International Journal of Advanced Science and Technology
Vol. 29, No. 12s, (2020), pp. 2736-2748

A Cloud based Anomaly Detection Model using Machine Learning Algorithm

R.BARONA[1], E. A. MARY ANITA [2]

[1]*Research Scholar, Department of ICE, Anna University, Chennai, India.*

[2]*Professor, CSE, S.A. Engineering College, Chennai, India.*

Abstract

Cloud Computing has developed as a recent technology in IT business as a major segment, whereas security concerns remain the essential hindrance to full-scale appropriation. Introduction of cloud based architectures reducing the computing cost has led to the usage of cloud computing. As a result, individual users to large scale organizations, they are turning their attention towards the cloud to store the huge volume of data. But security is one of the major challenges in cloud data storage. In this paper, we have proposed an optimal classification algorithm to classify the anomalies in the data. For optimal classification, we have used Random Forest (RF) classification algorithm with feature selection by Chi-square algorithm. To improve the performance of the classification task, we have proposed a node selection algorithm through a load balancing technique. The experimental validation was carried out with four publicly available sensor datasets from Intel lab. The results showed that the proposed system has high accuracy in anomaly detection.

Keywords: *Random Forest, Cloud computing, Chi-square, Anomaly detection*

1. Introduction

The computational difficulties in today's world are solved effectively with the help of cloud computing. The cloud computing offers various services to the users such as application software services, platform services and infrastructure services. All these services are provided to the users based on their demand. The cloud servers receive a huge amount of requests due to the increase in demand for cloud computing services[22]. Load balancing plays an important role in improving the performance of the cloud nodes. The load balancing approaches are grouped as static and dynamic based on how the changing load and resource status is considered [23, 24, 25]. Based on the resources required by the tasks the nodes are selected from the available nodes to deploy the task. In our work, we have used a node selection algorithm to select the optimal nodes to distribute the tasks.

Anomalies are also referred to as abnormalities, novelties, and outliers among other similar terms[27]. Anomaly detection methods areused to determine such abnormalities in data. Machine Learning algorithms are used today to detect anomalies in data and it has the greatest advantage of having a high detection rate and continuous learning and updating [30–32].The label with an instance denotes whether that instance is normal or anomalous [28]. The anomaly detection techniques that can be operated in three modes are supervised anomaly detection, semi-supervised anomaly detection and unsupervised anomaly detection [28]. In a supervised model, the training set instances are labelled as normal or abnormal classes. The predictive model will identify the anomalous instances depending on the labels in the testing process. In the unsupervised model, there is no labelling of instances and no distinct training process is required. Anomaly detection techniques based on the unsupervised model works on the assumption that the instances are normal and always are far from anomalies. In the semi-supervised model, the predictive model is only for all the normal patterns in the training set, and if the instances in the testing set couldn't be recognized, then it may be denoted as an anomaly.In the semi-supervised model, there is a predictive model only for all normal patterns in the training set, and the instances in the testing set may be an anomaly if it couldn't be recognized. In our work, we have used Random Forest (RF) Machine learning algorithm.RF is one of the supervised machine learning models used to classify anomalies.RF is a collaborative classifier used to improve accuracy. One of the major advantages of random forest is that it yields low classification errors when compared with other traditional classifiers[34] but the training time required for large data sets is higher. The selection of

optimal feature-set is required to reduce the computational overhead and avoid over-fitting problems [33]. Hence, we have introduced the Chi-square feature selection algorithm to reduce the dimensionality of data. This result in less amount of training time for the classifier. Feature selection is used to select optimal features for model construction.The feature selection process calculates the score of each feature and identifies the best set of features. Hence the accuracy of the model will get increased. In our proposed model the classification accuracy is considerably increased with the Chi-square feature selection algorithm. Mesleh has implemented the Support Vector Machines (SVM) algorithm using chi-square as a feature selection algorithm in the preprocessing step [26]. Thereby increased the performance of their proposed model.

The contribution of our work is as follows:

- We have proposed Node selection algorithm through a load balancing approach to select the optimal number of slave nodes in the cloud for processing.
- Chi-square algorithm is used for feature selection followed by RF classifier to classify the anomalies.
- The experimental validation is done with four publicly available sensor datasets from Intel Lab[9]. The resultant values are compared against RF with Feature Selection (FS) and RF without Feature Selection (FS) technique.

The structure of this paper is organized as follows: the related work on anomaly detection based on machine learning techniques discussed in Section 2. Then the functional model for the proposed system is presented in Section 3. Section 4 presented the experiment results and discussion. Finally, the conclusion and the future work discussed in Section 5..

