Proceedings of IOE Graduate Conference, 2019-Winter
Peer Reviewed
Year: 2019 Month: December Volume: 7
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Anomaly Detection in Wireless Sensor Network using Inverse Weight Clustering and C5.0 Decision Tree

Pramod Kumar Chaudhary [a], Arun Kumar Timalsina [b]

[a] *Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal*
Corresponding Email: [a] cpramodkumar10@gmail.com, [b] t.arun@ioe.edu.np

Abstract
Wireless Sensor Network is a network of integrated sensors for environmental sensing, data processing and communication with other sensors and the base station. At the same time WSNs are vulnerable to security breaches, attacks and information leakage. Anomaly detection techniques are used to detect such activities over the network that does not conform to the normal behavior of the network communication.
In this paper, two machine learning algorithms Iinverse Weight Clustering (IWC) and C5.0 decision tree for classifying anomalous and normal activities in a wireless sensor network have been used. The IWC clustering method is first used to partition the training instances into k clusters using Euclidean distance similarity. On each cluster, representing a density region of normal or anomaly instances, decision trees are built using C5.0 decision tree algorithm. The decision tree on each cluster refines the decision boundaries by learning the subgroups within the cluster. The experiment carried out on three different datasets (University of North Carolina Greensboro(UNCG), Intel Berkeley Research Lab(IBRL) and Bharatpur Airport WSN); the results indicated that proposed method achieved detection rate of 98.9% at false alarm-rate of 0.31% on IBRL; detection rate of 99.57% at false alarm-rate of 0.35% on Bharatpur Airport.

Keywords
WSN, Anomaly detection, IWC clustering, C5.0 decision tree

1. Introduction

Wireless sensor networks (WSNs) have become a popular area of research in recent years due to their huge potential to be used in various applications. The important applications of WSN include environmental monitoring, personal healthcare, enemy monitoring and so on [1]. WSNs are highly vulnerable to attacks, due to their open and distributed nature and limited resources of the sensor nodes [2].

The information obtained from the WSNs has to be accurate and complete. Inaccurate or incomplete data measurements of WSN are often known as WSN anomalies [1]. Anomaly detection systems (ADS) monitor the behaviour of a system and flag significant deviations from the normal activity as anomalies [3].

In this paper spatiotemporal dataset have been used, it is a temperature and humidity related data which contain different temperature and humidity values collected by sensor nodes for an area, and tried to find out Contextual anomalies in the dataset. Here a combination of Inverse Weighted Clustering (IWC) with C5.0 decision tree algorithms have been used as a model where IWC is used to cluster sample data into groups, label them and then used a C5.0 classifier to train and test the model to distinguish between the normal and abnormal behaviour.

2. Related Works

The researchers used different techniques to detect anomaly such as statistical-based, machine learning-based, and data mining-based approaches. R. M. Elbasiony, T. E. Eltobely and et al. (2013)[4], used weighted K-means and Random Forest classification, the experiment worked very well except that KDD CUP99 dataset was used and the results were 98.3% Detection Rate and 1.6% false alarm rate.

M. Wazid and A. K. Das (2016)[5], proposed a robust and efficient secure intrusion detection approach which uses the K-means clustering in order to extend the lifetime of a WSN. The authors assess the

approach over a WSN dataset that is created using Opnet modeler, which contains a range of attributes, such as end-to-end delay, traffic sent and traffic received. Authors claim that proposed scheme achieves 98.6% detection rate and 1.2% false positive rate and the proposed technique has the ability to detect two types of malicious nodes: blackhole and misdirection nodes.

Y. Li and J. Xia (2012)[6], proposed an efficient Intrusion detection system based on Support vector machines and gradually feature removal method, combination of clustering method, ant colony algorithm and support vector machine.

V. Golmah (2014) [7], proposed an efficient hybrid intrusion detection method based on C5.0 and SVM. This method achieves a better performance compared to the individual SVM. Evaluated the proposed method using DARPA dataset. W. Yassin, N. I. Udzir and et al. (2013)[8], proposed integrated machine algorithms and Naïve Bayes to minimize false alarm rate and improve accuracy rate. The results show significant improvement in accuracy rate with 99.0% when compared with previous studies with the same approach. However, false alarm rate was high at 2.2%.

In H. M. Tahir, W. Hassan and et al. (2015)[9], K-means clustering algorithms was combined with support vector machine to formed hybrid intelligent system, the Authors were able to obtain 96.24% accuracy and 3.715% alarm rate. K. H. Rao, G. Srinivas and et al. (2011)[10] proposed a technique by cascading K-means with different classification techniques, this removes the anomalies from K-means using id3, it overcome the disadvantage of both ID3 and K-means but integrating K-means +id3 is a time consuming process.

