International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

DCNN: The Density, Cluster Centers and Nearest Neighbors using Intrusion Detection Algorithm

M Lavanya, K Munivara Prasad

Abstract: ***Most current intrusion detection system employ signature based methods or data mining based methods which rely on labeled training dat. This training data is typically expensive to produce. Intrusion detection aims to detect intrusion behavior and serves as a complement to firewalls. It can detect attack types of malicious network communications and computer usage that cannot be detected by idiomatic firewalls. Many intrusion detection methods are processed through machine learn- ing. Previous literature has shown that the performance of an intrusion detection method based on hybrid learning or integration approach is superior to that of single learning technology. However, almost no studies focus on how additional representative and concise features can be extracted to process effective intrusion detection among massive and complicated data. In this paper, a new hybrid learning method is proposed on the basis of features such as density, cluster centers, and nearest neighbors (DCNN). In this algorithm, data is represented by the local density of each sample point and the sum of distances from each sample point to cluster centers and to its nearest neighbor. k-NN classifier is adopted to classify the new feature vectors. Our experiment shows that DCNN, which combines K-means, clustering-based density, and k-NN classifier, is effective in intrusion detection.***

Keywords: ***intrusion detection; DCNN; density; cluster center; nearest neighbor, hybrid learning method.***

I. INTRODUCTION

Intrusion detection is a security system that serves as a supplement to firewalls, which defend the computer system against attacks [1].prepared as per journal the template. 3. Contents of the paper are fine and satisfactory. Author (s) can make rectification in the final paper but after the final submission to the journal, rectification is not possible. Intrusion detection detects external intrusions and supervises unauthorized activities of internal users by identifying and responding to malicious network communication and com- puter usage behavior. Intrusion detection aims to detect intrusions by studying the process and characteristics of intrusion behavior, there- by enabling a real-time response to intrusion events and the invasion process. Two basic intrusion detection technologies exist, namely, anomaly detection and misuse detection [2].Currently, most of the relevant literature focuses on intrusion detection based on ma- chine learning and combines different criteria to improve detection performance, such as accuracy, detection rate, and false alarm. Although numerous advanced detection methods have been proposed, only a few studies focus on how simple and relatively large correlation values can be used to represent a large amount of data.

In this paper, a new characteristic value representation method is proposed for intru- sion detection. This new feature vector uses the local density of each sample point in the dataset, the distance from the sample point to the cluster center, and the distance to the near- est neighbor. Thus, the method is known as DCNN. This paper is organized as follows: In the second section, we summarize the related technologies in intrusion detection. The pro- posed DCNN algorithm is detailed in the third section. The experimental setup and the results are provided in the fourth section. Finally, a conclusion is given.

II. LITERATURE REVIEW

At present, intrusion detection technology based on rule matching, machine learning, and Devaraju S et al. used intrusion detection technology based on rule matching[3]. Association rule mining algorithm (ARMA) was used to detect the attack types on KDD Cup 99 datasets. Zhang Yi et al. used data mining technolo- gy[4]. This paper described the use of associa- tion rules and its optimization algorithm. The design and implementation of an intrusion de- tection system were based on feature analysis and knowledge discovery of log files. As mentioned above, intrusion detection based on rule matching fails to recognize unknown malicious behaviors, which can be solved with machine learning. Many studies have been conducted on this subject[5][6][7]. A.S. Eesa et al. used cuttlefish algorithm to obtain a new feature and then used decision tree for classification [5].

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Ms. iM iLavanya, iM.Tech iDept. iof iCSE, iChadalawada iRamanamma iEngineering iCollege, iTirupati, iIndia. ilavanya.machupalli@gmail.com
Mr. iK iMunivara iPrasad, iAssociate iProfessor, iDept. iof iCSE, iChadalawada iRamanamma iEngineering iCollege, iTirupati, iIndia. iprasadkmv27@gmail.com

Tsai CF and Lin CY used intrusion de- tection technology based on hybrid machine learning[7]. First, k-means clustering was per- formed to obtain cluster centers. Then, the tri- angular area related to two cluster centers with one data from the given dataset is calculated, thereby forming a new feature signature of the data. Finally, the k-NN classifier was used to classify similar attacks on the basis of the new feature represented by the triangular area[7]. Lin WC, Ke SW, and Tsai CF also used in- trusion detection technology based on hybrid machine learning[8]. They first used k-means clustering to obtain cluster centers and nearest neighbors, and then they obtained a new fea- ture by calculating the distance between data points and cluster centers or nearest neighbors. Finally, k-NN classifier and SVM were used to classify based on the new feature. They achieved a detection accuracy close to that of the k-NN classification on original features with low cost.

