Acknowledgements

The first thankful is due to Allah Almighty, the Almighty, who granted me knowledge. Without his support and guidance, this work would not have been possible, so all praise is to ALLAH.

Then I would like to acknowledge my family for all their efforts since my birth to these moments and they were the true supporter after God Almighty to overcome all difficulties.

I am pleased to thank everyone who advised me, guided me, directed me, or contributed with me in preparing this research by sending me to the references and sources required at any stage of the research, preparation, and I especially thank my esteemed professor Dr. Ibrahim Al-Merhag for his support and guidance in advice and correction and for my help in choosing the title and topic.

Also, my thanks go to the administration of the Faculty of Sciences at Zawia University. Also, my special thanks go to the head and faculty members of the Computer Department for their efforts to provide the best environment for teaching students.

.

Abstract

Intrusion detection systems (IDS) effectively complement other security mechanisms by detecting malicious activities on a computer or network, and their development is evolving at an extraordinary rate. The anomaly-based IDS, which uses learning algorithms, allows detection of unknown attacks. Unfortunately, the major challenge of this approach is to minimize false alarms while maximizing intrusion detection and accuracy rates. To overcome this problem, a hybrid learning approach is proposed through the combination of feature selecting techniques and K-Means clustering and Naïve Bayes classification. Feature selection techniques choose the most important feature and remove redundant and irrelevant features. K-Means clustering is used to cluster all data into the corresponding group based on data behavior, malicious and non-malicious. While the Naïve Bayes classifier is used to classify clustered data into correct categories, i.e. R2L, U2R, Probe, DoS and Normal. Experiments have been carried out to evaluate the performance of the proposed approach using 10%KDD Cup ’99 dataset. The results showed that proposed hybrid model significantly improves the accuracy, detection rate up to 94.06% and 99.49%, respectively with BestFirst and GreedyStepwise Search Method, while decreasing false alarms to 0.15%.

الملخص

يعتبر نظام كشف التسلل (IDS) مكمل وبشكل فعال لاليات امن المعلومات الأخرى وذلك من خلال دوره في اكتشاف الأنشطة الضارة والبرمجيات الخبيثة سواء كانت تستهدف جهاز كمبيوتر أو الشبكة ، والتي اصبحث تتطور وتتزايد بمعدل غير مسبوق ، وبالتالي فقد سمح لنظام كشف التسلل (IDS) القائم على اكتشاف الشذوذ في سلوك حزم البيانات والذي يستخدم خوارزميات التعلم باكتشاف اي هجمات او برمجيات خبيثة غير معروفة، لسوء الحظ ، يتمثل التحدي الرئيسي لهذا النهج في امكانية خفض نسبة الإنذارات الكاذبة و زيادة معدلات الكشف والدقة في نفس الوقت ، وللتغلب على هذه المشكلة تم اقتراح نهجًا هجينًا قادر على التعلم من خلال الجمع بين تقنيات اختيار الميزات وخوارزمية التجميع K-Means وخوارزمية التصنيف Naïve Bayes ، حيث تقوم تقنيات اختيار الميزة باختيار الميزات الأكثر أهمية وإزالة الميزات الزائدة وغير ذات الصلة، وتقوم خوارزمية التجميع-K Meansبتجميع كل البيانات ذات السلوك المتشابه في مجموعات مشابهة ،وعادا تنقسم الى مجموعة لحزم البيانات التي بها سلوك مشابه للبرمجيات خبيثة ومجموعة اخرى لحزم البيانات الطبيعية ، بينما تقوم خوارزمية التصنيف Naïve Bayes بتصنيف مجموعات البيانات إلى فئات صحيحة ، مثل R2L و U2R و Probe و DoS و Normal ، قمنا باجراء التجارب على النهج المقترح باستخدام مجموعة بيانات10٪ KDD Cup '99وقد أظهرت النتائج أن النهج المقترح يحسن بشكل كبير الدقة ومعدل الكشف حتى 94.06٪ و 99.49٪ على التوالي باستخدام طريقتي البحث BestFirst و GreedyStepwise ، بينما يقلل الإنذارات الكاذبة إلى 0.15٪.

