

UNIVERSITY OF TECHNOLOGY, SYDNEY  
AUSTRALIA

DOCTORAL THESIS

---

**FEATURE SELECTION USING MUTUAL  
INFORMATION IN NETWORK INTRUSION  
DETECTION SYSTEM**

---

*Supervisor:*

Prof. Xiangjian He

*Author:*

Mohammed AMBUSAIDI

*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 AMBUSAIKI

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 financial assistance.

# Contents

<b>Declaration of Authorship</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>Contents</b>	<b>vii</b>
<b>List of Figures</b>	<b>xi</b>
<b>List of Tables</b>	<b>xii</b>
<b>Abbreviations</b>	<b>xiv</b>
<b>Publications from this Thesis</b>	<b>xvi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Motivations and Objectives . . . . .	5
1.3 Thesis Contributions . . . . .	7
1.4 Thesis Structure . . . . .	9
<b>2 Related Work</b>	<b>11</b>
2.1 Anomaly Detection System . . . . .	12
2.2 Dependency Measures for Anomaly Detection . . . . .	14
2.2.1 Correlation Coefficient . . . . .	15
2.2.2 Mutual Information . . . . .	18



---

2.2.3	Nonlinear Correlation Coefficient . . . . .	22
2.3	Feature Selection Based on Mutual Information . . . . .	24
2.3.1	Supervised Feature Selection . . . . .	26
2.3.2	Existing Supervised Feature Selection Methods . . . . .	27
2.4	Unsupervised Feature Selection Methods . . . . .	33
2.4.1	Unsupervised Feature Selection . . . . .	33
2.4.2	Existing Unsupervised Feature Selection Methods . . . . .	34
2.5	Description of the Benchmark Datasets for Intrusion Detection . . . . .	40
2.5.1	KDD Cup 99 Dataset . . . . .	40
2.5.2	NSL-KDD Dataset . . . . .	41
2.5.3	Kyoto 2006+ Dataset . . . . .	45
2.6	Summary . . . . .	45
<b>3</b>	<b>Anomaly Detection System Based on Nonlinear Correlation Measure</b>	<b>48</b>
3.1	Linear and Nonlinear Correlation Analysis . . . . .	50
3.1.1	Pearson's Correlation Coefficient . . . . .	50
3.1.2	Nonlinear Correlation Coefficient . . . . .	51
3.2	Intrusion Detection Based on Correlation Coefficient . . . . .	52
3.3	Experimental Results and Analysis . . . . .	58
3.3.1	Dataset Selection . . . . .	58
3.3.2	Performance Evaluation . . . . .	60
3.3.3	Results and Discussion . . . . .	61
3.3.4	Comparative Study . . . . .	65
3.4	Summary . . . . .	67
<b>4</b>	<b>Supervised Filter-based Feature Selection Algorithm for IDS</b>	<b>68</b>
4.1	Filter-based Feature Selection . . . . .	70
4.1.1	Flexible Mutual Information based Feature Selection . . . . .	71
4.1.2	Feature Selection Based on Linear Correlation Coefficient . . . . .	75
4.2	Intrusion Detection Framework-based on Least Square Support Vector Machine . . . . .	78
4.3	Experimental Results and Analysis . . . . .	83
4.3.1	Experimental Setup . . . . .	83
4.3.2	Performance Evaluation . . . . .	84
4.3.3	Results and Discussion . . . . .	87
4.3.4	Comparative Study . . . . .	89
4.3.5	Additional Comparison . . . . .	92

---

4.4	Summary . . . . .	95
<b>5</b>	<b>Supervised Hybrid Feature Selection Algorithm for IDS</b>	<b>98</b>
5.1	Improved Forward Floating Selection . . . . .	99
5.2	Proposed Hybrid Feature Selection . . . . .	101
5.2.1	Filter Method for Feature Pre-selection . . . . .	101
5.2.2	Wrapper-based IFFS for Feature Selection Using LS-SVM . . . . .	103
5.2.2.1	Backtracking . . . . .	104
5.2.2.2	Replacing the Weak Feature . . . . .	105
5.3	Intrusion Detection Framework Based on LS-SVM . . . . .	105
5.4	Experiments and Results . . . . .	107
5.4.1	Results and Discussion . . . . .	109
5.4.2	Comparative Study . . . . .	112
5.5	Summary . . . . .	114
<b>6</b>	<b>Unsupervised Feature Selection Algorithm for IDS</b>	<b>116</b>
6.1	Laplacian Score . . . . .	118
	The Algorithm . . . . .	119
6.2	Modified Laplacian Score . . . . .	120
6.3	Intrusion Detection Based on Unsupervised Feature Selection . . . . .	123
6.4	Experiments and results . . . . .	125
6.4.1	Experimental settings . . . . .	126
6.4.2	Benchmark Datasets . . . . .	127
6.4.3	Results on UCI datasets . . . . .	129
6.4.4	Results on IDS datasets . . . . .	134
6.4.5	Comparison with LGFS and E-LGFS . . . . .	135
6.5	Summary . . . . .	140
<b>7</b>	<b>Conclusion and Future Work</b>	<b>142</b>
7.1	Summary of Contributions . . . . .	142
7.2	Future work . . . . .	146
<b>A</b>	<b>Least Squares Support Vector Machine</b>	<b>148</b>
<b>B</b>	<b>Estimating Mutual Information</b>	<b>151</b>

