EVALUATION OF MACHINE LEARNING TECHNIQUES FOR IMBALANCED DATA IN IDS

SHAHRAM MOKARAMIAN

A project submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Computer Science (Information Security)

Faculty of Computing
Universiti Teknologi Malaysia

AUGUST 2013

I dedicate my thesis to my family. A special feeling of thankfulness to my loving wife, “Fatemeh” whose without her help and support, this work could not be done. To my parents whose words of inspiration and push for endurance ring in my ears.

ACKNOWLEDGEMENTS

I would like to truly appreciate **Dr. Anazida Zainal** for taking the time out of her busy schedule to encourage and guide me through this project. Her deep knowledge and valuable experience inspired me and illuminated my path. Also for her great advices and technical points that she gave me with attractiveness.

Besides I should be appreciative of authority of Universiti Teknologi Malaysia (UTM) for providing me with a good environment and creative area.

I am deeply indebted to my wife, Fatemeh, for her continuous support and the inspiration throughout the journey. She is the best companion. Special thank you to my daughter, Ghazal who has inspired me to keep on striving to complete the study.

ABSTRACT

Network Intrusion Detection System (IDS) is an automated system that can detect a malicious traffic and it plays a critical role in a network. In recent years, machine learning algorithms have been developed and used to detect network intrusion. Most standard machine learning algorithms often give high overall accuracy. However, they favor on majority class when dealing with imbalanced data. Unfortunately, IDS deals with highly imbalanced data distribution and most machine learning algorithms have poor detection on R2L and U2R classes, which include malicious attacks. Therefore, it requires a resampling technique to balance the data. The purpose of this study is to investigate performance of three machine learning algorithms which are Support Vector Machine (SVM), Decision Tree (DT) and Fuzzy Classifier (FC) for imbalanced data in IDS and after the rebalanced the data which was achieved using Synthetic Minority Over-sampling TEchnique (SOMTE). The performance of the three machine learning algorithms was evaluated with the new rebalanced data. The benchmark DARPA KDDCup 1999 IDS dataset was used. SMOTE was implemented with two imbalance ratio, one is 1:4 another one is 1:1. After analysis the results of before and after resampling showed that FC performs better with imbalance ratio of 1:1. The accuracy of FC with balanced data was Normal traffic (99.19%), Denial of Service attacks (99.35%), Probe attacks (99.51%), Remote to Local attacks (99.67%) and User to Root attacks (99.41%). In addition, the data with imbalance ratio of 1:1 get the better results on all classes with these three machine learning algorithms.

ABSTRAK

Intrusion Detection System (IDS) Rangkaian adalah sistem automatik yang boleh mengesan trafik yang berniat jahat dan ia memainkan peranan penting dalam rangkaian. Pada tahun-tahun kebelakangan ini, algoritma pembelajaran mesin telah dibangunkan dan digunakan untuk mengesan pencerobohan rangkaian. Kebanyakkan algoritma pembelajaran mesin yang piawai sering memberi ketepatan keseluruhan yang tinggi. Namun, mereka seriang memihak kepada kelas majoriti apabila berurusan dengan data yang tidak seimbang. Malangnya, IDS menawarkan pengagihan data yang sangat tidak seimbang dan kebanyakkan algoritma pembelajaran mesin memberikan pengesanan yang rendah untuk kelas R2L dan U2R, termasuk serangan berbahaya. Oleh itu, ia memerlukan teknik persempelan semula untuk mengimbangi data tersebut. Tujuan kajian ini adalah untuk menyelidik prestasi tiga algoritma pembelajaran mesin iaitu *Support Vector Machine* (SVM), *Decision Tree* (DT) dan *Fuzzy Classifier* (FC) untuk ketidakseimbangan data dalam IDS dan data yang telah diseimbangkan yang dapat dicapai melalui Synthetic Minority Over-sampling TEchnique (SOMTE). Prestasi ketiga-tiga algoritma pembelajaran mesin kemudian dinilai dengan data baru yang telah diseimbangkan. Penanda aras set data DARPA KDDCup 1999 IDS telah digunakan. SMOTE telah dilaksanakan dengan dua nisbah ketidakseimbangan, iaitu 1:4 dan 1:1. Setelah menganalisis keputusan sebelum dan selepas pengsempelan semula, ia menunjukkan bahawa FC menunjukkan keputusan yang lebih baik dengan nisbah ketidakseimbangan 1:1. Ketepatan FC dengan data seimbang untuk trafik Normal adalah (99.19%), serangan *Denial of Service* (99,35%), serangan *Probe* (99,51%), serangan *Remote to Local* (99.67%) dan serangan *User to Root* (99.41%). Di samping itu, data dengan nisbah ketidakseimbangan 1:1 mencapai keputusan terbaik untuk ketiga-tiga kelas algoritma pembelajaran mesin.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
DoS	Denial of Service
DT	Decision Tree
FAR	False Alarm Rate
FC	Fuzzy Classifier
FNR	False Negative Rate
FPR	False Positive Rate
HIDS	Host-based Intrusion Detection System
IDS	Intrusion Detection System
IR	Imbalance Ratio
KNN	K-Nearest Neighbor
MLP	Multi-Layer Perceptron
NIDS	Network-based Intrusion Detection System
R2L	Remote to Local
SMOTE	Synthetic Minority Over-sampling TEchnique
SOM	Self-Organizing Map
SVM	Support Vector Machine
TNR	True Negative Rate
TPR	True Positive Rate
U2R	User to Root