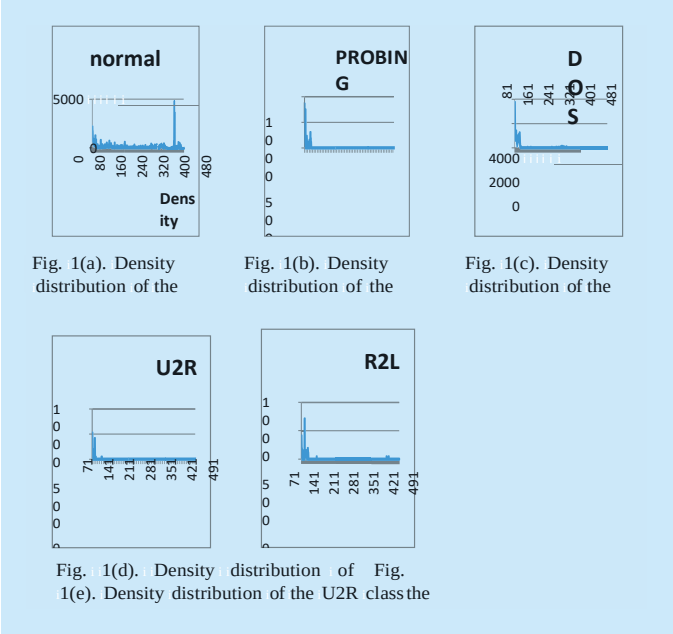

Fig. 1(a). Density distribution of the

Fig. 1(b). Density distribution of the

Fig. 1(c). Density distribution of the

Fig. 1(d). Density distribution of Fig. 1(e). Density distribution of the U2R class the

Fig.1 Density distribution

In this paper, we propose an intrusion de- tection algorithm based on the study of Lin WC et al.[8]. Clustering can be conducted based on measures, such as distance and densi- ty. Which a type of clustering is better depends on the characteristics being addressed. In in- trusion detection, the measure that performs best is uncertain. In other words, the distance cannot be fully representative of the feature. In addition to the distance, the density is a rep- resentative value of the feature. Therefore, we suggest taking density into account as a new representation of the feature.

III. DCNN ALGORITHM

DCNN FRAMEWORK

To illustrate the validity of density, the density distribution of the KDD Cup 99 corpus of normal and abnormal classes are calculated as shown in Figure 1. The definition of local den- sity will be a subject of focus later. Figure 1 indicates that the density distribution of each data type is different. The local density of normal data is most distributed at higher values of approximately 497–500, and the attack data tend to distribute with lower densities. Hence, density can be used as an effective distinguishing feature. The frame of intrusion detection algorithm based on DCNN is shown in Figure 2. Hybrid learning is applied to network data. First, clustering is used to obtain distances and density related to network data to form a new feature vector with low dimension. k-NN classifier is developed based on the new feature vectors, and the label is output the class of data Clustering on training set T aims to obtain the cluster center (Ci), the nearest neighbors of each sample point (Ni), and the local density of each sample point (ρi): Intrusion detection is a classification problem. Therefore, the number of categories in classi- fication should be determined first. Distance is adopted to measure the similarity between the unlabeled sample point and each type of attack data. Hence, the number of clusters should be equal to the number of classification catego- ries. New feature vector formation: The dis- tance between the sample point and all cluster centers (d1), and the distance between the data point and nearest neighbor point in this cluster (d2) are calculated. Then, two distances are added to obtain the new data Di. The training dataset T=(x1,x2,…,xn) is replaced by a new two-dimensional feature vector T'=(D,ρ) formed by distance Di and local density ρi.

