UNIVERSITY OF TECHNOLOGY, SYDNEY
AUSTRALIA

DOCTORAL THESIS

FEATURE SELECTION USING MUTUAL INFORMATION IN NETWORK INTRUSION DETECTION SYSTEM

Author:
Mohammed AMBUSAIDI

Supervisor:
Prof. Xiangjian He
Co-supervisor:
Dr. Priyadarsi Nanda
Co-supervisor:
A/Prof. Jinjun Chen

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

SCHOOL OF COMPUTING AND COMMUNICATIONS
THE FACULTY OF ENGINEERING AND INFORMATION TECHNOLOGY

December 2015

Declaration of Authorship

I, Mohammed AMBUSAIDI, certify that the work in this thesis has not been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signed: Production Note: Signature removed prior to publication.

Date: 3/12/2015

Abstract

FEATURE SELECTION USING MUTUAL INFORMATION IN NETWORK INTRUSION DETECTION SYSTEM

by Mohammed Ambusaidi

Network technologies have made significant progress in development, while the security issues alongside these technologies have not been well addressed. Current research on network security mainly focuses on developing preventative measures, such as security policies and secure communication protocols. Meanwhile, attempts have been made to protect computer systems and networks against malicious behaviours by deploying Intrusion Detection Systems (IDSs). The collaboration of IDSs and preventative measures can provide a safe and secure communication environment. Intrusion detection systems are now an essential complement to security project infrastructure of most organisations. However, current IDSs suffer from three significant issues that severely restrict their utility and performance. These issues are: a large number of false alarms, very high volume of network traffic and the classification problem when the class labels are not available.

In this thesis, these three issues are addressed and efficient intrusion detection systems are developed which are effective in detecting a wide variety of attacks and result in very few false alarms and low computational cost. The principal contribution is the efficient and effective use of mutual information, which offers a solid theoretical framework for quantifying the amount of information that two random variables share with each other. The goal of this thesis is to develop an IDS that is accurate in detecting attacks and fast enough to make real-time decisions.

First, a nonlinear correlation coefficient-based similarity measure to help extract both linear and nonlinear correlations between network traffic records is used. This measure is based on mutual information. The extracted information is used to

develop an IDS to detect malicious network behaviours. However, the current network traffic data, which consist of a great number of traffic patterns, create a serious challenge to IDSs. Therefore, to address this issue, two feature selection methods are proposed; filter-based feature selection and hybrid feature selection algorithms, added to our current IDS for supervised classification. These methods are used to select a subset of features from the original feature set and use the selected subset to build our IDS and enhance the detection performance.

The filter-based feature selection algorithm, named Flexible Mutual Information Feature Selection (FMIFS), uses the theoretical analyses of mutual information as evaluation criteria to measure the relevance between the input features and the output classes. To eliminate the redundancy among selected features, FMIFS introduces a new criterion to estimate the redundancy of the current selected features with respect to the previously selected subset of features.

The hybrid feature selection algorithm is a combination of filter and wrapper algorithms. The filter method searches for the best subset of features using mutual information as a measure of relevance between the input features and the output class. The wrapper method is used to further refine the selected subset from the previous phase and select the optimal subset of features that can produce better accuracy.

In addition to the supervised feature selection methods, the research is extended to unsupervised feature selection methods, and an Extended Laplacian score EL and a Modified Laplacian score ML methods are proposed which can select features in unsupervised scenarios. More specifically, each of EL and ML consists of two main phases. In the first phase, the Laplacian score algorithm is applied to rank the features by evaluating the power of locality preservation for each feature in the initial data. In the second phase, a new redundancy penalization technique uses mutual information to remove the redundancy among the selected features. The final output of these algorithms is then used to build the detection model.

The proposed IDSs are then tested on three publicly available datasets, the KDD Cup 99, NSL-KDD and Kyoto dataset. Experimental results confirm the effectiveness and feasibility of these proposed solutions in terms of detection accuracy, false alarm rate, computational complexity and the capability of utilising unlabelled data. The unsupervised feature selection methods have been further tested on five more well-known datasets from the UCI Machine Learning Repository. These newly added datasets are frequently used in literature to evaluate the performance of feature selection methods. Furthermore, these datasets have different sample sizes and various numbers of features, so they are a lot more challenging for comprehensively testing feature selection algorithms. The experimental results show that ML performs better than EL and four other state-of-art methods (including the Variance score algorithm and the Laplacian score algorithm) in terms of the classification accuracy.

Acknowledgements

I am pleasure to sincerely thank to my supervisor, **Professor Xiangjian He** for his continuous support, advice, help and invaluable suggestions throughout my PhD journey. His excellent guidance, constant motivation, steadfast encouragement and expert guidance make this journey a rewarding experience in my life that I will never forget. I owe my research achievements to his experienced supervision.

I would like to thanks my co-supervisor, **Dr. Priydarsi Nanda** for his friendly guidance, valuable suggestions and feedback. His encouragement and support have been a great help and kept me moving ahead at a critical time. I gratefully acknowledge the useful discussions with him. I would also like to thanks my co-supervisor, **A/Prof. Jinjun Chen** for his support and friendly advice which has been a great help.

I am extremely thankful to my fellow research colleagues and the staff of the school, especially those people listed below for providing various assistance for the completion of this research work.

