DOCUMENT

paper

SCORE

78 of 100

ISSUES FOUND IN THIS TEXT

62

PLAGIARISM

11%

Contextual Spelling 3

Unknown Words	2
Misspelled Words	1

Grammar 4

Determiner Use (a/an/the/this, etc.)	2
Incorrect Noun Number	1
Wrong or Missing Prepositions	1

Punctuation 10

Punctuation in Compound/Complex Sentences	6
Comma Misuse within Clauses	4

Sentence Structure 1

Misplaced Words or Phrases	1

Style 44

Passive Voice Misuse	38
Unclear Reference	4
Inappropriate Colloquialisms	1
Improper Formatting	1

Vocabulary enhancement

No errors

paper

Wireless Network Intrusion Detection using
K-Means Clustering Algorithm

V. Akshay – 2015503005
J. Timothy Jones Thomas – 2015503561
S. Vikraman – 2015503564

Abstract -This project is to create an Intrusion Detection System (IDS). An Intrusion Detection System is an application that monitors the network for malicious activities and unauthorized access to device information as well as personal data. The Intrusion Detection System is based [1] on the k-means clustering algorithm. This algorithm partitions n observations into k clusters, [2] in which each observation [3] belongs to the cluster [4] with the nearest mean, thus serving as a prototype for that cluster [5]. The initial data set is partitioned [6] into such clusters [8] [7] and by making use of these, the cluster [9] to which the test data belongs can be predicted [10]. Based on this predicted [11] cluster [12], the application notifies whether the test data is normal [13] or suspicious. If a new or unknown type of attack takes place, the application [14] will find the data to be deviant from the rest and will mark it as suspicious. The initial training data space is obtained [15] from the KDDCUP99 dataset.

Keywords – intrusion detection system; network security; k-means clustering; packet sniffing;

I. INTRODUCTION

With growing susceptibility to attacks, user data is prone to huge [16] risks. Hence, network security is of paramount importance. Valuable resources having network access should be permanently protected [17] from all attempts to destroy, expose, alter, disable, steal or gain unauthorized access and/or [18] usage. Resources confidentiality, integrity [20] and availability [21] have to remain intact.

1 Passive voice
2 Unoriginal text: 8 words
userpages.umbc.edu/~gobbert/papers/...
3 Repetitive word: *observation*
4 Repetitive word: *cluster*
5 Repetitive word: *cluster*
6 Passive voice
7 Repetitive word: *clusters*
8 [clusters,]
9 Repetitive word: *cluster*
10 Passive voice
11 Repetitive word: *predicted*
12 Repetitive word: *cluster*
13 Overused word: *normal*
14 Repetitive word: *application*
15 Passive voice
16 Overused word: *huge*
17 Passive voice

Intrusion detection system (IDS) is a system specially designed to detect such malicious attempts. As traditional 19 IDS's 22 are mainly signature-based, detecting only known attacks, their biggest problem is the inability to detect 23 new or variant attacks. One topic that intuitively stands out as a potential solution for 25 solving this problem is 24 k-means clustering.

18 ~~and/or~~ → and
19 Unoriginal text: 24 words
pdfs.semanticscholar.org/aaaa/95a4c0...
20 integrity,
21 availability,
22 ~~IDS's~~ → IDS
23 Repetitive word: *detect*
24 Unoriginal text: 15 words
pdfs.semanticscholar.org/aaaa/95a4c0...
25 Possibly confused preposition

II. RELATED WORKS

Neural Network (NN) is the most popular AI algorithm used for intrusion detection compared to other algorithms. However, training these networks takes a lot of time to achieve a reasonable level of performance, and also their adaptability is unsatisfactory [1]. Recent IDSs based on Naïve Bayes and Decision Trees seem promising, with better accuracy and performance. Genetic Algorithms and Support Vector Machines (SVM) are also being used 26, though it has been stipulated 27 that the accuracy of SVMs are on the lower side [1].

Bisyron 28 Wahyudi Masduki [2] used the following machine learning algorithms - KNN, SVM and Dempster Shafer theory. Firstly, a few features from KDDCUP are selected 29 as training data. KNN and SVM are used to classify attack data 30 and the output of those two different methods is combined 31 with Dempster Shafer Theory. The performance of the classification process is good 32 in overall, but the results are not optimal to detect R2L and U2R attacks categories.