(Galar *et al.*, 2012; Anderson, 1980; Denning, 1987; Beghdad, 2009; Vapnik, 1998; Phoungphol *et al.*, 2012; Yu *et al.*, 2007; Zadeh, 1965; Chawla *et al.*, 2002; Yen and Lee, 2009; Yoon and Kwek, 2007; Kubat and Matwin, 1997; Barandela *et al.*, 2004; Fernandez *et al.*, 2010)

CHAPTER 1

INTRODUCTION

1.1 Overview

The class imbalance problem is a difficult challenge faced by machine learning and data mining, and it has attracted a significant amount of research in these years. A classifier affected by the class imbalance problem for a specific dataset would see strong accuracy overall but very poor performance on the minority class. This study will evaluated three machine learning algorithms for imbalanced data problem before and after rebalancing the dataset in intrusion detection system (IDS).

1.2 Background of the Problem

Nowadays, cyber-crime has become one of the most important problems in the computer world. All over the world companies and governments are increasingly dependent on their computer networks and communications, hence need to protect these systems from attack. Find the best possible way to protect all information system is needed. The prevention techniques such as encryption, Virtual Private Network (VPN) and firewall alone seem to be inadequate (Zainal, 2011) . It is

important to have a detecting and monitoring system to protect important data. Intrusion detection is identifying unauthorized users in a computer system.

Intrusion Detection System (IDS) is an automated system that can detect a computer system intrusion either by using the audit trail provided by an operating system or by using the network monitoring tools. The main goal of intrusion detection is to detect unauthorized use, misuse and abuse of computers by both system insiders and external intruders. In IDS, misuse and anomaly are the two types of detection approaches. Misuse detection can detect known attacks by constructing a set of signatures of attacks while anomaly detection recognizes novel attacks by modeling of normal behaviors (Xu and Wang, 2005; Zainal, 2011).

Intrusion detection is a tool of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problem. The network traffic is made up of attack and normal traffic. The number of attacks on the network is typically a very small fraction of the total traffic. On the basis of this the attacks can also be categorized into two classes, minority and majority attack class. The normal data, Denial of Service (DoS) and Probe attacks belong to majority class whereas User-to-Root (U2R) and Remote-to-Local (R2L) belong to minority class also called as rare class of attacks. In real world environment, the minority attacks are more dangerous than the majority (Sharma and Mukherjee, 2012; Lopez *et al*., 2012).

Intrusion detection systems goal is to detect malicious action in close to real-time and raise an alert. Operators can then take proper actions to decrease any impact of the activity. Intrusion detection systems also can categories by either HIDS (host-based) or NIDS (network-based). Network-based IDS analyze network traffic to monitor entire computer networks. IDSs also can be further labeled as anomaly-based or misuse-based (Davis and Clark, 2011).

Additionally, intrusion detection techniques can be mapped into two classes: anomaly detection and misuse detection. Anomaly detection consists of establishing normal behavior profile for user and system activity also observing significant deviations of actual user activity with respect to the established habitual pattern. Misuse detection, refers to intrusions that follow well-defined attack patterns that exploit weaknesses in system and application software (Beghdad, 2009).

The difficulty faced by IDS is highly imbalanced data distribution (Wu and Banzhaf, 2010; Zainal, 2011). Imbalanced data further make difficult the anomaly detection cases. Most of the studies implementing supervised method like LGP (Prasad *et al.*, 2008) and Neuro-Fuzzy (Toosi and Kahani, 2007) on KDD Cup 1999 Intrusion Detection Datasets, reported poor results especially on R2L (Remote to Local) and U2R (User to Root) attacks. This is because R2L and U2R constitute the least data in the experimental dataset (KDD Cup 1999) compared to other classes of traffic (Normal, Denial of Service (DoS) and Probe).