- Professor Massimo Piccardi, Professor Doan B. Hoang, Associate Professor Qiang Wu, Dr. Min Xu, Dr. Wenjing Jia, Dr. Zhiyuan Tan, Dr. Aruna Jamdagni, Dr. Chao Zeng, Khaled Aldebei, Ahmed Mian Jan, Shaukat Abedi, Sari Awwad, Huiling Zhou, Sheng Wang, Minqi Li, Guopeng Zhang, Liangfu Lu and Wenbo Wang.

Special thanks to my wife, Intisar Alsabari, for her patience, understanding and assistance. I also thank my father Mr. Abdullah Ambusaidi and my mother Mrs. Azza Alharasi for the freedom to study for the long time necessary to complete postgraduate studies. I also would like to thank my sons, Awab Ambusaidi and Yassin Ambusaidi, for their patience. This thesis could not have been completed without the support and encouragement of my siblings. My special thanks go to my friends for their continuous support and encouragement. I would like to thank

Mr. John Hazelton for his English corrections. Last but not least, I would like to thank my sponsor, the Ministry of Higher Education, Oman, for providing me this opportunity to complete my PhD and for the fiancial assistance.

Contents

List of Figures

List of Tables

Abbreviations

ANN	**A**rtificial **N**eural **N**etworks
DoS	**D**enial **o**f **S**ervice
DR	**D**etection **R**ate
DT	**D**ecision **T**rees
E-LGFS	**E**xtended **L**ocal and **G**lobal structure preserving based on **F**eature **S**election
FE	**F**eature **E**xtraction
FMIFS	**F**lexible **M**utual **I**nformation **F**eature **S**election
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
FS	**F**eature **S**election
GSAD	**G**eometrical **S**tructure **A**nomaly **D**etection
HIDS	**H**ost based **I**ntrusion **D**etection **S**ystem
ICMP	**I**nter **C**ontrol **M**essage **P**rotocol
IDS	**I**ntrusion **D**etection **S**ystem
IFFS	**I**mproved **F**orward **F**loating **S**election
IT	**I**nformation **T**heory
LCC	**L**inear **C**orrelation **C**oefficient
LGFS	**L**ocal and **G**lobal structure preserving based on **F**eature **S**election
LSSVM	**L**east **S**quare **S**upport **V**ector **M**achine

MARS	**M**ultivariate **A**daptive **R**egression **S**plines
MCA	**M**ultivariate **C**orrelation **A**nalysis
MI	**M**utual **I**nformation
MIFS	**M**utual **I**nformation **F**eature **S**election
MIFS-U	**M**utual **I**nformation **F**eature **S**election-**U**niform information distribution
MMIFS	**M**odified **M**utual **I**nformation **F**eature **S**election
mRMR	**M**in-**R**edundancy **M**ax-**R**elevance
NCC	**N**onlinear **C**orrelation **C**oefficient
NIDS	**N**etwork based **I**ntrusion **D**etection **S**ystem
NI	**N**ormalised **M**utual **I**nformation
NLGFS	**N**ormalised **L**ocal and **G**lobal structure preserving based on **F**eature **S**election
NMIFS	**N**ormalised **M**utual **I**nformation **F**eature **S**election
PCA	**P**rincipal **C**omponent **A**nalysis
PCC	**P**earson's **C**orrelation **C**oefficient
PCC-R	**P**earson's **C**orrelation **C**oefficients-**R**ank
PDF	**P**robability **D**ensity **F**unction
R2U	**R**emote to **U**ser
RP	**R**edundancy **P**enalization
SVM	**S**upport **V**ector **M**achine
TANN	**T**riangle **A**rea based **N**earest **N**eighbours
TCP	**T**ransmission **C**ontrol **P**rotocol
TN	**T**rue **N**egative
TP	**T**rue **P**ositive

Publications from this Thesis

JOURNAL PUBLICATIONS:

[1] A. M. Ambusaidi, Z. Tan, X. He, P. Nanda, L. F. Lu, A. Jamdagni, Intrusion detection method based on nonlinear correlation measure, International Journal of Internet Protocol Technology 8 (2) (2014) 77-86.

[2] A. M. Ambusaidi, X. He, Z. Tan, P. Nanda, Building an intrusion detection system using a filter-based feature selection algorithm, IEEE Transactions on Computers, undergo a major revision.

[3] A. M. Ambusaidi, X. He, P. Nanda, Chen, J, Unsupervised Feature Selection Method for Machine Learning, Journal of Network and Computer Applications, to be submitted.

CONFERENCE PUBLICATIONS:

[1] A. M. Ambusaidi, L. F. Lu, X. He, Z. Tan, A. Jamdagni, P. Nanda, A nonlinear correlation measure for intrusion detection, in: International Conference on Frontier of Computer Science and Technology (FCST-12), IEEE, 2012, pp. 1-7.

[2] A. M. Ambusaidi, X. He, Z. Tan, P. Nanda, L. F. Lu, U. T. Nagar, A novel feature selection approach for intrusion detection data classification, International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom-14), IEEE, 2014, pp. 82-89.

[3] A. M. Ambusaidi, X. He, P. Nanda, Unsupervised Feature Selection Method for Intrusion Detection System, International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom-15), IEEE, 2015, Accepted.

Dedicated to My Family