An Investigation of Scalable Anomaly Detection Techniques for a Large Network of Wi-Fi Hotspots

Pheeha Machaka[1] and Antoine Bagula[2]

[1]Council for Scientific and Industrial Research, Modelling and Digital Science, Meiring Naude Rd, Pretoria, 0184, South Africa
University of the Western Cape, Robert Sobukwe Road, Bellville, 7535, South Africa
[1]pmachaka@csir.co.za, [1]bbagula@uwc.ac.za

Abstract. The paper seeks to investigate the use of scalable machine learning techniques to address anomaly detection problem in a large Wi-Fi network. This was in the efforts of achieving a highly scalable preemptive monitoring tool for wireless networks. The Neural Networks, Bayesian Networks and Artificial Immune Systems were used for this experiment. Using a set of data extracted from a live network of Wi-Fi hotspots managed by an ISP; we integrated algorithms into a data collection system to detect anomalous performance over several test case scenarios. The results are revealed and discussed in terms of both anomaly performance and statistical significance.

Keywords: Performance Monitoring, Neural Networks, Artificial Immune Systems, Bayesian Networks, Anomaly Performance Detection, Multilayer Perceptron, Naive Bayes, AIRS2.

1 Introduction

Wireless Fidelity (Wi-Fi) is a wireless networking technology that uses radio waves to provide high-speed wireless internet connections. Wi-Fi is based on the IEEE 802.11 standards and builds upon a fast, easy and inexpensive networking approach [1] where Access Points (APs) are used to broadcast signals to Wi-Fi-capable client devices (laptops and Smartphone devices) within their range, and connect to the Internet.

Performance monitoring is an important task upon which large Wi-Fi network deployment depends. As traditionally implemented, performance monitoring is based on a reactive network approach where the operating system software only warns the network administrators when a problem occurs. This approach leads to both the halting of important network processes and the hampering of critical business processes of the organization.

Pre-emptive network monitoring provides the potential to prevent the occurrence of faults by analyzing the status of the network components to create a fail-safe network status or allow a smooth migration from a faulty to fail-safe network status. Wi-Fi technology has become so popular and this lead to large scale deployment of thousands of hotspots networks. These hotspots generate huge amounts of monitoring

data, thus there is a call for efficient data handling methods that would analyze data and recognize anomalous hidden patterns and implement fault tolerance mechanism. While statistical analysis methods have been deployed in many cases to address this issue, soft computing methods borrowed from the human immune system are emerging as powerful tools used in anomaly detection and security monitoring systems.

1.1 Related Work

There has been work done in the field of anomaly detection, and in this paper, three soft computing methods were identified, viz. Artificial Neural Networks, Artificial Immune Systems and Bayesian Networks. With Artificial Neural Network (ANN), the work has focused on employing ANN for anomaly detection on network traffic data [2-4]. Artificial Immune Systems (AIS) was used for intrusion detection, and detection of computer viruses [5-7].Bayesian Networks were also used for anomaly detection for disease outbreak [4] and also in detecting and analyzing anomaly behavior in network-based FTP services [8].

The three machine learning techniques have gained success in anomaly detection and in this paper; we would like to employ them on a large network of Wi-Fi hotspots for intrusion detection. The work done in this paper furthers the work by authors in [9-11] and the efforts to find out which method works best for large data networks, and how each methods performs under network intrusion for the test cases set out in this paper.

1. Which method performs better for monitoring a large Wi-Fi network?
2. How do these methods perform under different test cases and network thresholds?

The remainder of the paper is organized as follows: in section 2, the machine learning techniques used in this article are briefly described. Section 3 will describe the research and experiment design. Section 4 will reveal and discuss the experiment results while section 5 brings the article to a conclusion.

2 Algorithms

Artificial Neural Networks - ANN's are mathematical or computational models that get their inspiration from biological neural systems. In this paper the neural network model, Multilayer Perceptron (MLP) was used to conduct experiments. The MLP is a feed forward neural network model in which vertices are arranged in layers. MLP have one or more layer(s) of hidden nodes, which are not directly connected to the input and output nodes [12]. For the purpose of this experiment we employed Weka's Multilayer Perceptron implementation.