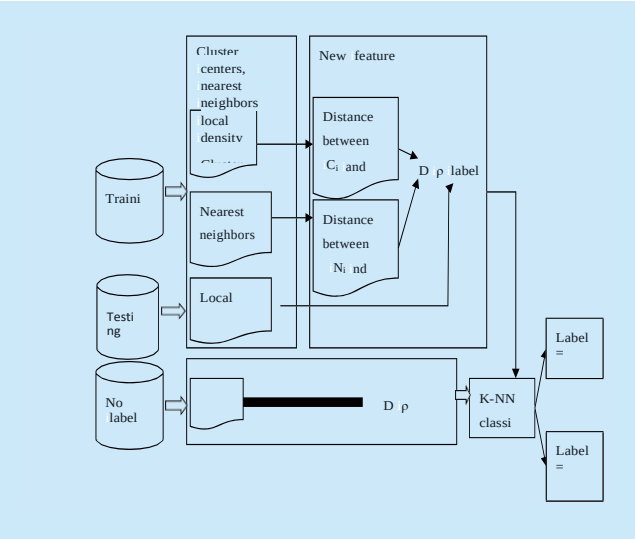

Fig.2 Frame of DCNN

Training and testing: The above step is repeated on test dataset S. A new two-dimen- sional feature vector S' is obtained. Then, k-NN classifier is used for intrusion detection on T' and S' datasets.

IV. EXPERIMENTS

A. Experimental setup

Dataset:

The training and testing datasets used in this paper are all KDD Cup 99 corpus[11]. KDD Cup 99 corpus is the dataset used in the Knowledge Discovery and Data Mining (KDD) contest held in 1999. Although the data are old, they are widely recognized and used by researchers. Each network connection in the KDD Cup 99 dataset is marked as normal or attack. The attacks are divided into 4 categories and 39 species. The four types of attack are DOS, R2L, U2R, and PROBING. Table II describes the classification tags for the five types of data. The KDD Cup 99 dataset has 41 dimen- sional of feature descriptions and one dimen- sion of category label for a total of 42 dimen- sions. Similar to the work of Zhang et al. [11], 19 dimensional characteristics are selected. After taking out 19 dimensional data, quanti- tative data need to be normalized. Afterwards, we need to remove duplicate data to obtain a single dataset. the composition of the remaining data. The training dataset has 119845 data, and the training and testing data- sets have 177463 data.

B. Experimental results Original data classification:

Intrusion detection using k-NN classifier is performed for the original 19 dimensional KDD Cup datasets. Results of the five types of data are shown in Table I. K is set to 21. The total accuracy is 84.36%.

Table-I: Result of k-NN classifier (K=21)

actual\ Predicted	Normal	PROBING	DOS	U2R	R2L	Accuracy
Normal	68741	1852	6124	593	79	88.83%
PROBING	126	1594	21	4	11	90.77%
DOS	8147	1218	30287	0	0	76.38%
U2R	5	1	19	27	0	51.92%
R2L	241	44	254	5	452	45.38%

CANN:

Results of the five data types of CANN are shown in Table II. The K value of the baseline classifier k-NN is set to 21. The total accuracy is 89.79%, thereby indicating that the perfor- mance is better than that of k-NN classification on original features.

Table-II: Result of CANN classifier (K=21)

actual\ Predicted	Normal	PROBING	DOS	U2R	R2L	Accuracy
Normal	69630	240	7079	80	360	89.97%
PROBING	94	1595	60	2	5	90.83%
DOS	3056	793	35549	118	136	89.65%
U2R	10	0	9	33	0	63.16%
R2L	162	2	31	0	801	80.42%

DCNN:

Results of the five data types of DCNN are shown in Table III. The K value of the baseline classifier k-NN is set to 21. The total accuracy is 96.74%, thereby indicating the best perfor- mance.