Network traffic contains class imbalanced problem. The class imbalanced problem arises when some particular classes are represented with too many instances and the other some classes have very few instances (Zainal, 2011). Usually the classification is biased towards the classes with majority instances (Liao, 2008). Most reported works in IDS (Hossain *et al.*, 2003; Xu and Wang, 2005; Lee *et al.*, 2006; Shafi and Abbass, 2009; Jemili *et al.*, 2007; Zainal, 2011) reported poor detection on U2R and R2L classes.

As an example, let consider a data set whose imbalance ratio is 1:100 (i.e., for each example of the positive class, there are 100 negative class examples). A classifier tries to maximize the accuracy of its classification rule, may obtain an accuracy of 99% just by the ignorance of the positive examples, with the classification of all instances as negatives.

As stated by Galar *et al*. (2012), in recent years, class imbalance problem has emerged as one of the challenges in data mining community and a number of solutions have been proposed at the data and algorithm levels and trying to address the imbalanced data problem. The problem of imbalanced data is not properly addressed. Most machine learning algorithms are influenced towards the class with more instances and give poor detection performance for minority class and give out high false alarm rate.

Fuzzy Classifier (FC) has the ability to handle datasets with overlapping and imbalancing problem that is a good potential solution since IDS datasets are usually extremely skewed (Ali *et al.*, 2011).

The main capability of fuzzy classifier is better than the standard classifiers, which proposed by many researchers (Ali *et al.*, 2011; Visa, 2006). When compare to other classifiers, the FC is a better candidate for classification of imbalanced data. More precisely, the fuzzy classifier recognizes better the minority class while also achieving better overall accuracy then neural network and RF (random Forest) (Ali *et al.*, 2011).

SVM (Support Vector Machine) and Decision Trees (DT) are also two popular machine-learning algorithms, which are widely used for classification with imbalanced data (Phoungphol *et al.*, 2012; Chandrasekhar and Raghuveer, 2013; Teng *et al.*, 2010; Liu *et al.*, 2010). SVM was introduced by Vapnik (1998) is one of the most fascinating recent developments in classifier design. SVMs have several important properties including the ability to model complex nonlinear decision boundaries, good performance in a wide variety of applications, less prone to overfitting, and a compact description of the learning models. Decision trees use simple knowledge representation to classify examples into a limited number of classes. In a standard setting, the tree nodes denote the attributes, the edges represent

the possible values for a particular attribute and the leaves are assigned with class labels.

These machine learning algorithms cannot get satisfactory results with severely imbalanced data. So, imbalanced data is needed to balance the data by using one resampling technique. One of the popular resampling approach is SMOTE (Synthetic Minority Over-sampling TEchnique), which adds information to the training set by introducing new, non-replicated minority class examples (Chawla *et al.*, 2002). The results show that the SMOTE approach can improve the accuracy of classifiers for a minority class.

1.3 Statement of the Problem

One of the main challenges in intrusion detection system is that, few attacks are rarely happened. IDS deal with highly imbalanced data distribution. This would lead to significantly disparate or too small training dataset for determined classes. Most of the standard machine learning techniques are influenced towards the majority classes and give poor detection performance for classes with very less data samples during training which giving out high false alarm rate. This research gives a primary focus on this imbalanced issue.

1.4 Purpose of the Research

The aim of this study is to improve detection accuracy for imbalanced class in IDS especially U2R and R2L. Generally this will improve detection accuracy as well.

1.5 Objectives of the Study

The particular objectives of this study are:

i. To study and investigate performance of three machine learning algorithms (FC, SVM and DT) for imbalanced data in IDS.
ii. To rebalance the data by an up-sampling algorithm (SMOTE) to deal with severely imbalanced problem and evaluate the performance of FC, SVM and DT techniques.
iii. To compare the results of these machine learning algorithms with imbalanced data before the rebalancing and after the rebalancing dataset.