Bayesian Networks - Bayesian Networks can be described briefly as acyclic directed graph (DAG) which defines a factorisation of a joint probability distribution over the variables that are represented by the nodes of the DAG, where the factorisation is given by the direct links of the DAG [13]. The NaiveBayes algorithm was used for the experiments. It makes a strong assumption that all attributes of the

examples are independent of each other given the context of the class. The Weka's NaiveBayes implements this probabilistic Naïve Bayes classifier [14].

Artificial Immune Systems - The AIS takes inspiration from the robust and powerful capabilities of the Human Immune System's (HIS) capabilities to distinguish between self and non-self [7].The Algorithm employed in this paper's experiments is the Weka's Artificial Immune Recognition System (AIRS) learning algorithm [15].The AIRS is a supervised AIS learning algorithm that has shown significant success on a broad range of classification problems [5-7].

3 The Research Design

The section that follows will describe the methods and techniques used to carry out the research presented in this paper.

3.1 The Wi-Fi Network

The experiment network connected more than 400 hotspots around the Cape Town area, with more than 615 Cisco WRT54GL gateway devices connected to the network. For data collection and monitoring, a Syslog daemon program was installed on each gateway device, and was left to run 2-3 months collecting monitoring data at every hour's interval.

3.2 Network Performance Monitoring

The network was monitored based on three performance metrics. This includes:

- Uptime and Downtime (%) - This metric measures the availability, stability and reliability of the communication device when used in the network.
- Load Average (%) - Measures the "congestion rate" for the device based on the number of users connected to the device.
- Radio Noise (in dB) - Wi-Fi uses the shared 2.4 GHz spectrum band and the proliferation of devices using the spectrum leads to congestion and noisy Wi-Fi devices.
- Standard deviation - To detect aberrant behavior in performance, statistical confidence bands were used to measure deviations in a time series. A deviation depends on the Delta (δ) parameter whose sensible values were taken between 0 and 3.
- Encoding and Selection - Three levels of performance were used to describe performance. A 3-bit encoded nominal value was used to describe performance. This type of encoding was also used by authors in [11].

3.3 Performance Evaluation Techniques

The effectiveness of the methods is evaluated based on their ability to make correct predictions. The following measures were used to quantify the performance of the algorithms:

- True Positive (TP) rate, also known as detection rate.
- False Positive (FP) rate, also known as false alarm rate.
- F-measure, it is a harmonic mean for precision and recall.
- Kappa Statistic - used to measure the agreement between predicted and observed categorization of a dataset, while correcting for agreement that occurs by chance.

3.4 Test Cases

For the test cases in this study, we followed a model suggested by the authors in [92-Wu Shelly]. We conducted experiments using four test case scenarios revealing Wi-Fi operating constraints from loose (e.g. rural setting where QoS is an issue) to the most stringent (e.g. Suburban setting where modern applications demand higher QoS).

The Weka machine learning software was used for the experiments, a stratified 10-fold cross-validation technique was used for training and testing the algorithms.

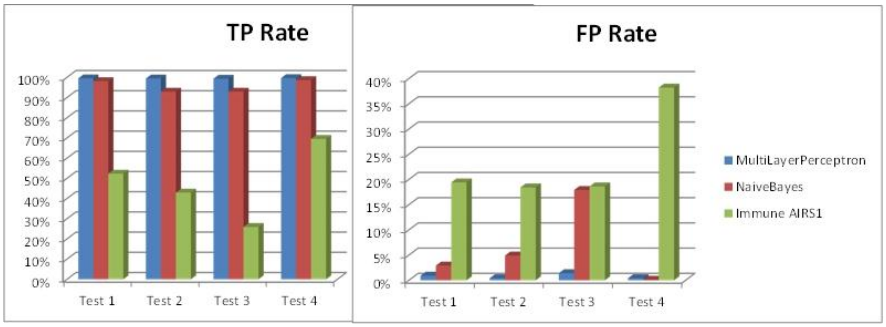

Fig. 1. Anomaly TP Rate and FP Rate performance

4 Results and Discussions

Using the test cases and methods described above; the experiments were conducted and results were revealed based on the algorithms': True Positive (TP), False Positive (FP), Kappa Statistic and F-measure performance. A graphical and t-test performance evaluation is used.