Table-III: Result of DCNN classifier (K=21)

actual\ Predicted	Normal	PROBING	DOS	U2R	R2L	Accuracy
Normal	76375	896	82	11	25	98.69%
PROBING	42	1711	3	0	0	97.44%
DOS	1992	502	36955	69	134	93.20%
U2R	14	0	3	35	0	67.31%
R2L	67	15	48	4	862	86.55%

Experiments on real data sets:

The detailed information about the real data sets used in this paper. All real data include numerical attributes (Abualigah et al., 2017). In Fig. 7, the accuracy of our proposed method is much more than that of base density methods such as DPC-KNN (Rodriguez & Laio, 2014), DPC-PCA (Du et al., 2016) and NCAR (Pourbahrami et al., 2018), on real data. After normalizing the data, the performance of NCARD algorithm on real data has been presented. NCARD has better result than other algorithms on Pen-based digits, Sonar and Waveform. The accuracy of NCARD is better than the remaining three algorithms in high dimension data sets. NCAR is the second best, followed by DPC-KNN and DPC-PCA. In our implementation, p is the parameter which is used to define the number of neighbours in the first step process in NCARD, NCAR, DPC- KNN and DPC-PCA algorithms. In our algorithm, we select parameter p from a fixed value [5%]. For real data sets, the average run times of NCARD and NCAR are 58 and 47 seconds, respectively.

Experiments iion iiartificial iidata iisets:

We compared our proposed neighborhood construction algorithm with kNN, ε-neighborhood, NC, and NCAR. In kNN value k is of 5% to 10% of points in the data set, where we donate these configurations by kNN1 and kNN2, respectively. The purpose is to make sure that the value of k is at least the size of the smallest cluster. Inkaya et al. procedure is used to set the value of ε and k. The k-distance graph (k=4) is used to determine ε value for each data set. Two groups of data sets (Group1 and Group 2) are used:

Group 1 data sets: The first group includes twenty three artificial data sets which are used in our implementation experiments in order to validate the performance of NCARD algorithm in comparison with other algorithms. Group 1 is composed of 2-dimensional and higher dimensional data sets compiled by Inkaya et al., (2015).

Group 2 data sets: The second group includes twenty four artificial data sets which are used in our implementation experiments in order to validate the performance of NCARD algorithm in comparison with other algorithms. Each data set is composed of four clusters, namely the letters S, A, O and E. The data set D- 000 to D-1211 of Group 2 with 5-18 target points, 2780-2783 data points is used to illustrate the performance of the NCARD algorithm. Group 2 is composed of 2-dimensional and higher dimensional data sets compiled by Inkaya et al., (2015). The results of comparing the accuracy of NCARD clustering algorithm with NCAR, kNN, NC algorithms. According to the results in Table 5, the proposed algorithm works equally well with high and low amounts of data, and it increases the accuracy of NCAR algorithm and removes its weaknesses with high amounts of data. Our proposed method has higher level of accuracy than other geometric methods mentioned above such as NCAR, NC, as well as base distance methods such as kNN and ε-neighborhood on artificial data sets. NC, NCAR, and NCARD have significantly higher JI and RI values in comparison with KNN1, KNN2, and ε-neighborhood. The minimal NCARD performance is much higher than the other algorithms. according to the maximum performance of all algorithms is at least one algorithm in group 1 of data sets in which the target clusters would be obtained. For group 1 the min performance of NCAR and NC is not the best among the six algorithms of competing approaches, instead NCARD wins in this case. In fact, in group 1 NCARD has the best performance. The separation of data sets in the clusters of group 2 is smaller than their compactness. In other words, with a decrease in the distance among the clusters, there would be more cluster mix, which is the major limitation of NCAR and NC. Furthermore, in group 2, KNN1 and KNN2 fail to locate their target clusters in the data sets of group 2. Meanwhile, NCARD, NCAR, ε-neighborhood, and NC show their best performance and find their target clusters. The smallest neighborhood variance has been reported for NCARD algorithm; this shows it leads to more homogeneous neighborhood in data sets. Furthermore, it has been reported that out of the three algorithms, NCARD, NCAR, and NC have shown rather homogeneous neighborhood. The basis for judging statistical significance of differences is the mean value of measured performance. In the two data set groups, NCARD yields significantly better performance than the other neighbourhood algorithms. NCAR, NC, and ε-neighborhood are the next best algorithms, respectively. NCARD shows better strong results in cluster mixes due to its rather low standard deviation values. JI and RI of NCARD algorithm in group 1 data set perform better than the JI and RI of NC and NCAR indicating that cluster mixes are more frequent in NCARD than NCAR. However, divided clusters are more frequent in NCARD than NCAR. The small number of cluster mixes in NCARD facilitates effective clustering. KNN1 and KNN2 have lower JI and RI values than the others.