1.6 Scope of the Study

The scope of this study will be limited to following:

i. The study will use KDD Cup 1999 Intrusion Detection data set (http://kdd.ics.uci.edu/databases/kddcup99) as widely used by other researchers in the field of IDS (Jemili *et al.*, 2007; Shafi and Abbass, 2009; Zainal, 2011; Abraham *et al.*, 2007; Tajbakhsh *et al.*, 2009; Farid *et al.*, 2010).
ii. Classification of attacks are based on four established dominant categories which are Denial of Service (DoS), Probe, User to Root (U2R) and Remote to Local (R2L) as widely used in other studies in the field of IDS (Abraham *et al.*, 2007; Shafi and Abbass, 2009; Zainal, 2011; Tajbakhsh *et al.*, 2009; Farid *et al.*, 2010; Teng *et al.*, 2010).
iii. It is assumed that the cost implications for making decisions are the same for all type of attacks as widely assumed by other researchers in

the field of IDS (Abraham *et al.*, 2007; Shafi and Abbass, 2009; Zainal, 2011).

1.7 Significant of the Study

The research is important and significant from theoretical and practical perspectives. The rationale and motivation for this research is imbalanced data, which is commonly found in intrusion detection domain, has reduced the performance of machine learning based IDS.

The research findings are expected to lead to better understanding on the nature of computer network security and provide a better approach deal with imbalanced data IDS. As such, they should benefit both researchers and practitioners.

1.8 Research Contributions

The main contribution is to evaluate of the using three machine learning algorithms (FC, SVM and DT that widely used by other researchers) on imbalanced data in IDS before rebalancing the dataset and after rebalancing the dataset and which can be more accurate for imbalanced data problem.

1.9 Research Methodology

This part quickly presents the research methodology in this study. The details will be offered in Chapter 3. Phase 1 dealt with implementing and testing three machine learning algorithms (FC, SVM and DT) on imbalanced data and compare the results. In phase 2, procedures to rebalanced the data by an up-sampling algorithm and test it. Finally, Phase 3 dealt with evaluating and comparing the performance of these three techniques with rebalance data and without rebalance.

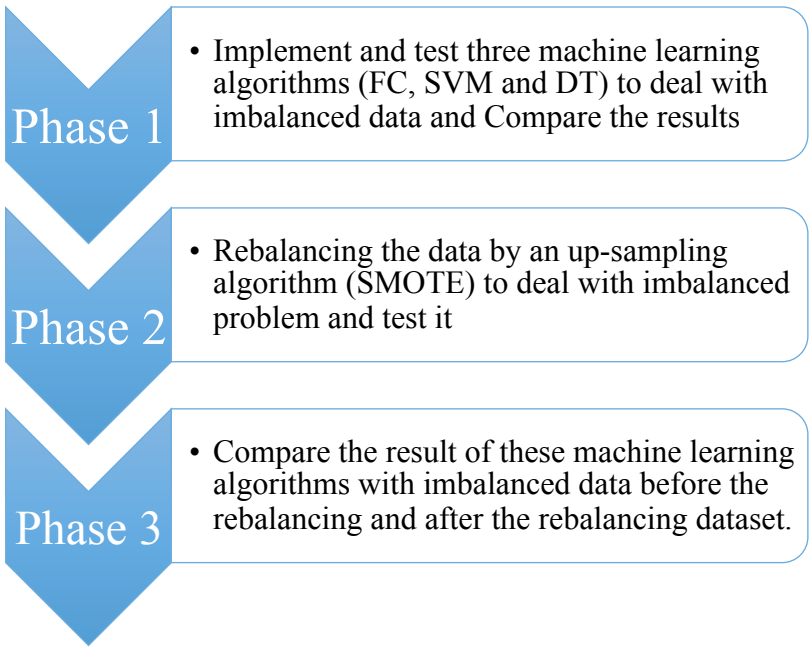

Figure 1.1: Design and development phases of the study

1.10 Organization of the Study

This study is organized into four chapters as shown in Figure 1.2.