P. C. Yong, C. Xiang and et al. (2008)[11], proposed a multiple-level hybrid classifier, a novel intrusion detection system, which combines the supervised tree classifiers and unsupervised Bayesian clustering to detect intrusion. This approach provides the high detection rate and false alarm rate in comparison of Kernel miner, Three level tree classifier, Bagged boosted C5.0 trees. G. Kim and S. Lee (2014)[12], presented a new hybrid intrusion detection method hierarchically integrates a misuse detection and anomaly detection in a decomposed structure. The misuse detection model is built based on C4.5 decision tree algorithm and is used to decompose the normal training data into smaller subsets. The one-class SVM is used to create anomaly detection for the decomposed region.

J. Wang, Q. Yang and et al. (2009)[13], presented an intrusion detection system based on decision tree technology. In the process of constructing intrusion rules, information gain ratio is used in place of information gain. The experiment results show that the C4.5 decision tree is feasible and effective, and has a high accuracy rate. His experimental study gives almost 90% of classifier accuracy.

A. P. Muniyandi, R. Rajeshwori and et al. (2012)[14], presents an anomaly detection method using K-Means+C4.5, a method to cascade k-means clustering and C4.5 decision tree methods. This method achieves better performance in comparison to the K-Means, ID3, Naïve Bayes, K-NN, and SVM. In short, various techniques have been proposed in the field of intrusion detection, but there is still room to improve detection rate and accuracy, and reducing false alarm rate. In contrast, the proposed model has been tested on the WSN dataset of IBRL and Bharatpur Airport to demonstrate that it is able to increase detection rate while minimizing the false alarm rate. In this paper, Inverse Weighted clustering and C5.0 have been used. The reason for choosing Inverse Weighted Clustering is that it removes the initial selection of data as seen in traditional K-means and the reason of choosing C5.0 is it is more efficient, its decision tree is smaller in cooperation with C4.5.

3. Proposed Method

Anomaly Detection in WSN datasets using IWC+C5.0

In this paper an anomaly detection model using two machine learning algorithms IWC and C5.0 have been proposed. Initially IWC was used for partitioning the dataset into K closest cluster using Euclidean distance formula and then C5.0 techniques was applied on each closest cluster to built decision tree for each cluster and classify the each instance into normal or anomaly using decision tree result. There are two modules in IWC+C5.0; namely the pre-classification module and the classification module. The first module, involving Inverse Weighted Clustering iteration function where similar data are grouped into several clusters based on their behavior. The entire data are labeled with the K-th clusters set accordingly. Next, the labeled clustered data are classified into abnormal and normal classes using the Decision tree classifier to recover

the misclassified data from the first module. It was found that IWC+C5.0 is able to classify the abnormal and normal data more accurately at the subsequent classification module.

Inverse Weighted Clustering (IWC)

IWC in essence is built upon K-means algorithm. However, it relies on running the k-means many times until the centroids and clusters become stable. In other words, while the k-means stops once k centroids and clusters are formulated, IWC takes the resulting centroids and rerun k-means over the same data by computing the distance between each record and the centroids [15].

In this paper, K=2 have been predefined for WSN data representing Cluster 1 and Cluster 2. Certain activities or data are alike to either normal or abnormal behaviour. The IWC algorithm is unable to differentiate this behaviour precisely. Thus, Decision tree classifier applied to re-classify clustered labeled data to improve the shortcoming.

C5.0 Decision tree

C5.0 supports sampling and cross-validation. C5.0 models are quite robust in the presence of problems such as missing data and large numbers of input fields. It does not require long training times to estimate. In addition, it is easier to understand than some other model types, since the rules derived from the model have a very straight forward interpretation. C5.0 tree or rule sets are usually smaller than C4.5 [16]. C5.0 is faster than C4.5. Memory usage is more efficient in C5.0 than C4.5. The C5.0 rule sets have lower error rates on unseen cases. So comparing with C4.5 the accuracy of result is good with C5.0 algorithm.

A C5.0 model is based on the information theory. Decision trees are built by calculating the information gain ratio.

It is based on the entropy measure commonly used in information theory [17].