Table of Content

No	Title	Page
	ACKNOWLEDGEMENT	I
	ABSTRACT	Ii
	الملخص	Iii
	List Of Figures	iv
	List Of Tables	Ix
	List Of Abbreviations	iv
Chapter One	Introduction	1
1.1	Overview	1
1.2	Problem Statement	2
1.3	The Goal	2
1.4	The Objectives	3
1.5	Scope and Limitation of the Study	3
1.6	Significance of the study	3
1.7	Thesis Outline	4
Chapter Tow	Background of Intrusion Detection Systems	5
2.1	Introduction	5
2. 2	Network Security	5
2.3	Intrusion Detection Systems	6
2.3.1	Classification of Intrusion Detection Systems	6
2.3.1.1	Classification Based on the Source of the Data	6
2.3.1.2	Classification Based on Detection Methods	7
2.4	Attacks Overview	8
2.5	Data Mining	9
2.6	Machine Learning	10
2.6.1	Unsupervised Learning	11
2.6.1.1	Clustering	11
2.6.1.1.1	Clustering Methods	11
2.6.1.1.2	K-Means Clustering Algorithm	12
2.6.2	Supervised Learning	13
2.6.2.1	Classification	14
2.6.2.1.1	Naïve Bayes Classification	14
2.7	Evaluation of Machine Learning Algorithms	15
2.7.1	Holdout Method and Random Subsampling	15

No	Title	Page
2.7.2	Random Subsampling	16
2.7.3	Cross-Validation Method	16
2.8	Overfitting and Under-fitting with Machine Learning Algorithms	16
2.9	Metrics to Evaluate Machine Learning Algorithms	16
2.9.1	Confusion Matrix	17
2.9.3	Sensitivity and Specificity	17
2.9.4	Precision and Recall	18
2.10	Feature Engineering	18
2.10.1	Feature Engineering Techniques	18
2.11	Feature Selection	20
2.11.1	General Procedure of Feature Selection	20
2.11.2	Feature Selection Method	21
2.11.2 .1	Correlation-Based Feature Selection	22
2.11.2.2	Information Gain	22
2.11.2.2	Information Gain	22
2.11.2.3	Gain Ratio	23
2.12	Intrusion Detection System Dataset	23
2.12.1	DARPA 1998 Dataset for Intrusion Detection Evaluation	23
2.12.2	KDD Cup 1999 Dataset Description	24
2.13	WEKA	25
2.14	Related Work	25
Chapter Three	Research Methodology	30
3.1	Introduction	30
3.2	Choice of Dataset	30
3.3	Choice of Data Mining Software	31
3.4	Choice of Data Mining Algorithms	31
3.5	Trial and Error Methodology	33
3.6	Research Methodology	34
3.6.1	Data Preprocessing Phase	35
3.6.2	Research Experiment - Stage (1):	36
3.6.3	Research Experiment - Stage (2):	37
3.6.4	Research Experiment - Stage (3):	38
3.7	Evaluation Metrics	38
3.8.	Dataset Validity	38

No	Title	Page
3.9	Dataset Reliability	39
3.10	Dataset Objectivity	39
Chapter Four	Experimental Results and Evaluation	41
4.1	Introduction	41
4.2	Experimental Setup	42
4.2.1	Experimental Environments and Tools	41
4.2.2	Experimental Measurements	42
4.2.3	KDD'99 Fragmentation	42
4.3	Data Preprocessing	44
4.4	Research Experiment - Stage (1)	51
4.4.1	Random Seeds and Sum of Squared Errors	51
4.4.2	Comparison of Evaluation Result between K-mean algorithm, Naive Bayes algorithm as single operator and hybrid proposed approach	52
4.5	Research Experiment - Stage (2)	53
4.5.1	Attribute Selection	53
4.5.2	A Technical Perspective	53
4.5.3	Correlation-Based Feature Selection (CFS)	54
4.5.4	Information Gain Attribute Evaluation	58
4.5.5	Gain Ratio Attribute Evaluator	59
4.5.6	Correlation Attribute Evaluation	60
4.6	Implementation and Result Discussion	61
4.6.1	The features and the most important features to detect attacks	61
4.6.2	Visual investigation for Distribution histograms of all features	64
4.7	Research Experiment - Stage (3)	67
4.7.1	The Evaluation Result of K-Mean clustering with feature selection techniques	67
4.7.2	Evaluation Result of Single Naïve Bayes with feature selection techniques	68
4.7.3	Evaluation Result of K-mean with Naïve Bayes classifier KM+NB as hybrid model with feature selection techniques	69
4.8	Diagnosis of Overfitting Phenomenon	84
4.9	Evaluation findings and discussion	86
Chapter Five	Conclusions and Recommendations	90
5.1	Conclusions	90
5.2	Recommendations	90

No	Title	Page
	References	92
	Appendix A	95

Table of Figures

No	Title	Page
Figure 2.1	CIA triad model	5
Figure 2.2	Classification of intrusion detection systems	6
Figure 2.3	Data mining as a step in the process of knowledge discovery.	10
Figure 2.4	Feature selection	21
Figure 2.5	Feature selection methods	22
Figure 3.1	Methodology diagram	35
Figure 4.1	Script to transform and replace attribute values in WEKA	47
Figure 4.2	Data from Original dataset	48
Figure 4.3	Original dataset after conversion	48