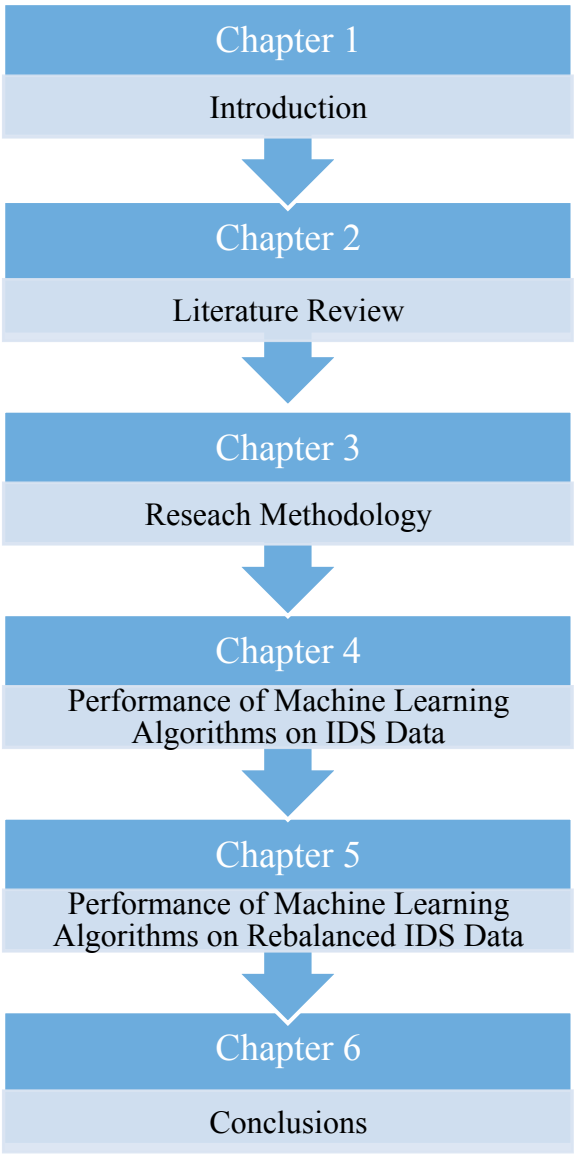

Figure 1.2: Organization of the thesis

Chapter 1 is an outline to this study. Chapter 2 will be provided a literature review that leads to understand the research problem and get information of related work by other researchers. Chapter 3 will provide the research methodology. Chapter 4 will be provided the performance of three machine learning algorithms which are SVM, DT and FC on IDS data. The performance of these three machine learning algorithms after rebalancing by SOMTE technique and before rebalancing will be provided in Chapter 5. Finally, Chapter 6 is the conclusion of this work.

REFERENCES

Abadeh, M. S., Habibi, J., Barzegar, Z. and Sergi, M. (2007). A Parallel Genetic Local Search Algorithm for Intrusion Detection in Computer Networks. *Engineering Applications of Artificial Intelligence.* 20(8)**,** 1058-1069.

Abraham, A., Grosan, C. and Martin-Vide, C. (2007). Evolutionary Design of Intrusion Detection Programs. *International Journal of Network Security.* 4(3)**,** 328-339.

Ali, A., Shamsuddin, S. M., Ralescu, A. L. and Visa, S. (2011). Fuzzy Classifier for Classification of Medical Data. *Proceedings of the 2011 IEEE nternational Conference on Hybrid Intelligent Systems (HIS)*. 5-8 Dec. 2011. IEEE, 173-178.

Anderson, J. P. (1980). *Computer Security Threat Monitoring and Surveillance.* Fort Washington, Pennsylvania: James P. Anderson Co.

Barandela, R., Valdovinos, R. M., Sánchez, J. S. and Ferri, F. J. (2004). The Imbalanced Training Sample Problem: Under or over Sampling? In Fred, A., Caelli, T., Duin, R. W., Campilho, A. and de Ridder, D. (Eds.). *Structural, Syntactic, and Statistical Pattern Recognition.* (pp. 806-814) Springer Berlin Heidelberg.

Beghdad, R. (2009). Efficient Deterministic Method for Detecting New U2r Attacks. *Computer Communications.* 32(6)**,** 1104-1110.

Breiman, L. (1996). Bagging Predictors. *Mach. Learn.* 24(2)**,** 123-140.

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45(1)**,** 5-32.

Chandrasekhar, A. M. and Raghuveer, K. (2013). Intrusion Detection Technique by Using K-Means, Fuzzy Neural Network and Svm Classifiers. *Proceedings of the 2013*.

Chari, S. N. and Cheng, P. C. (2003). Bluebox: A Policy-Driven, Host-Based Intrusion Detection System. *ACM Transactions on Information and System Security.* 6(2)**,** 173-200.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research.* 16**,** 321-357.

Chawla, N. V., Japkowicz, N. and Kotcz, A. (2004). Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explor. Newsl.* 6(1)**,** 1-6.

Chen, Y., Abraham, A. and Yang, B. (2007). Hybrid Flexible Neural-Tree-Based Intrusion Detection Systems. *International Journal of Intelligent Systems.* 22(4)**,** 337-352.

Chimphlee, W. (2008). *Hybrid Fuzzy Techniques for Unsupervised Intrusion Detection System.* PhD Thesis, Universiti Tecknologi Malaysia.

