyse the requirements and data of the project. I need to learn machine learning algorithms fast so that I can catch up with my companies. I join every analysis and coding of every algorithm we examine.

**Qi Tang – developer**

My main responsibility in this project is coding. In order to do a better job, I have to learn many machine learning algorithms and try to implement them in Python. Considering the implementation and debug time, my work time is nearly 260 hours.

**Bilal Waris – developer**

I have been active in all parts of the project, because my purpose of joining this project was to learn and not to be concise in specific part of the project so for that I took part in coding, presenting (marketing our project in demola final pitching event), documenting, communicating and attending meetings.

**Yibin Qian – developer**

I help prepare materials for mid-pitching, and gave the speech for Mid-pitching, also I did some documentation work, and test our program.

# Development ideas and suggestions for execution

**How would you develop your solution, working methods, and communication with your project partner, teacher and Demola?**

In this project, we did not communicate much with our teacher. Because we want to challenge ourselves and check if we can finish this project independently as a team. But in the end, we found we also missed the opportunity to learn knowledge from our teacher. On the final meeting, our teacher, Sari, mentioned if we ever considered publishing our algorithm. That was a so great idea which we never thought about. However, if we did communication with Sari during the project, there would be definitely a better result. Cause we would prepare more with that idea in this project.

**What working methods have you used?**

* We use Facebook group to manage team members and share information and learning materials
* We use Git to manage our source codes.
* We use email to communicate with our sponsor Intel.

# Learning experiences

**What have you learned during the project from personal aspect, for each team member? Personal experiences, acquired skills, working as a team, etc…**

Yunkun Niu

This is my first project in a none Chinese language country. I met so many difficulties and also learnt so much. This is a very good opportunity for me to apply my machine learning algorithm knowledge into application and more I enhanced my understand in this field. I also learnt how to manage a team to reach the goal by overcoming all the difficulties during the project. My abilities, such as learning, communicating, analysis and coordinating all got practiced and improved in this project.

Bilal Waris

I learned a lot from this project infect everything is learning while working in demola, its my first project in demola before that I wasn’t know about demola and its procedure but now I will suggest everyone to join for his favorite project in demola, I have learned following skills through this project, management skills, machine learning skills, coding skills, networking with other teams, coordinating with project partner and demola.

Yibin Qian

I have learned a lot through this project. I improved the skills to work individually as well as teamwork. Also I consolidated the knowledge which is quite related to my study. Besides, I understand the project process inside a big company such as Intel.

Qi Tang

In this project, I gained experience in trying many machine algorithms to use solve particular problem. As a result, I deeply understand the machine learning algorithms that we learnt so far, and figure out how to choose correct algorithm to solve different problems. Besides, I learnt how to collaborate with each other in multicultural environment.

# Appendix

**Describe the results in one or two pages**, (create a short presentation from your project, or you can write an NABC pitch from your project)

In this project, we tried three machine learning algorithms to solve the anomaly detection

problem. First we used SOM, the basis of GHSOM, but the result is not satisfied. Because there

are rare anomalies in real system environment. Thus SOM is not suitable for it. Then we used

GMM which is designed for anomaly detection. The first attempt of GMM is perfect, 100%

accuracy. But if the detecting data are generated by slightly modifying good data manually, GMM

failed. Cause the artificial data do not correspond to the distribution of the real data. Even though,

GMM is a good algorithm for anomaly detection and deserved to be tested in the real system

environment. Finally, we designed and implemented a new algorithm, RP. RP works fairly well on

the artificial data. And more it is high efficient and can detect new requests in real--‐time. Overall,

different algorithms match with different data. Each of them has its specific application situation, RP, designed and implemented for the given data by us, performs best for this project.