Let T is the training dataset

$X(c)$ is the class I where $c = 1,2,3,\ldots,n$

$$\mathrm{I}(T_1,T_2,\ldots,T_n) = -\sqrt{p_c \log_2(p_c)} \tag{1}$$

T_c is the number of samples in c

$$P_c = \frac{T_c}{T} \tag{2}$$

let attribute A has v distant values Entropy=E(A) is

$$\sum \frac{T_{1j}+T_{2j}+\cdots+T_{nj}}{T} \times \mathrm{I}(T_1,T_2,\ldots,T_n) \qquad j=1 \tag{3}$$

Where T_{cj} is the sample in class c and subset j of attribute

$$\mathrm{I}(T_{1j},T_{2j},\ldots,T_{nj}) = -\sum p_{cj} \log_2(p_{cj}) \tag{4}$$

$$\mathrm{Gain}(A) = \mathrm{I}(T_1,T_2,\ldots,T_n) – \mathrm{E}(A) \tag{5}$$

C5.0 decision tree algorithm [18]:

Step 1: The C5.0 generates a either a decision tree or a rule set

Step 2: Pick the most informative attribute

Step 3: Find the partition with the highest information gain using Eq (5)

Step 4: At each resulting node, repeat step 1 and 2

4. Experimental Results

Dataset Description

To see the performance of proposed method, IBRL and Bharatpur Airport WSN datasets were used. The dataset have 11 attributes but only 2 attributes was taken for classification of instance into normal and abnormal, IWC applied on the dataset for portioning the data into K clusters , here K=2, number of iteration =10 was taken. After running IWC algorithm on input WSN data of Intel lab and Bharatpur Airport datasets, the entire data was grouped into two classes, one in red belonging to class-1 and the other in green belonging to class-2 for labeling the data. These two values represent the cluster to which each reading (i.e. temperature and humidity) belongs.

On IBRL datasets

The IBRL datasets contain information collected from 54 sensors deployed in the IBRL, between 28 February and 5April 2004. Mica2Dot sensors with weatherboards collected time-stamped topology information, along with humidity, temperature, light, and voltage values once every 31 s[19]. Here, a portion of datasets with 7526 reading instances of two attributes (i.e. humidity and temperature) were taken for the experimental purpose.

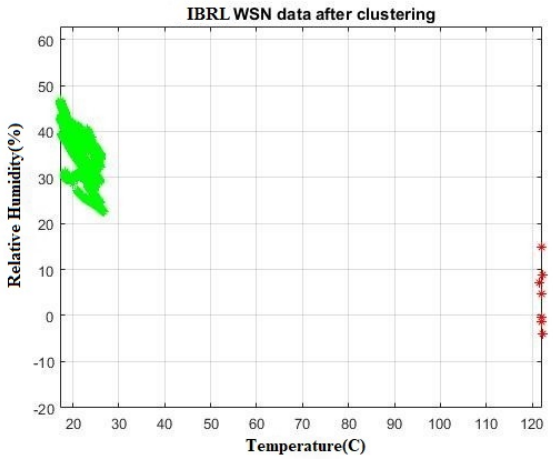

Figure 1: Intel lab WSN data after clustering

On Bharatpur Airport datasets

The datasets consist of humidity and temperature measurements collected during 1-month period at intervals of 15 m. The data were collected during January 2016. Here also, a portion of datasets was utilized for experimental purpose. Here 164 anomalies (spatiotemporal data) were manually added to 2880 readings of Bharatpur Airport WSN data thus making 3044 input readings to make the input data somehow labeled and qualified for the research purpose. Intel lab used the same way to manipulate WSN input data by manually adding records with anomalies in the dataset. The manually introduced anomaly in Bharatpur Airport data were spatiotemporal data to find out the contextual anomalies.

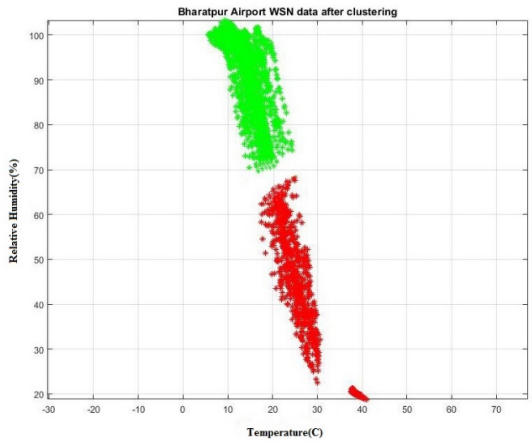

Figure 2: Bharatpur Airport WSN data after clustering

Figures 1 and Figure 2 show the results of the data after clustering humidity and temperature readings. Before constructing the decision tree, data was split into two groups constituting 66% of the training group and 34% of the testing data of the clustered data.