2. Related Work

Kolias [1] performed intrusion detection on the AWID dataset by using various machine learning algorithms. The attribute selection was done manually and selected 20 features to implement 8 classifiers. The overall accuracy of these various classifiers range from 89% to 96%. But, the manual feature selection process is complex and time consuming.Machine learning techniques have been extensively used in the area of anomaly detection techniques. Buczak and Guven [6] performed a detailed literature survey of machine learning techniques for cyber security anomaly detection. Mingjian et al.[4] develop a machine learning based anomaly detection methodology in which they have used neural networksto reconstruct the benchmark and scaling data byusing the k-means clustering and the cyberattack is estimated by the naive Bayes classification.Ibrahim Alrashdi et al.[7] proposed an Anomaly Detection loT (AD-IoT) system to address the loT cybersecurity threats in a smart city. They are used Random Forest machine learning algorithm for anomaly detection and obtained the accuracy of 99.34%. However, the precision rate is lower for their model. HeshanKumarage et al.[8] performed Distributed anomaly detection based on incremental fuzzy cluster evaluation by using the Intel Lab sensor datasets. They have achieved higher classification accuracy rate.Malicious attacks abusing the wireless communication medium on WSNs [10] which enable for eavesdropping, illegal modification of data and data fabrication resulting in confidential information being accessible to unauthorized parties [11].Chan et al. [12] proposed a Secure Information Aggregation (SIA) protocol robust to the stealthy attack in data aggregation perspective.Habeeb et al. [13] performed focused survey on real-time big data processing, anomalous detection, and machine learning algorithms. They have explored the challenges associated with real time big data processing using machine learning for anomalous detection.Rettig et al. [14]proposed a work to detect anomaly over big data streams. Their technique works well on both numerical and categorical data. The two key conditions they have satisfied through their work are generality and scalability. Zhang et al. [15] detects outliers in multidimensional data through their proposed methodology. The anomaly detection is performed by measuring the distance between data points in different subspaces. They have showed that changes may be observable on one dimension, over a subset of dimensions, or overall for multi-dimensional data.Kwon et al. [16] presented deep anomaly detection (DAD) techniques for cyber-intrusion detection.Mohammadi et al. [17]introduced DAD

techniques for the Internet of Things (IoT) and big-data anomaly detection. Ball et al. [18] reviewed Sensor networks anomaly detection.Kiran et al. [19] presented deep learning based methods for video anomaly detection in their proposed work. Li et al.[20]presented the Principal Component Analysis (PCA) approach for linear model-based unsupervised anomaly detection methods.Fabrizio et al. [21] used K-Nearest Neighbor (KNN) algorithm which computes the anomaly scores by calculating the average distance to its k nearest neighbors. Ramachandra et al.[27] performed a detailed survey on video anomaly detection and stated that, to develop a new model and to achieve high detection rate at low false positive rate the researchers should use more complex datasets with large variety of anomaly types , so that the model can be practical and can be used in real time applications. The same can be applied for data anomaly detection to develop a practical model suitable for real time applications. Fan et al. [29] proposed an unsupervised AD method for high dimensional input dataset, which first integrates convolutional auto-encoder and Gaussian process regressionto extract features and to remove anomalies from noisy data as well. They have evaluated the performance of the proposed model by using four datasets and showed that their model is suitable for high dimensional input data. Krishnaveni et al.[35] proposed a model to differentiate normal and abnormal traffic by using Support Vector Machine (SVM) classifier with Information Gain Ratio (IGR) method as a feature selection method and obtained 96.24% accuracy.

To sum up, the developments of machine learning and cloud computing technologies have brought new openings to the development of anomaly detection techniques. We adopt these advantages to design a more sophisticated, intelligent and flexible anomaly detection tool.

3. Functional Model of the system

The functional model of our system consists of the components such as an end-user, Trusted Source (TS), Master and Slave nodes. The TS has the following roles: (i) User Authentication, (ii) Submission of a job to the master and (iii) Send back the result to the user. Initially, the user submits the job to the TS in the cloud. From the TS, the job is now moved to Master node which is also in the cloud. Upon receiving the job from the TS, the master splits the job into the number of tasks depending on the number of slave nodes selected. The slave nodes perform the classification task to perform analysis of the data. The overall execution time or the response time is improved with the selection of execution nodes or slave nodes by using the node selection algorithm through load balancing. Hence node selection plays a vital role in the proposed work of anomaly detection. The classification is done by using the RF algorithm. Instead of performing feature selection by the RF algorithm, Chi-square algorithm is used to select the effective features. The performance of the RF classification algorithm is improved greatly with the introduction of this feature selection algorithm. The anomalies are detected through this process and the resultant report is sent to the user.Fig.1 shows the functional model of the proposed system.