Chimphlee, W., Abdullah, A. H., Sap, M. N. M., Srinoy, S. and Chimphlee, S. (2006). Anomaly-Based Intrusion Detection Using Fuzzy Rough Clustering. *Proceedings of the 2006 Hybrid Information Technology (ICHIT'06)*. 329-334.

Crothers, T. (2002). *Implementing Intrusion Detection Systems: A Hands-on Guide for Securing the Network.* Wiely.

Davis, J. J. and Clark, A. J. (2011). Data Preprocessing for Anomaly Based Network Intrusion Detection: A Review. *Computers & Security.* 30(6–7)**,** 353-375.

Denning, D. E. (1987). An Intrusion-Detection Model. *Software Engineering, IEEE Transactions on.* SE-13(2)**,** 222-232.

Ektefa, M., Memar, S., Sidi, F. and Affendey, L. S. (2010). Intrusion Detection Using Data Mining Techniques. *Proceedings of the 2010 Information Retrieval & Knowledge Management, (CAMP), 2010 International Conference on*. 17-18 March 2010. 200-203.

Farid, D. M., Harbi, N. and Rahman, M. Z. (2010). Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection. *International Journal of Network Security and Its Applications.* 2(2)**,** 12-25.

Fernandez, A., Jesus, M. J. d. and herrera, F. (2010). On the 2-Tuples Based Genetic Tuning Performance for Fuzzy Rule Based Classification Systems in Imbalanced Data-Sets. *Information Sciences.* 180**,** 1268-1291.

Fraichard, T. and Garnier, P. (2001). Fuzzy Control to Drive Car-Like Vehicles. *Robotics and Autonomous Systems.* 34(1)**,** 1-22.

Galar, M., Ferna, x, ndez, A., Barrenechea, E., Bustince, H. and Herrera, F. (2012). A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on.* 42(4)**,** 463-484.

Gao, M., Hong, X., Chen, S. and Harris, C. J. (2011). A Combined Smote and Pso Based Rbf Classifier for Two-Class Imbalanced Problems. *Neurocomputing.* 74(17)**,** 3456-3466.

Guan, J., Liu, D.-x. and Wang, T. (2004). Applications of Fuzzy Data Mining Methods for Intrusion Detection Systems. In Laganá, A., Gavrilova, M., Kumar, V., Mun, Y., Tan, C. J. K. and Gervasi, O. (Eds.). *Computational Science and Its Applications – Iccsa 2004.* (pp. 706-714) Springer Berlin Heidelberg.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation.* Prentice Hall PTR.

Hossain, M., Bridges, S. M. and Vaughn Jr, R. B. (2003). Adaptive Intrusion Detection with Data Mining. *Proceedings of the 2003 IEEE International Conference on Systems, Man and Cybernetics*. 5-8 Oct. 2003. IEEE, 3097-3103 vol.4.

Japkowicz, N. (2001). Concept-Learning in the Presence of between-Class and within-Class Imbalances. In Stroulia, E. and Matwin, S. (Eds.). *Advances in Artificial Intelligence.* (pp. 67-77) Springer Berlin Heidelberg.

Japkowicz, N. and Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.* 6(5)**,** 429-449.

Jeatrakul, P. and Kok-Wai, W. (2012). Enhancing Classification Performance of Multi-Class Imbalanced Data Using the Oaa-Db Algorithm. *Proceedings of the 2012 Neural Networks (IJCNN), The 2012 International Joint Conference on*. 10-15 June 2012. 1-8.

Jemili, F., Zaghdoud, M. and Ben Ahmed, M. (2007). A Framework for an Adaptive Intrusion Detection System Using Bayesian Network. *Proceedings of the 2007 IEEE Intelligence and Security Informatics*. 23-24 May 2007. IEEE, 66-70.

Kabiri, P. and Ghorbani, A. A. (2005). Research on Intrusion Detection and Response: A Survey. *International Journal of Network Security.* 1(2)**,** 84-102.

kailashiya, D. and Jain, D. R. C. (2011). Improve Intrusion Detection for Decision Tree with Stratified Sampling. *Int.J.Computer Technology & Applications.* Vol 3(3)**,** 1209-1216.

Kendall, K. (1998). *A Database of Computer Attacks for Evaluation of Intrusion Detection Systems.* PhD Thesis, Massachusetts Insitute of Technlogy,Boston.

Koblar, V. (2012). *Optimizing Parameters of Machine Learning Algorithms.* PhD Jozef International Postgraduate School.

Kohonen, T. (1990). The Self-Organizing Map. *Proceedings of the IEEE.* 78(9)**,** 1464-1480.

Kressel, U. H. G. (1999). Pairwise Classification and Support Vector Machines. 255-268.