Decision tree was trained using labeled training dataset consisting of 66% of input dataset records and then tested against 34% of testing dataset, but without giving it the cluster ID that shows to which cluster each record in the testing dataset belongs.

The classification phase was performed using C5.0 decision tree after applying the entire pre-processing step. The resulting predictions were compared with the cluster ID of each record in the labeled testing dataset. C5.0 Decision tree divides the network behavior to normal and abnormal and assigns the abnormal behavior to its specific category.

5. Results and Discussion

The confusion matrix was realized from the classification of the proposed model using synthetic labeled UNCG [20], real Intel Berkeley Research Laboratory (IBRL) and Bharatpur Airport WSN datasets. Table 1, 2 and 3 show the confusion matrix obtained in testing the proposed approach.

Table 1: Results on UNCG dataset [4690 records] with C5.0

X = 1407	Predicted anomalies	Predicted normality
Actual anomalies	TP = 16	FN = 1
Actual normality	FP = 2	TN = 1388

Table 2: Results on Intel lab dataset [7526 records] with IWC + C5.0

X = 2535	Predicted anomalies	Predicted normality
Actual anomalies	TP = 583	FN = 1
Actual normality	FP = 6	TN = 1945

Table 3: Results on Bharatpur Airport dataset [3044 records] with IWC+C5.0

X = 1035	Predicted anomalies	Predicted normality
Actual anomalies	TP = 462	FN = 1
Actual normality	FP = 2	TN = 570

From the confusion matrix Tables, the results obtained clearly indicated a high rate of detection.

Table 4: Performance Evaluation formulas

Metric	Formula
Accuracy	(TP+TN) / (TP+TN+FP+FN)
Detection Rate	(TP) / (TP+FP)
False Alarm Rate	(FP) / (FP+TN)

The performance evaluation of the proposed approach was carried out by comparing the result of the proposed approach and different hybrid intelligent approaches.

Table 5: Performance Evaluation averaged over 5 trials for 2 attributes

Classifier Algorithms	Datasets	Performance Measures in %			
		Accuracy	False Alarm	Detection Rate	F–Measure
C5.0	UNCG	0.9979	0.0014	0.8889	0.9143
IWC+ C5.0	IBRL	0.9972	0.0031	0.9898	0.994
	Bharatpur Airport	0.9971	0.0035	0.9957	0.9966

As in Table 5, anomaly detection in two attributes (i.e. temperature and humidity) on labeled UNCG WSN datasets with C5.0, and on both IBRL and Bharatpur Airport WSN data with combined IWC+C5.0 whose detection rate and false alarm show that when two of the best classifier is combined, the detection rate exceeds 98% with very low false alarm rate below 1%.

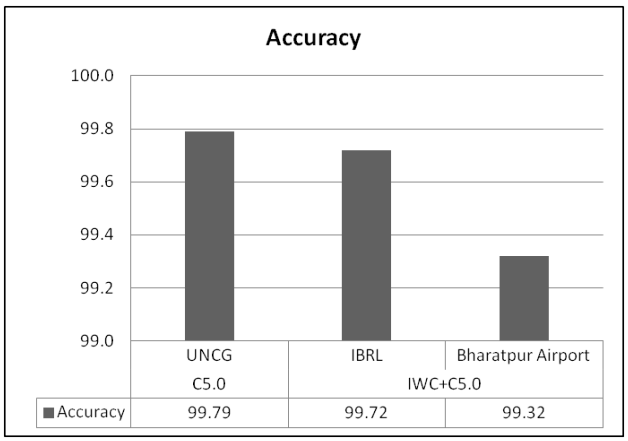

Figure 3: Accuracy on different datasets

As in Figure 3, we can see the accuracy of C5.0 on labeled UNCG datasets is 99.79%, accuracy of integrated IWC+C5.0 on Intel Berkeley Research Lab dataset is 99.72% and on Bharatpur airport WSN datasets is 99.71% respectively.