Yanjie Zhaossss [3] studied classifier model on a training dataset which has very different class distribution. They proposed PN rule, a two-stage general-to-specific framework of learning a rule-based model. This method can detect only 10.7% of attacks in the attack class R2L despite a lot of false alarms generated. A real shortage of this method 33 is that the rule is determined automatically,

26 Passive voice
27 Passive voice
28 ~~Bisyron~~ → Byron
29 Passive voice
30 data,
31 Passive voice
32 Overused word: *good*

which makes it dependent on the dataset.

33 Repetitive word: *method*

Lofti Mhamdi [4] used Multilayer Perceptron(MLP) which is a supervised learning algorithm based on the feed-forward neural network with one or more layers between input and output layer. 34 Tuan A Tang [4] used MLP for anomaly detection, where the proposed model is a single hidden layer neural network. However, the proposed classifier achieved only 7.32% detection and 2.5% false alarm rates for the R2L attacks.

34 Unoriginal text: 22 words
pdfs.semanticscholar.org/aaaa/95a4c0...

III. PROPOSED WORK

Initially, the knowledge base is created 35 by using pre-existing datasets (KDDCUP99) to form clusters. This clustering is done by the k-means clustering algorithm 36. Each cluster 37 represents the different types of network access.

Fig 1: Topology of a network with transfer of malicious packets

A sample network is shown 38 in Fig. 1, where the device that sends the malicious packets and the device 39 that receives them, are on the same network 40. The receiver has an Intrusion Detection System, which scans the packets that are received, fetches the required parameters, and sends it to the classifier. If the classifier deems the packet 41 as suspicious, the user is notified 42. Else, the process is repeated 43 for the next packet. The architecture for the Intrusion Detection System is represented 44 in Fig. 2.

35 Passive voice

36 Passive voice

37 Repetitive word: *cluster*

38 Passive voice

Fig. 2: Architecture for the intrusion detection system

39 Repetitive word: *device*

40 Repetitive word: *network*

IV. ALGORITHM AND MATHEMATICS INVOLVED

k-means [45] clustering is used to model the knowledge base. Each observation belongs to the cluster with the nearest mean. These clusters [46] are plotted [47] on a plane. If an observation [48] is plotted [50] [49] near a cluster [51], it is assumed to belong to that cluster [52]. Hence, these clusters [53] should be placed [54] in such way that there is no ambiguity in association [55]. To achieve this [56], it is better to place [57] the clusters [58] as far as possible from each other.

We develop cluster models based on the training data set for multiple values of k, which represents the number of clusters in the model. The k value for which the variance of clusters [59] is minimum is chosen [60] as the best model. This [61] is equivalent to minimizing the pair-wise squared deviations of points in the same cluster [62]. This [63] is also equivalent to maximizing the squared deviations [64] between points [65] in different clusters [66]. This [67] is done [68] by using the training dataset once again.

Each Observation is a tuple with n values, and the ith value in the tuple is represented [69] as obs[i]. The deviation or distance between two observations in the cluster model is given [70] as,

deviation=i=0nobs1i-obs2i212

To find which cluster a given observation belongs to, the model calculates the deviations between the cluster centroids and the given observation [71]. The observation [72] is said to belong to the cluster [73] to which it is nearest [74] i.e., the [75] deviation [76] is minimum. This [77] is the basis for classification.

By finding the deviation between each record in the training dataset and the centroid of the cluster to which it belongs, we get the variance for that record [78]. Each record [79] is a tuple consisting of n values.

variance=i=0ncentroidi-inputi212

41 Repetitive word: *packet*
42 Passive voice
43 Passive voice
44 Passive voice
45 ~~k-means~~ → K-Means
46 Repetitive word: *clusters*
47 Passive voice
48 Repetitive word: *observation*
49 Passive voice
50 Repetitive word: *plotted*
51 Repetitive word: *cluster*
52 Repetitive word: *cluster*
53 Repetitive word: *clusters*
54 Passive voice
55 the association or an association
56 Dangling modifier
57 Repetitive word: *place*
58 Repetitive word: *clusters*
59 Repetitive word: *clusters*
60 Passive voice
61 Unclear antecedent
62 Repetitive word: *cluster*
63 Unclear antecedent
64 Repetitive word: *deviations*
65 Repetitive word: *points*
66 Repetitive word: *clusters*
67 Unclear antecedent
68 Passive voice
69 Passive voice
70 Passive voice

The mean of the variances, for all N records of the training set, is taken as the clustering score.