**Bibliography**

**153**

# List of Figures

2.1	Nearest and farthest neighbourhood graph. . . . .	37
3.1	Overall procedures of the proposed intrusion detection framework . . . . .	53
3.2	The flow chart of the proposed algorithm . . . . .	59
3.3	Matrices expressions of two different measures for normal profiles . . . . .	63
3.4	FPRs for various threshold values on the training dataset . . . . .	65
4.1	The framework of the LS-SVM-based intrusion detection system . . . . .	80
4.2	Building and testing time using all features and FMIFS, respectively, on three datasets. . . . .	89
4.3	Comparison results of $F$ -measure rate on the corrected labels of KDD Cup 99 dataset . . . . .	95
4.4	Comparison results of classification accuracy on KDDTest <sup>-21</sup> . . . . .	96
5.1	The overall scheme of the proposed hybrid feature selection . . . . .	102
5.2	The overall procedure of the proposed wrapper algorithm-based IFFS . . . . .	104
5.3	The framework of the LS-SVM-based intrusion detection system . . . . .	106
5.4	Building and testing times using all features and FMIFS and the proposed method, respectively, on KDD Cup 99. . . . .	111
6.1	The framework of the proposed intrusion detection system . . . . .	124
6.2	Effect of number of selected features on UCI datasets . . . . .	133
6.3	Effect of number of selected features on IDS datasets with the two classifiers . . . . .	136

# List of Tables

2.1	The different group of features in KDD Cup 99 and NSL-KDD dataset	42
2.2	The different types of attacks and description	43
2.3	A list of all features in Kyoto 2006+ dataset	44
3.1	Sample distribution on the training dataset	60
3.2	Sample distribution on the testing dataset	60
3.3	Confusion matrix	62
3.4	DRs for various threshold values on the training dataset	64
3.5	Confusion matrix for NCC-training set	65
3.6	Comparison of detection and false alarm between different IDS using NSL-KDD dataset	66
4.1	Comparison of feature ranking	86
4.2	Performance classification for all attacks based on the three datasets	88
4.3	Feature ranking results for the four types of attacks on the KDD Cup 99 dataset	90
4.4	Comparison results in terms of accuracy rate with other approaches based on the KDD Cup 99 dataset (n/a means not available by authors)	91
4.5	Comparison results based on NSL-KDD dataset (n/a means not available by authors)	91
4.6	Comparison performance of classification on the Kyoto 2006+ dataset (the days 2007, Nov. 1,2 and 3), #I is the number of Iteration	92
4.7	Accuracy, building time (min) and testing time (min) for all different classes on corrected labels of KDD Cup 99 dataset compare with PLSSVM proposed by Amiri in [1].	93
4.8	Detection rate (%) for different algorithm performances on the test dataset with corrected labels of KDD Cup 99 dataset (n/a means not available by authors)	94
5.1	Comparison of feature ranking	108

---

5.2	Performance of classification based on the evaluation data on KDD Cup 99 . . . . .	110
5.3	Performance of classification based on Kyoto 2006+ data . . . . .	111
5.4	Comparison results in terms of accuracy rate with other approaches based on the evaluation dataset . . . . .	112
5.5	Performance of classification based on the corrected labels of KDD Cup 99 data (n/a means not available by authors) . . . . .	113
6.1	General information and summary of datasets used in the experiments	127
6.2	A comparison of classification accuracies using three feature selection algorithms on UCI datasets based on 1NN . . . . .	131
6.3	A comparison of classification accuracies using three feature selection algorithms on seven datasets based on SVM . . . . .	132
6.4	A comparison of classification accuracies using three feature selection algorithms on IDS datasets based on 1NN . . . . .	135
6.5	A comparison of classification accuracies using three feature selection algorithms on seven datasets based on SVM . . . . .	137
6.6	A comparison of classification accuracies using three feature selection algorithms on four UCI datasets based on 1NN . . . . .	138
6.7	A comparison of classification accuracies using three feature selection algorithms on four UCI datasets based on SVM . . . . .	139

# Abbreviations

<b>ANN</b>	<b>Artificial Neural Networks</b>
<b>DoS</b>	<b>Denial of Service</b>
<b>DR</b>	<b>Detection Rate</b>
<b>DT</b>	<b>Decision Trees</b>
<b>E-LGFS</b>	<b>Extended Local and Global structure preserving based on Feature Selection</b>
<b>FE</b>	<b>Feature Extraction</b>
<b>FMIFS</b>	<b>Flexible Mutual Information Feature Selection</b>
<b>FN</b>	<b>False Negative</b>
<b>FP</b>	<b>False Positive</b>
<b>FS</b>	<b>Feature Selection</b>
<b>GSAD</b>	<b>Geometrical Structure Anomaly Detection</b>
<b>HIDS</b>	<b>Host based Intrusion Detection System</b>
<b>ICMP</b>	<b>Inter Control Message Protocol</b>
<b>IDS</b>	<b>Intrusion Detection System</b>
<b>IFFS</b>	<b>Improved Forward Floating Selection</b>
<b>IT</b>	<b>Information Theory</b>
<b>LCC</b>	<b>Linear Correlation Coefficient</b>
<b>LGFS</b>	<b>Local and Global structure preserving based on Feature Selection</b>
<b>LSSVM</b>	<b>Least Square Support Vector Machine</b>

---

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