List Of Tables

No	Title	Page
Table 2.1	A confusion matrix whose size is 2 x 2	17
Table 4.1	Number of Samples in KDD CUP 99 Data Sets	43
Table 4.2	Categorization of attributes with four labels.	43
Table 4.3	Class wise detail of KDD Cup99 data set attributes.	44
Table 4.4	Mapping of Attack Class with Attack Type	46
Table 4.5	Numeric values of (Training and Test set) Features	47
Table 4.6	Classes and References	47
Table 4.7	Minimum, maximum, and distinct values of some features of 10% KDD99	49
Table 4.8	Description of redundancy in (10% KDD Cup'99 subset)	49
Table 4.9	Description of redundancy in (KDD corrected Test subset)	49
Table 4.10	Sample dataset after redundancy removal	50
Table 4.11	shown trial and error methodology for proper selection of the initial seed	52
Table 4.12	shown the result of Comparison between Clustering, classification and hybrid proposed approach	52
Table 4.13	Selected Features Obtained by CFS + BestFirst {forward} technique to distinguish between Normal Network Traffic and Attacks	55
Table 4.14	Selected Features Obtained by CFS + GeneticSearch technique to distinguish between Normal Network Traffic and Attacks	56
Table 4.15	Selected Features Obtained by CFS + GreedyStepwise technique to distinguish between Normal Network Traffic and Attacks	57
Table 4.16	Selected Features Obtained by CFS + RankSearch technique to distinguish between Normal Network Traffic and Attacks.	58
Table 4.17	Selected Features Obtained by InfoGainAttributeEval + Ranker technique to distinguish between Normal Network Traffic and Attacks.	59
Table 4.18	Selected Features Obtained by GainnRatio + Ranker technique to distinguish between Normal Network Traffic and Attacks.	60
Table 4.19	Selected Features Obtained by CorrelationAttributeEval+ Ranker technique to distinguish between Normal Network Traffic and Attacks.	61
Table 4.20	shown the results for each feature selection technique with category contribution	62

No	Title	Page
Table 4.21	Summary of clustering accuracy, Detection Rate, False Alarm using K-means on dataset before and after feature selection Rate	68
Table 4.22	Summary of clustering accuracy, Detection Rate, False Alarm Rate using Naïve Bayes on dataset before and after feature selection	69
Table 4.23	Evaluation Parameters For hybrid model (KM+NB) on test data set with Correlation-Based Feature Selection and BestFirst as Search Method	70
Table 4.24	Classification result for K-Means with Naïve Bayes and CfsSubsetEval + Best First as proposed hybrid approach using Test data set	70
Table 4.25	shown the results across Binary category classes obtained from Hybrid approach with CfsSubsetEval+ BestFirst for the Normal and Attacks classes using Test data set	71
Table 4.26	. Shown Accuracy, Detection Rate and False Alarm Rate Results of Hybrid approach classifier with CfsSubsetEval+ BestFirst in	71
Table 4.27	Evaluation Summary for Single Naïve Bayes, Hybrid Model (K-mean+ Naïve Bayes) on Testing Dataset with CfsSubsetEval+BestFirst	71
Table 4.28	Evaluation Parameters For hybrid model (KM+NB) on test data set with Correlation-Based Feature Selection and GeneticSearch as Search Method	72
Table 4.29	Classification result for K-Means with Naïve Bayes and CfsSubsetEval + GeneticSearch as proposed hybrid approach using Test data set	72
Table 4.30	shown the results across Binary category classes obtained from Hybrid approach with CfsSubsetEval+ GeneticSearch for the Normal and Attacks classes using Test data set	72
Table 4.31	. Shown Accuracy, Detection Rate and False Alarm Rate Results of Hybrid approach classifier with CfsSubsetEval+ GeneticSearch in Corrected KDD Testset	73
Table 4.32	Evaluation Summary for Single Naïve Bayes, Hybrid Model (K-mean+ Naïve Bayes)on Testing Dataset with CfsSubsetEval+	73
Table 4.33	Evaluation Parameters For hybrid model (KM+NB) on test data set with Correlation-Based Feature Selection and GreedyStepwise as Search Method	74
Table 4.34	Classification result for K-Means with Naïve Bayes and CfsSubsetEval + GreedyStepwise as proposed hybrid approach using Test data set	74