4.1 True Positive Rate Performance

In the bar graph representation of figure 1 above indicates a bar graph representation of TP rate and that the MLP had an average TP rate of 99.45%, while NaiveBayes had an average TP rate of 95.62% across all test cases. The AIRS1 algorithm's performance was lower in recognising classes correctly with average TP rate of 47.65%.

Table 1. Results for True Positive Rate T-test for Paired Two Samples for Means

	Multilayer Perceptron	NaiveBayes
Mean	0.9810	0.9563
Variance	0.0001	0.0007
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	13.8587	
P(T<=t) two tailed	0.0000	

	Multilayer Perceptron	AIRS2
Mean	0.9810	0.4519
Variance	0.0001	0.0524
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	32.9818	
P(T<=t) two tailed	0.0000	

For the test:

True Positive Rate: H_0: $\mu_{MLP} - \mu_{NaiveBayes} = 0$ (1)
True Positive Rate: H_1: $\mu_{MLP} - \mu_{NaiveBayes} \neq 0$

In table 1, the value of the t-Statistic is 13.858 and its two-tailed p-value is 5.147E-31. At the 5% confidence level, the test is highly significant and there is overwhelming evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly True Positive Rate for the MLP and NaiveBayes algorithms.

For the test:

True Positive Rate: H_0: $\mu_{MLP} - \mu_{AIRS2} = 0$ (2)
True Positive Rate: H_1: $\mu_{MLP} - \mu_{AIRS2} \neq 0$

The value of the t-Statistic is 32.981 and its two-tailed p-value is 1.34191E-82. At the 5% confidence level, the test is highly significant and there is overwhelming evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly True Positive Rate for the MLP and AIRS2 algorithms.

4.2. False Positive Rate Performance

The bar graph representation in figure 1 indicates that the MLP and NaiveBayes had very low average FP rate, 0.77% and 6.45% respectively. A poor performance was seen with AIRS1 technique; it had an average FP rate of 23.65%, and had a high FP rate of 38.2% in test case 4.

For the test:

False Positive Rate: H_0: $\mu_{MLP} - \mu_{NaiveBayes} = 0$ (3)
False Positive Rate: H_1: $\mu_{MLP} - \mu_{NaiveBayes} \neq 0$

In table 2 below, the value of the t-Statistic is -2.188 and its two-tailed p-value is 0.0298. At the 5% confidence level, the test is significant and there is strong evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly False Positive Rate for the MLP and NaiveBayes algorithms.

Table 2. Results for False Positive Rate T-test for Paired Two Samples for Means

	Multilayer Perceptron	NaiveBayes
Mean	0.0231	0.0301
Variance	0.0009	0.0012
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	-2.1882	
P(T<=t) two tailed	0.0298	

	Multilayer Perceptron	AIRS2
Mean	0.0231	0.3507
Variance	0.0009	0.0407
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	-22.4821	
P(T<=t) two tailed	0.0000	

For the test:

False Positive Rate: H_0: $\mu_{MLP} - \mu_{AIRS2} = 0$ (4)
False Positive Rate: H_1: $\mu_{MLP} - \mu_{AIRS2} \neq 0$

The value of the t-Statistic is -22.482 and its two-tailed p-value is 1.57884E-56. At the 5% confidence level, the test is highly significant and there is overwhelming evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly False Positive Rate for the MLP and AIRS2 algorithms.