V. CONCLUSION

In this paper, we proposed a new hybrid ma- chine learning-based intrusion detection method called DCNN, which effectively reduces the feature dimension of the original dataset into a simple and representative two-dimen- sional vector. It saves time and improves the accuracy in our experiment on the KDD Cup 99 dataset. Experimental results show that DCNN can successfully detect intrusions. Additional work is needed in the future. For example, we used only k-means clustering algorithm to obtain cluster centers. We can change the selection method of initial cluster centers to improve the clustering accuracy. In addition, the density calculation is not very ac- curate, In the future, we can change the densi- ty to the density in the point's cluster. Finally, only the k-NN classifier is used as the baseline classifier in this paper. Wide comparisons can be conducted with the use of other baseline classifiers for intrusion detection.

REFERENCES

1. Yao Lan, Wang Xinmei. Present situation and development trend of intrusion detection system. Telecommunications Science. 2002. (12):31-35.
2. YAO Jun-lan. Intrusion detection technology and its development trend. Information Tech- nology. 2006.(4):172-175
3. Devaraju S, Ramakrishnan S. Detection of Attacks for IDS using Association Rule Min- ing Algorithm. IETE JOURNAL OF RESEARCH. 2015 .61(6) 624-633.
4. ZHANG Yi, LIU Yan-heng et al. Intrusion detec- tion system based on association rules. Journal of Jilin University. 2006. (2).
5. A.S. Eesa, Z. Orman, A.M.A. Brifcani. A novel fea- ture-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. EXPERT SYSTEMS WITH APPLICA- TIONS. 2015. 42(5):2670–2679.
6. E. de la Hoz, E. de la Hoz, A. Ortiz, J. Orte- ga, A. Martinez-Alvarez. Feature selection by multi-objective optimization: application to net- work anomaly detection by hierarchical self-or- ganising maps. KNOWLEDGE-BASED SYSTEMS. 2014. 71(SI):322–338.
7. Tsai CF, Lin CY. A triangle area based nearest neighbors approach to intrusion detection. PATTERN RECOGNITION. 2010. 43(1): 222-229.
8. Lin WC, Ke SW, Tsai CF. CANN: An intrusion detection system based on combining clus- ter centers and nearest neighbors. KNOWL- EDGE-BASED SYSTEMS. 2015. 78: 13-21.
9. Wang FN, Wang SS. Solving the intrusion detec- tion problem with KPCA-RVM. DESIGN, MANU- FACTURING AND MECHATRONICS. 2016. 520- 527.
10. Rodriguez A, Laio A. Clustering by fast search and find of density peaks. SCIENCE. 2014. 344(6191):1492-1496.