Kruegel, C., Mutz, D., Valeur, F. and Vigna, G. (2003). On the Detection of Anomalous System Call Arguments. *LECTURE NOTES IN COMPUTER SCIENCE.* (2808)**,** 326-344.

Kubat, M. and Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *In Proceedings of the Fourteenth International Conference on Machine Learning.* (pp. 179-186) Morgan Kaufmann.

Lazarevic, A., Srivastava, J. and Kumar, V. (2005). *Managing Cyber Threats.* Springer US.

Le, G. H. N. (2011). *Machine Learning with Informative Samples for Large and Imbalanced*

Datasets. PhD Thesis, School of Electrical, Computer and Telecommunications Engineering, University of Wollongong.

Lecouteux, B., Linarès, G. and Oger, S. (2012). Integrating Imperfect Transcripts into Speech Recognition Systems for Building High-Quality Corpora. *Computer Speech & Language.* 26(2)**,** 67-89.

Lee, H., Chung, Y. and Park, D. (2006). An Adaptive Intrusion Detection Algorithm Based on Clustering and Kernel-Method. *Proceedings of the 2006 Springer-Verlag Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*. Singapore: Springer-Verlag, 603-610.

Liao, T. W. (2008). Classification of Weld Flaws with Imbalanced Class Data. *Expert Systems with Applications.* 35(3)**,** 1041-1052.

Lin, S.-J., Chang, C. and Hsu, M.-F. (2013). Multiple Extreme Learning Machines for a Two-Class Imbalance Corporate Life Cycle Prediction. *Knowledge-Based Systems.* 39(0)**,** 214-223.

Lin, S.-W., Lee, Z.-J., Chen, S.-C. and Tseng, T.-Y. (2008). Parameter Determination of Support Vector Machine and Feature Selection Using Simulated Annealing Approach. *Applied Soft Computing.* 8(4)**,** 1505-1512.

Liu, A., Ghosh, J. and Martin, C. (2007). Generative Oversampling for Mining Imbalanced Datasets. *Dmin.* (pp. 66-72).

Liu, Y., Li, N., Shi, L. and Li, F. (2010). An Intrusion Detection Method Based on Decision Tree. *Proceedings of the 2010 E-Health Networking, Digital Ecosystems and Technologies (EDT), 2010 International Conference on*. 17-18 April 2010. 232-235.

Lopez, V., Fernandez, A., Jesus, M. J. d. and Herrera, F. (2012). A Hierarchical Genetic Fuzzy System Based on Genetic Programming for Addressing Classification with Highly Imbalanced and Borderline Data-Sets. *Knowledge-Based Systems.* 38**,** 85-104.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Murphey, Y. L., Wang, H., Ou, G. and Feldkamp, L. A. (2007). Oaho: An Effective Algorithm for Multi-Class Learning from Imbalanced Data. *Proceedings of the 2007* 406-411.

Panigrahi, S., Kundu, A., Sural, S. and Majumdar, A. K. (2009). Credit Card Fraud Detection: A Fusion Approach Using Dempster–Shafer Theory and Bayesian Learning. *Information Fusion.* 10(4)**,** 354-363.

Phoungphol, P., Zhang, Y., Zhao, Y. and Srichandan, B. (2012). Multiclass Svm with Ramp Loss for Imbalanced Data Classification. *Proceedings of the 2012 Granular Computing (GrC), 2012 IEEE International Conference on*. 11-13 Aug. 2012. 376-381.

Prasad, G. V. S. N. R. V., Dhanalakshmi, Y., Kumar, V. V. and Babu, I. R. (2008). Modeling an Intrusion Detection System Using Data Mining and Genetic Algorithms Based on Fuzzy Logic. *International Journal of Computer Science and Network Security.* 8(7)**,** 319-324.

Quah, J. T. S. and Sriganesh, M. (2008). Real-Time Credit Card Fraud Detection Using Computational Intelligence. *Expert Systems with Applications.* 35(4)**,** 1721-1732.

Qureshi, A. A. (2006). *Network Intrusion Detection Using an Innovative Statistical Approach.* Florida Institute of Technology.

Raskutti, B. and Kowalczyk, A. (2004). Extreme Re-Balancing for Svms: A Case Study. *SIGKDD Explor. Newsl.* 6(1)**,** 60-69.

Schapire, R. E. (1990). The Strength of Weak Learnability. *Mach. Learn.* 5(2)**,** 197-227.

