Notice of the Final Oral Examination
for the Degree of Doctor of Philosophy

of

AMIR FARZAD

MSc (Shahrood University of Technology, 2016)
BSc (Shahrood Azad University, 2010)

“Log Message Anomaly Detection Using Machine Learning”

Department of Electrical and Computer Engineering

Wednesday, June 23, 2021
1:30pm
Remote Defence

Supervisory Committee:
Dr. T. Aaron Gulliver, Department of Electrical and Computer Engineering, University of Victoria (Supervisor)
Dr. Kin Fun Li, Department of Electrical and Computer Engineering, UVic (Member)
Dr. Miguel Nacenta, Department of Computer Science, UVic (Outside Member)

External Examiner:
Dr. Abdul Bais, Department of Electronic Systems Engineering, University of Regina

Chair of Oral Examination:
Dr. Jianping Pan, Department of Computer Science, UVic

Dr. Stephen Evans, Acting Dean, Faculty of Graduate Studies
Abstract

Log messages are one the most valuable sources of information in the cloud and other software systems. These logs can be used for audits and ensuring system security. Many millions of log messages are produced each day which makes anomaly detection challenging. Automating the detection of anomalies can save time and money as well as improve detection performance. In this dissertation, Deep Learning (DL) methods called Auto-LSTM, Auto-BLSTM and Auto-GRU are developed for log message anomaly detection. They are evaluated using four data sets, namely BGL, Openstack, Thunderbird and IMDB. The first three are popular log data sets while the fourth is a movie review data set which is used for sentiment classification. The results obtained show that Auto-LSTM, Auto-BLSTM and Auto-GRU perform better than other well-known algorithms.

Dealing with imbalanced data is one of the main challenges in Machine Learning (ML)/DL algorithms for classification. This issue is more important with log message data as it is typically very imbalanced and negative logs are rare. Hence, a model is proposed to generate text log messages using a Sequence Generative Adversarial Network (SeqGAN) network. Then features are extracted using an Autoencoder and anomaly detection is done using a GRU network. The proposed model is evaluated with two imbalanced log data sets, namely BGL and Openstack. Results are presented which show that oversampling and balancing data increases the accuracy of anomaly detection and classification.

Another challenge in anomaly detection is dealing with unlabeled data. Labeling even a small portion of logs for model training may not be possible due to the high volume of generated logs. To deal with this unlabeled data, an unsupervised model for log message anomaly detection is proposed which employs Isolation Forest and two deep Autoencoder networks. The Autoencoder networks are used for training and feature extraction, and then for anomaly detection, while Isolation Forest is used for positive sample prediction. The proposed model is evaluated using the BGL, Openstack and Thunderbird log message data sets. The results obtained show that the number of
negative samples predicted to be positive is low, especially with Isolation Forest and one Autoencoder. Further, the results are better than with other well-known models. A hybrid log message anomaly detection technique is proposed which uses pruning of positive and negative logs. Reliable positive log messages are first identified using a Gaussian Mixture Model (GMM) algorithm. Then reliable negative logs are selected using the K-means, GMM and Dirichlet Process Gaussian Mixture Model (BGM) methods iteratively. It is shown that the precision for positive and negative logs with pruning is high. Anomaly detection is done using an Long Short-Term Memory (LSTM) network. The proposed model is evaluated using the BGL, Openstack, and Thunderbird data sets. The results obtained indicate that the proposed model performs better than several well-known algorithms.

Last, an anomaly detection method is proposed using radius-based fuzzy C-means (FCM) with more clusters than the number of data classes and a multilayer perceptron (MLP) network. The cluster centers and a radius are used to select reliable positive and negative log messages. Moreover, class probabilities are used with an expert to correct the network output for suspect logs. The proposed model is evaluated with three well-known data sets, namely BGL, Openstack and Thunderbird. The results obtained show that this model provides better results than existing methods.