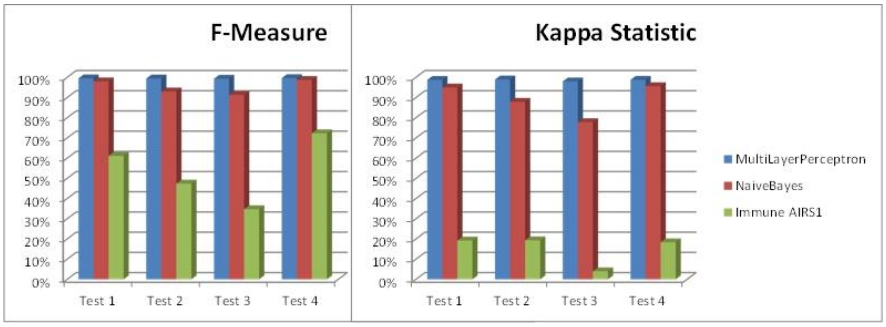

Fig. 2. Anomaly F-measure and Kappa Statistic performance

4.3 F-Measure Performance

The MLP is shown to be, on average, the most accurate of the techniques with an average F-measure of 99.45%. The NaiveBayes had an average F-measure of 95.25%

across all test cases. This is indicated by the bar graph representation in figure 2. AIRS1 revealed poor results with an average F-measure of 53.88%.

Table 3. Results for F-Measure T-test for Paired Two Samples for Means

	Multilayer Perceptron	NaiveBayes
Mean	0.9804	0.9528
Variance	0.0001	0.0010
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	12.8538	
P(T<=t) two tailed	1.9720	

	Multilayer Perceptron	AIRS2
Mean	0.9804	0.4851
Variance	0.0001	0.0482
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	32.1631	
P(T<=t) two tailed	0.0000	

For the test:

$$\text{F-Measure: } H_0: \mu_{MLP} - \mu_{NaiveBayes} = 0 \qquad (5)$$
$$\text{F-Measure: } H_1: \mu_{MLP} - \mu_{NaiveBayes} \neq 0$$

The value of the t-Statistic is 12.853 and its two-tailed p-value is 6.32577E-28. At the 5% confidence level, the test is highly significant and there is overwhelming evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly F-measure for the MLP and NaiveBayes algorithms.

For the test:

$$\text{F-Measure: } H_0: \mu_{MLP} - \mu_{AIRS2} = 0 \qquad (6)$$
$$\text{F-Measure: } H_1: \mu_{MLP} - \mu_{AIRS2} \neq 0$$

The value of the t-Statistic is 32.163 and its two-tailed p-value is 9.08911E-81. At the 5% confidence level, the test is highly significant and there is overwhelming evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly F-measure for the MLP and AIRS2 algorithms.

4.4 Kappa Statistic Performance

Indicated by the bar graph in figure 2, the MLP and NaiveBayes had an average Kappa statistic of 98.59% and 89.05%, respectively. AIRS1 technique had an average Kappa statistic of 15.22%, revealing poor accuracy and precision.

Table 4. Results for Kappa Statistic T-test for Paired Two Samples for Means

	Multilayer Perceptron	NaiveBayes
Mean	0.9512	0.8906
Variance	0.0005	0.0052
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	12.5824	
P(T<=t) two tailed	1.9720	

	Multilayer Perceptron	AIRS2
Mean	0.9512	0.1551
Variance	0.0005	0.0157
Observations	200	200
Hypothesised Mean Difference	0	
t Stat	91.0005	
P(T<=t) two tailed	0.0000	

For the test:

$$\text{Kappa Statistic: } H_0: \mu_{MLP} - \mu_{NaiveBayes} = 0 \quad (7)$$
$$\text{Kappa Statistic: } H_1: \mu_{MLP} - \mu_{NaiveBayes} \neq 0$$

The value of the t-Statistic is 12.58 and its two-tailed p-value is 4.29687E-27. At the 5% confidence level, the test is highly significant and there is overwhelming evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly Kappa Statistic for the MLP and NaiveBayes algorithms.