11. X.-Q. Zhang, C.H. Gu, J.J. Lin. Intrusion detection system based on feature selection and support vector machine. International Conference on Communications and Networking in China. 2006. pp. 1–5.
12. O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, ''An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks,'' Expert Syst. Appl., vol.29, no.4, pp.713–722,2005.
13. G. Wang, J. Hao, J. Ma, and L. Huang, ''A new approach to intrusion detection using artificial neural networks and fuzzy clustering,'' Expert Syst. Appl., vol. 37, no. 9, pp. 6225–6232, 2010. [3] E. De la Hoz, E. De La Hoz, A. Ortiz, J. Ortega, and B. Prieto,''PCA filtering and probabilistic SOM for network intrusion detection,'' Neuro computing, vol. 164, pp. 71– 81, Sep. 2015.
14. R.Singh, H.Kumar, and R.K.Singla, ''An intrusion detection system using network traffic profiling and online sequential extreme learning machine,'' Expert Syst. Appl., vol. 42, no. 22, pp. 8609–8624, 2015.
15. A. S. Eesa, Z. Orman, and A. M. A. Brifcani, ''A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems,''Expert Syst. Appl., vol.42, no.5, pp.2670– 2679,2015.
16. S.-Y. Ji, B.-K. Jeong, S. Choi, and D. H. Jeong, ''A multi-level intrusion detection method for abnormal network behaviors,'' J. Netw. Comput. Appl., vol. 62, pp. 9–17, Feb. 2016.
17. W.-C. Lin, S.-W. Ke, and C.-F. Tsai, ''CANN: An intrusion detection system based on combining cluster centers and nearest neighbors,'' Knowl.Based Syst., vol. 78, pp. 13–21, Apr. 2015.
18. S. M. H. Bamakan, H. Wang, T. Yingjie, and Y. Shi, ''An effective intrusion detection framework based on MCLP/SVM optimized by time varying chaos particle swarm optimization,'' Neuro computing, vol. 199, pp.90–102, Jul. 2016.
19. K. Hwang, M. Cai, Y. Chen, and M. Qin, "Hybrid Intrusion Detection with Weighted Signature Generation over Anomalous Internet Episodes," Vol. 4, pp. 41 – 55, Feb. 2007.
20. [10] Y.Gao, Y. Jin, J. Chen, and H. Wu, " A Novel SemiSupervised Learning Approach for Network Intrusion Detection on Cloud-Based Robotic System," Vol. 6, pp. 50927 – 50938, Sep.2018.
21. K. Wu, Z. Chen, and W. Li, "A Novel Intrusion Detection Model for a Massive Network Using Convolutional Neural Networks," Vol. 6, pp. 50850 – 50859, Sep.2018.
22. M. Amini, J. Rezaeenour A and E. Hadavandi, "A Neural Network Ensemble Classifier for Effective Intrusion Detection Using Fuzzy Clustering and Radial Basis Function Networks," Vol. 25, Apr 2016.
23. W. Hu, Wei Hu and Steve Maybank, "AdaBoost-Based Algorithm for Network Intrusion Detection," Vol. 38, pp. 577- 583, Apr. 2008.
24. R. Amir,X. Zhao Wang and H. Abbas, "Fuzziness based semi-supervised learning approach for intrusion detection system," Vol. 6,pp. 484- 497,Apr.2016.
25. J. Zhang, M. Zulkernine, and A. Haque, "Random-ForestsBased Network Intrusion Detection Systems," Vol. 38, pp. 649 – 659, Sep. 2008.
26. A. Das, D. Nguyen, and J. Zambreno, "An FPGA-Based Network Intrusion Detection Architecture," Vol. 3, pp. 316 – 317, Mar.2008.
27. H. Sadreazami, A. Mohammadi and A. Asif, "Distributed Graph-based Statistical Approach for Intrusion Detection in Cyber-Physical Systems," Vol. 4 pp. 137- 147, Sep.2017.
28. K. Huang, Q. Zhang, C. Zhou and N. Xiong, "An Efficient Intrusion Detection Approach for Visual Sensor Networks Based on Traffic Pattern Learning," Vol. 47, pp. 2704 – 2713, May. 2017.
29. F. Yie Leu, K.Lin Tsai, and Y. Ting Hsiao, "An Internal Intrusion Detection and Protection System by Using Data Mining and Forensic Techniques," Vol. 11,pp. 427 – 438, Apr.2015.
30. A. Hofmann and B. Sick, "Online Intrusion Alert Aggregation with Generative Data Stream Modeling," Vol.8, pp. 282 - 294 Aug.2009.

AUTHORS PROFILE

Ms. M Lavanya, M.Tech Student, Dept. of CSE, Chadalawada Ramanamma Engineering College, Tirupati, India. lavanya.machupalli@gmail.com

Mr. K Munivara Prasad, Associate Professor, Dept. of CSE, Chadalawada Ramanamma Engineering College, Tirupati, India. prasadkmv27@gmail.com