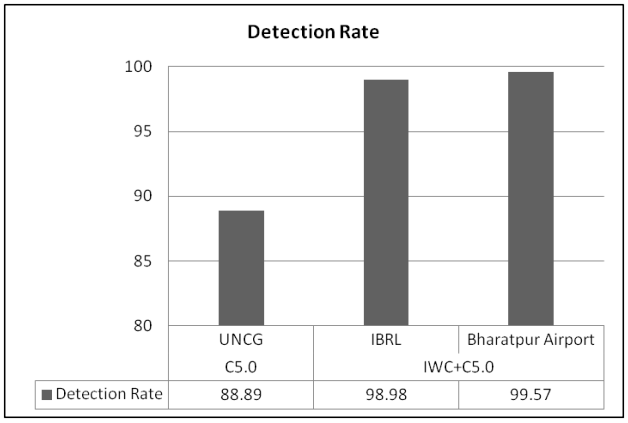

Figure 4: Detection rate on different datasets

In Figure 4, we can see the detection rate with C5.0 on labeled UNCG dataset is 88.89%, detection rate with IWC+C5.0 on IBRL datasets is 98.98% and on Bharatpur Airport WSN datasets is 99.57% respectively.

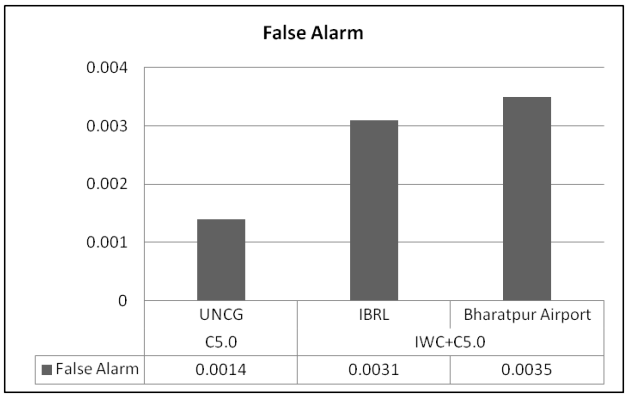

Figure 5: False Alarm results on different datasets

In Figure 5, we can see the false alarm rate with C5.0 on labeled UNCG data is 0.14% and with integrated IWC+C5.0 on IBRL is 0.31% and on Bharatpur Airport WSN is 0.35% respectively.

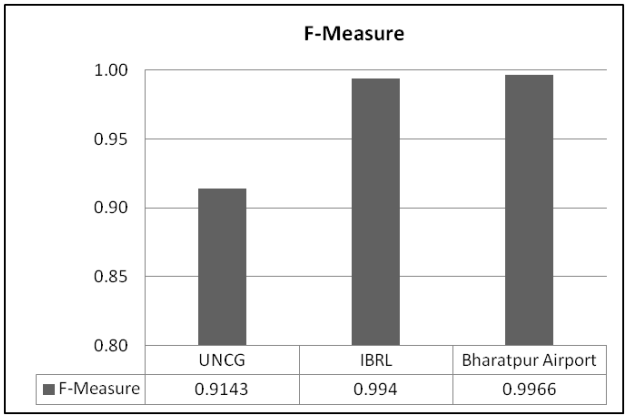

Figure 6: F-Measures on different datasets

Table 6: Different Approaches Vs. Proposed IWC+C5.0 Approach

Datasets	Techniques	Accuracy Rate	Detection Rate	Alarm Rate
IBRL	IWC + Naïve Bayes	98.20%	98.20%	1.80%
	IWC + KNN	98.64%	98.60%	1.36%
	IWC + SVM	99.38%	99.40%	0.62%
	IWC + C5.0	99.72%	98.90%	0.31%

In Figure 6, we can see F-Measures with C5.0 on labeled UNCG data (0.9143) and with integrated IWC+C5.0 on IBRL (0.994) and on Bharatpur Airport WSN is 0.9966 respectively.

Comparisons

The proposed approach was compared with four some of the hybrid intelligent approaches for WSN anomaly detection.

Table 6 gives the percentage of detection rate, accuracy and false alarm rate. The detection rate for proposed method is 98.98 which is greater than IWC+Naïve Bayes (98.2), IWC+KNN (98.6) except IWC+SVM (99.4). Similarly the accuracy for proposed algorithm is 99.72 and is greater than Naïve Bayes (98.2), KNN (98.64) and SVM (99.37). Also false alarm rate is 0.31 which is lesser than that of Naïve Bayes (1.79), KNN (1.35) and SVM (0.62).

6. Conclusions

The proposed model integrated IWC algorithm for clustering jobs and C5.0 decision tree for prediction jobs for anomaly detection in WSNs and tested in terms of accuracy, True positive (TP), True negative (TN), False positive (FP), False negative (FN), F-Measure. The Experimental results show that overall performance of the proposed approach improved in terms of detection rate and low false alarms rate. The result findings show that proposed IWC+C5.0 is the efficient technique for detecting anomaly on Intel Berkeley Research lab wireless sensor network, with a higher rate of accuracy (99.72%), and lower false alarm rate (0.31%).