For the test:

$$\text{Kappa Statistic: } H_0: \mu_{MLP} - \mu_{AIRS2} = 0 \quad (8)$$
$$\text{Kappa Statistic: } H_1: \mu_{MLP} - \mu_{AIRS2} \neq 0$$

The value of the t-Statistic is 91.00 and its two-tailed p-value is 4.14E-164. At 5% confidence level, the test is highly significant and there is overwhelming evidence to infer that the alternative hypothesis is true. Therefore we reject the null hypothesis and conclude that there is a difference in the mean anomaly Kappa Statistic for the MLP and AIRS2 algorithms.

5 Conclusions

The statistical hypothesis test experiments, tables 1 to 4, that were conducted for anomaly performance detection reveal that, in all algorithm performance measures, there is a significant mean difference among the three algorithms. One can safely conclude that there was a significant difference in mean performance measures for MLP, NaiveBayes and the AIRS2 algorithms.

The bar chart representations in figure 1 and 2 were carefully examined, and for all performance measures, the MLP had an overall good performance and came out with the highest (above 90%) algorithm performance measures. The NaiveBayes also had a good performance that was slightly lower than that of the MLP. On the other hand, the AIRS2 had a poor performance relative to the MLP and NaiveBayes.

When applying the algorithms to a large Wi-Fi networking problem, the MLP would be a better option as it would produce more accurate results. The NaiveBayes would also produce good results, but not better than that of the MLP. On the other

hand, the AIRS2 algorithm may produce mediocre performance results on a large Wi-Fi network monitoring problem.

6 References

[1] S. J. Vaughan-Nichols. The challenge of wi-fi roaming. *Computer 36(7),* pp. 17-19. 2003.

[2] J. Cannady. Artificial neural networks for misuse detection. Presented at National Information Systems Security Conference. 1998, .

[3] E. Cheng, H. Jin, Z. Han and J. Sun. "Network-based anomaly detection using an elman network," in *Networking and Mobile Computing*Anonymous 2005, .

[4] J. Zhang and M. Zulkernine. Anomaly based network intrusion detection with unsupervised outlier detection. Presented at Communications, 2006. ICC'06. IEEE International Conference On. 2006, .

[5] S. Forrest, S. A. Hofmeyr, A. Somayaji and T. A. Longstaff. A sense of self for unix processes. Presented at Security and Privacy, 1996. Proceedings., 1996 IEEE Symposium On. 1996, .

[6] D. Dasgupta and F. González. An immunity-based technique to characterize intrusions in computer networks. *Evolutionary Computation, IEEE Transactions On 6(3),* pp. 281-291. 2002.

[7] K. Luther, R. Bye, T. Alpcan, A. Muller and S. Albayrak. A cooperative AIS framework for intrusion detection. Presented at Communications, 2007. ICC'07. IEEE International Conference On. 2007, .

[8] B. Cha and D. Lee. Network-based anomaly intrusion detection improvement by bayesian network and indirect relation. Presented at Knowledge-Based Intelligent Information and Engineering Systems. 2007, .

[9] P. Machaka, A. Bagula and N. De Wet. A highly scalable monitoring tool for wi-fi networks. Presented at Wireless Systems (IDAACS-SWS), 2012 IEEE 1st International Symposium On. 2012, .

[10] P. Machaka and A. Bagula. "Preemptive performance monitoring of a large network of wi-fi hotspots: An artificial immune system," in *Wired/Wireless Internet Communications*Anonymous 2011, .

[11] P. Machaka, T. Mabande and A. Bagula. "Monitoring of a large wi-fi hotspots network: Performance investigation of soft computing techniques," in *Bio-Inspired Models of Networks, Information, and Computing Systems*Anonymous 2012, .

[12] R. A. Dunne. *A Statistical Approach to Neural Networks for Pattern Recognition* 2007*702*.

[13] U. B. Kjræulff and A. L. Madsen. *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis: A Guide to Construction and Analysis* 201*222*.

[14] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques* 2005.

[15] A. Watkins, J. Timmis and L. Boggess. Artificial immune recognition system (AIRS): An immune-inspired supervised learning algorithm. *Genetic Programming and Evolvable Machines 5(3),* pp. 291-317. 2004.