Shafi, K. and Abbass, H. A. (2009). An Adaptive Genetic-Based Signature Learning System for Intrusion Detection. *Expert Systems with Applications.* 36(10)**,** 12036-12043.

Shanmugam, B. (2010). *A Hybrid Intelligent Intrusion Detection System Using Fuzzy Logic and Improved Packet Tracing Model.* PhD Thesis, Universiti Teknologi Malaysia.

Sharma, N. and Mukherjee, S. (2012). A Novel Multi-Classifier Layered Approach to Improve Minority Attack Detection in Ids. *Procedia Technology.* 6(0)**,** 913-921.

Shuo, W. and Xin, Y. (2012). Multiclass Imbalance Problems: Analysis and Potential Solutions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on.* 42(4)**,** 1119-1130.

Shuo, W. and Xin, Y. (2013). Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures. *Knowledge and Data Engineering, IEEE Transactions on.* 25(1)**,** 206-219.

Sun, Y., Wong, A. K. C. and Kamel, M. S. (2009). Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence.* 23(04)**,** 687-719.

Tahir, M., Khan, A., Majid, A. and Lumini, A. (2013). Subcellular Localization Using Fluorescence Imagery: Utilizing Ensemble Classification with Diverse Feature Extraction Strategies and Data Balancing. *Applied Soft Computing.* (0).

Tajbakhsh, A., Rahmati, M. and Mirzaei, A. (2009). Intrusion Detection Using Fuzzy Association Rules. *Applied Soft Computing.* 9(2)**,** 462-469.

Teng, S., Du, H., Wu, N., Zhang, W. and Su, J. (2010). A Cooperative Network Intrusion Detection Based on Fuzzy Svms. *JOURNAL OF NETWORKS.* 5(4)**,** 475-483.

Thammasiri, D., Delen, D., Meesad, P. and Kasap, N. (2013). A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition. *Expert Systems with Applications.* (0).

Toosi, A. N. and Kahani, M. (2007). A New Approach to Intrusion Detection Based on an Evolutionary Soft Computing Model Using Neuro-Fuzzy Classifiers. *Computer Communications.* 30(10)**,** 2201-2212.

Valeur, F., Vigna, G., Kruegel, C. and Kemmerer, R. A. (2004). A Comprehensive Approach to Intrusion Detection Alert Correlation. *IEEE Transactions on Dependable and Secure Computing.* 1(3)**,** 146-168.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley-Interscience.

Visa, S. (2006). *Fuzzy Classifiers for Imbalanced Data Sets.* University of Cincinnati, Cincinnati.

Wasikowski, M. (2009). *Combating the Class Imbalance Problem in Small Sample Data Sets.* Master Thesis, University of Kansas School of Engineering

Weiss, G. M. (2004). Mining with Rarity: A Unifying Framework. *SIGKDD Explor. Newsl.* 6(1)**,** 7-19.

Wu, S. X. and Banzhaf, W. (2010). The Use of Computational Intelligence in Intrusion Detection Systems: A Review. *Applied Soft Computing.* 10(1)**,** 1-35.

Xu, X. and Wang, X. (2005). An Adaptive Network Intrusion Detection Method Based on Pca and Support Vector Machines. *Proceedings of the 2005 Springer-Verlag international conference on Advanced Data Mining and Applications*. Wuhan, China: Springer-Verlag, 696-703.

Yang, L., Wang, R. and Zeng, Y.-S. (2007). An Improvement of One-against-One Method for Multi-Class Support Vector Machine. *Proceedings of the 2007 Machine Learning and Cybernetics, 2007 International Conference on*. 19-22 Aug. 2007. 2915-2920.

Yen, S.-J. and Lee, Y.-S. (2009). Cluster-Based under-Sampling Approaches for Imbalanced Data Distributions. *Expert Systems with Applications.* 36(3 PART 1)**,** 5718-5727.

Yoon, K. and Kwek, S. (2007). A Data Reduction Approach for Resolving the Imbalanced Data Issue in Functional Genomics. *Neural Computing and Applications.* 16(3)**,** 295-306.

Yu, T., Jan, T., Simoff, S. and Debenham, o. (2007). A Hierarchical Vqsvm for Imbalanced Data Sets. *Proceedings of the 2007 International Joint Conference on Neural Networks*. August 12-17, 2007 Orlando, Florida, USA: 518-523.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control.* 8(3)**,** 338-353.

Zainal, A. B. (2011). *An Adaptive Intrusion Detection Model for Dynamic Network Traffic Patterns Using Machine Learning Techniques.* PhD Thesis, Universiti Teknologi Malasia.