No	Title	Page
Table 4.35	shown the results across Binary category classes obtained from Hybrid approach with CfsSubsetEval+ GreedyStepwise for the Normal and Attacks classes using Test data set	74
Table 4.36	Shown Accuracy, Detection Rate and False Alarm Rate Results of Hybrid approach classifier with CfsSubsetEval+ GreedyStepwise in Corrected KDD Testset	75
Table 4.37	Evaluation Summary for Single Naïve Bayes, Hybrid Model (K-mean+ Naïve Bayes) on Testing Dataset withCfsSubsetEval+ GreedyStepwise	75
Table 4.38	Evaluation Parameters For hybrid model (KM+NB) on test data set with Correlation-Based Feature Selection and RankSearch as Search Method	76
Table 4.39	Classification result for K-Means with Naïve Bayes and CfsSubsetEval + RankSearch as proposed hybrid approachusing Test data set	76
Table 4.40	shown the results across Binary category classes obtained from Hybrid approach with CfsSubsetEval+ RankSearch for the Normal and Attacks classes using Test data set	76
Table 4.41	. Shown Accuracy, Detection Rate and False Alarm Rate Results of Hybrid approach classifier with CfsSubsetEval+ RankSearch in Corrected KDD Testset	76
Table 4.42	Evaluation Summary for Single Naïve Bayes, Hybrid Model (K-mean+ Naïve Bayes)on Testing Dataset with CfsSubsetEval+RankSearch	77
Table 4.43	Evaluation Parameters For hybrid model (KM+NB) on test data set with Information Gain Attribute Evaluation and Ranker as Search Method	78
Table 4.44	Classification result for K-Means with Naïve Bayes and InfoGainAttributeEval + Ranker as proposed hybrid approach using Test data set	78
Table 4.45	shown the results across Binary category classes obtained from Hybrid approach with InfoGainAttributeEval + Ranker for the Normal and Attacks classes using Test data set	78
Table 4.46	shown Accuracy, Detection Rate and False Alarm Rate Results of Hybrid approach classifier with InfoGainAttributeEval + Ranker in Corrected KDD Testset	79
Table 4.47	Evaluation Summary for Single Naïve Bayes, Hybrid Model (K-mean+ Naïve Bayes) on Testing Dataset with InfoGainAttributeEval + Ranker	79
Table 4.48	Evaluation Parameters For hybrid model (KM+NB) on test data set with Gain Ratio Attribute Evaluator and Ranker as Search Method	80

No	Title	Page
Table 4.49	Classification result for K-Means with Naïve Bayes and GainRatioAttributeEval + Ranker as proposed hybrid approach using Test data set	80
Table 4.50	shown the results across Binary category classes obtained from Hybrid approach with GainRatioAttributeEval + Ranker for the Normal and Attacks classes using Test data set	80
Table 4.51	.shown Accuracy, Detection Rate and False Alarm Rate Results of Hybrid approach classifier with GainRatioAttributeEval + Ranker in Corrected KDD Testset	81
Table 4.52	Evaluation Summary for Single Naïve Bayes, Hybrid Model (K-mean+ Naïve Bayes) on Testing Dataset with GainRatioAttributeEval + Ranker	81
Table 4.53	Evaluation Parameters For hybrid model (KM+NB) on test data set with Correlation Attribute Evaluation and Ranker as Search Method	82
Table 4.54	Classification result for K-Means with Naïve Bayes and CorrelationAttributeEval + Ranker as proposed hybrid approach using Test data set	82
Table 4.55	shown the results across Binary category classes obtained from Hybrid approach with CorrelationAttributeEval + Ranker for the Normal and Attacks classes using Test data set	83
Table 4.56	.shown Accuracy, Detection Rate and False Alarm Rate Results of Hybrid approach classifier with CorrelationAttributeEval + Ranker in Corrected KDD Testset	83
Table 4.57	Evaluation Summary for Single Naïve Bayes, Hybrid Model (K-mean+ Naïve Bayes) on Testing Dataset with CorrelationAttributeEval + Ranker	84
Table 4.58	shown the Training Accuracy and validation dataset Accuracy for integration all feature selections with Naïve Bayes classifier	85
Table 4.59	Shown the Training Accuracy and validation dataset Accuracy for integration all feature selections with Naïve Bayes classifier and K-means.	86
Table 4.60	shown the Comparison of evaluation results accuracy, detection rate and False alarm rate for integration all feature selections with single Naïve Bayes classifier, single K-means clustering and Hybrid Model KM+NB	89

List of Abbreviations

AFRL	Air Force Research Laboratory
AR	Attribute Ratio
CFS	Correlation-based Feature Selection
DARPA	Defense Advanced Research Projects Agency
DOS	Denial of Service
DTM	Decision Table Majority
FFSA	Forward Feature Selection Algorithm
FN	False Negatives
FP	False Positives
GA	Genetic Algorithm
GNU	General Public License
GR	Gain Ratio
HIDS	Host Based Intrusion Detection System
IDS	Intrusion detection system
IG	Information Gain
IP	Internet Protocol
KDD	the knowledge discovery process in databases
LCFS	Linear Correlation Feature Selection
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NIDS	Network Based Intrusion Detection System
R2L	Remote to Local
SVM	Support Vector Machine
TCP	Transmission Control Protocol
TN	True Negatives
TP	True Positives
U2R	User to Root
UDP	User Datagram Protocol
WEKA	Waikato Environment for Knowledge Analysis