Future work is to improve the accuracy by combining different clustering algorithms such as Hierarchical clustering, Adaptive resonance (ART) Neural Network and Kohonen's Self-Organizing Maps with Decision tree C5.0. Furthermore, other partitioning algorithms are going to be used instead of IWC to find out whether better results can be achieved.

References

[1] S. S. Bhojannawar, C. M. Bulla, and et. al. Anomaly detection techniques for wireless sensor networks - a survey. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(10), Oct 2013.

[2] Zhen Feng, Jingqi Fu, Dajun Du, Fuqiang Li, and Sizhou Sun. A new approach of anomaly detection in wireless sensor networks using support vector data description. *International Journal of Distributed Sensor Networks*, 13(1), 2017.

[3] ME Elhamahmy, Hesham N Elmahdy, and Imane A Saroit. A new approach for evaluating intrusion detection system. *International Journal of Artificial Intelligent Systems and Machine Learning*, 2(11):290–298, 2010.

[4] Reda M Elbasiony, Elsayed A Sallam, Tarek E Eltobely, and Mahmoud M Fahmy. A hybrid network intrusion detection framework based on random forests and weighted k-means. *Ain Shams Engineering Journal*, 4(4):753–762, 2013.

[5] Mohammad Wazid and Ashok Kumar Das. An efficient hybrid anomaly detection scheme using k-means clustering for wireless sensor networks. *Wireless Personal Communications*, 90(4):1971–2000, 2016.

[6] Yinhui Li, Jingbo Xia, Silan Zhang, Jiakai Yan, Xiaochuan Ai, and Kuobin Dai. An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Systems with Applications*, 39(1):424–430, 2012.

[7] Vahid Golmah. An efficient hybrid intrusion detection system based on c5. 0 and svm. *International Journal of Database Theory and Application*, 7(2):59–70, 2014.

[8] Warusia Yassin, Nur Izura Udzir, Zaiton Muda, Md Nasir Sulaiman, et al. Anomaly-based intrusion detection through k-means clustering and naives bayes classification. In *Proc. 4th Int. Conf. Comput. Informatics, ICOCI*, number 49, pages 298–303, 2013.

[9] Hatim Mohamad Tahir, Wail Hasan, Abas Md Said, Nur Haryani Zakaria, Norliza Katuk, Nur Farzana Kabir, Mohd Hasbullah Omar, Osman Ghazali, and Noor Izzah Yahya. Hybrid machine learning technique for intrusion detection system. 2015.

[10] K Hanumantha Rao, G Srinivas, Ankam Damodhar, and M Vikas Krishna. Implementation of anomaly detection technique using machine learning algorithms. *International Journal of Computer Science and Telecommunications*, 2(3):25–31, 2011.

[11] Cheng Xiang, Png Chin Yong, and Lim Swee Meng. Design of multiple-level hybrid classifier for intrusion detection system using bayesian clustering and decision trees. *Pattern Recognition Letters*, 29(7):918–924, 2008.

[12] Gisung Kim, Seungmin Lee, and Sehun Kim. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Systems with Applications*, 41(4):1690–1700, 2014.

[13] Juan Wang, Qiren Yang, and Dasen Ren. An intrusion detection algorithm based on decision tree technology. In *2009 Asia-Pacific Conference on Information Processing*, volume 2, pages 333–335. IEEE, 2009.

[14] Amuthan Prabakar Muniyandi, R Rajeswari, and R Rajaram. Network anomaly detection by cascading k-means clustering and c4. 5 decision tree algorithm. *Procedia Engineering*, 30:174–182, 2012.

[15] Wesam Barbakh and Colin Fyfe. Inverse weighted clustering algorithm. *Inverse weighted clustering algorithm*, 11(2), 2007.

[16] Meesala Shobha Rani and S Basil Xavier. A hybrid intrusion detection system based on c5. 0 decision tree and one-class svm. *International journal of current engineering and technology*, 5(3):2001–2007, 2015.

[17] Lior Rokach and Oded Maimon. Decision trees. In *Data mining and knowledge discovery handbook*, pages 165–192. Springer, 2005.

[18] [Online] `http://db.csail.mit.edu/labdata/data.txt.gz`.

[19] [Online] `http://www.uncg.edu/cmp/downloads/`.