Clustering score=1N(i=0Nvariance[i])

Once the clustering scores are calculated [80] for multiple k values, the model with the lowest score is taken [81] as the best model. This model is used [82] for classifying the test data. The cluster to which most of the normal [83] data is mapped, [84] is found. If the test data is mapped [86] [85] to any other cluster [87], it is marked [88] as suspicious. Else, it is marked [89] as normal [90].

While evaluating IDS, for every possible test value there are two kinds of error: false positive (FP) and false negative (FN). FP occurs when an event is predicted [92] as [91] normal [94] [93] but it is, in fact, intrusive, while FN occurs [95] when a normal [96] event occurs without being recognized as one. On the other hand, true [98] positive (TP) measures the proportion of actual positives which are correctly identified as such, while true [99] negative (TN) measures the proportion [100] of negatives [101] which are correctly identified [103] [102] as such. The performance of [97] the classifier can be quantified using the detection rate (DR) and overall accuracy (OA) measures [1]. DR shows the percentage of the true [105] intrusions that have been successfully detected [106]:

DR [104]=TPTP+FN×100

OA is calculated [108] as the total number of correctly classified intrusions divided by the total number of observations:

OA=TP [107]+TNTP+TN+FP+FN×100

V. IMPLEMENTATION

The Intrusion Detection System is developed using Python

71 Repetitive word: *observation*
72 Repetitive word: *observation*
73 Repetitive word: *cluster*
74 [nearest,]
75 [i.e., ~~the~~ → i.e.; the]
76 Repetitive word: *deviation*
77 Unclear antecedent
78 Repetitive word: *record*
79 Repetitive word: *record*
80 Passive voice
81 Passive voice
82 Passive voice
83 Overused word: *normal*
84 [mapped,]
85 Passive voice
86 Repetitive word: *mapped*
87 Repetitive word: *cluster*
88 Passive voice
89 Passive voice
90 Overused word: *normal*
91 Unoriginal text: 29 words
pdfs.semanticscholar.org/aaaa/95a4c0...
92 Passive voice
93 Overused word: *normal*
94 [normal,]
95 Repetitive word: *occurs*
96 Overused word: *normal*
97 Unoriginal text: 36 words
pdfs.semanticscholar.org/aaaa/95a4c0...
98

programming language. The Apache Spark (pyspark [109]) Machine Learning library(mllib [110]) is imported [111], which contains the implementation of the k-means clustering algorithm. The following inbuilt functions were used [112]: KMeans.train(), KMeansModel.predict(). The train() function returns a KMeansModel object which is used [113] for prediction. The predict() function returns the predicted cluster index.

VI. RESULTS AND PERFORMANCE ANALYSIS

The Clustering scores for different values of k are shown in Table 1 and Graph 1. From the values [114], it can be deduced [115] that k = 70 gives the lowest score. Hence it is chosen for prediction, though higher values of k can give lower scores.

Table 1: Clustering Scores for different values of k

Value of k

Clustering Score

10

1.041418

20

0.771934

30

0.662832

40

0.619724

50

0.580507

60

0.533794

70

0.410462

80

0.523469

Graph 1: Clustering Scores

The test data set is given as input [116] and the Detection Rate and Overall Accuracy are calculated [117]. Table 2 displays

99 Overused word: *true*

100 Repetitive word: *proportion*

101 Repetitive word: *negatives*

102 Passive voice

103 Repetitive word: *identified*

104 Unoriginal text: 14 words
pdfs.semanticscholar.org/aaaa/95a4c0...

105 Overused word: *true*

106 Passive voice

107 Unoriginal text: 20 words
pdfs.semanticscholar.org/aaaa/95a4c0...

108 Passive voice

109 Unknown word: *pyspark*

110 Unknown word: *mllib*

111 Passive voice

112 Passive voice

113 Passive voice

114 Repetitive word: *values*

115 Passive voice

the results of running the test data set.

Table 2: Results of Prediction

Type of Prediction

Number of Observations

True Positive

50003

False Positive

9804

True Negatives

240632

False Positives

10590

Based on the Table 118 2, the Detection rate (DR) is calculated to be 82.52% 119 and the Overall Accuracy is 93.44%.

VII. CONCLUSION

Data security is a major 120 concern for everyone. In this paper, we discussed the k-means clustering algorithm conceptually, for classifying the test values as malicious or normal 121. This algorithm can be integrated into an application in a real-time system to monitor the network. In this way, network security can be ensured 122.

116 [input,]

117 Passive voice

118 [~~the~~ Table]

119 [82.52%,]

120 Overused word: *major*

121 Overused word: *normal*

122 Passive voice