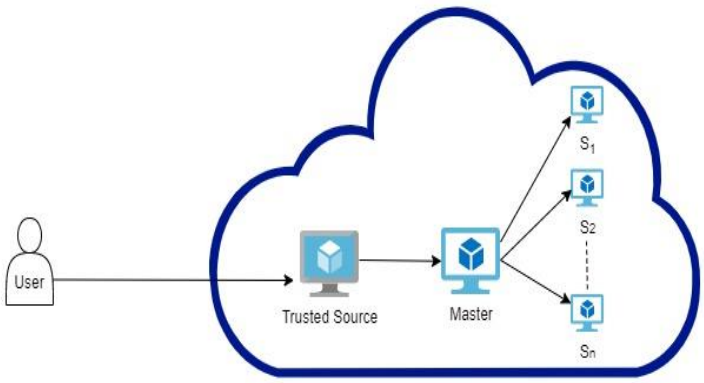

Figure.1 Functional model of the system

3.1. Formulation of node selection Algorithm

Fig.2 shows the proposed work flow of the system. The various steps involved in the process is given as follows:

(i) User authentication and job submission to the TS

The first step in the system is the authentication of user and job submission. The user authentication is performed by using the credentials supplied from the end-user. The credentials are verified to authenticate the user to submit the job. Upon validating, the user is allowed to submit the job. The job submitted by the user is heterogeneous data which may or may not contain anomalies. The TS now hand over the job from the user to the master to perform the analysis to detect anomaly data.

Node selection through load balancing

Node selection is one of the major tasks in the proposed framework. The overhead of the system is reduced greatly by selecting the optimal number of slave nodes required. The required number of nodes are selected through the load balancing technique. Hence,
no slave nodes will suffer by underutilization or overutilization of its resources by assigning tasks to them. The overutilization or underutilization of a specific node is the major issue in the response time or execution time of the task.

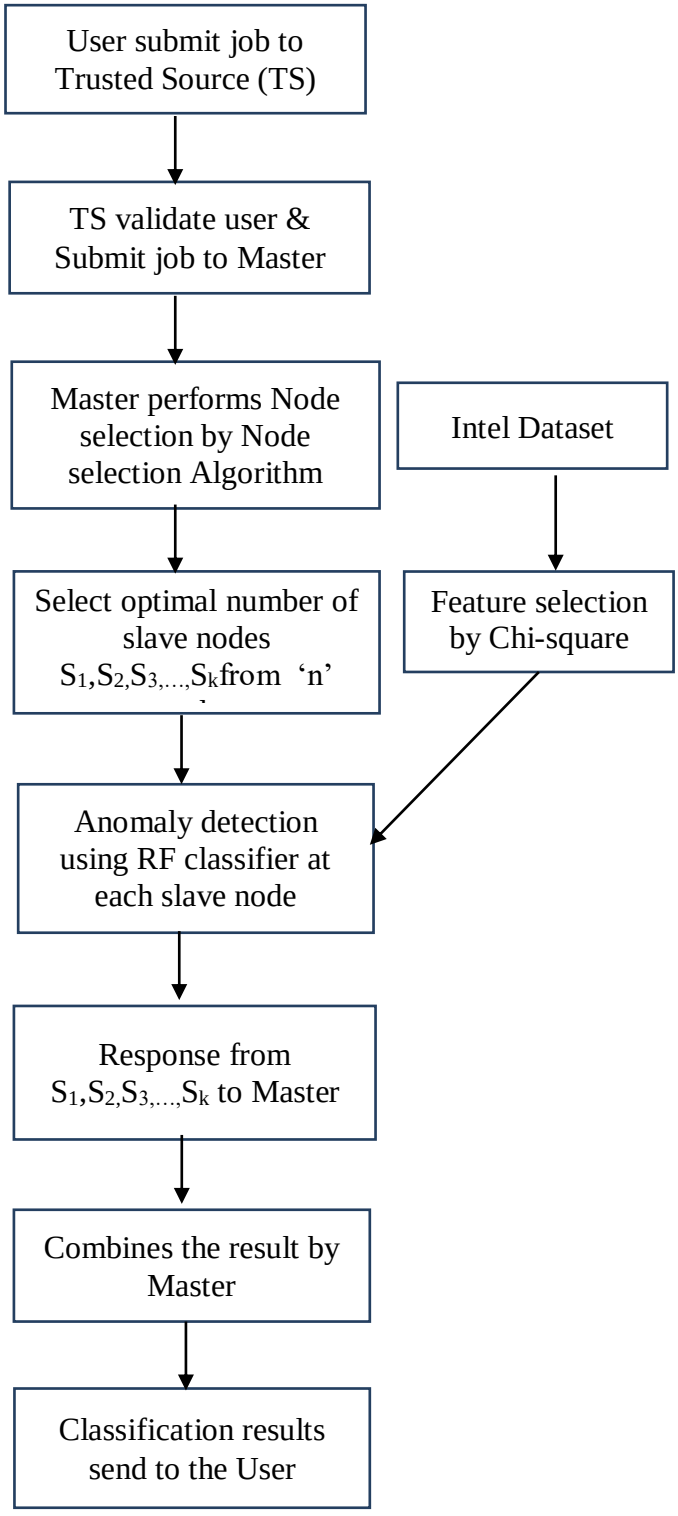

Figure.2 Proposed work flow of the system

Under utilization is the condition in which utilizing the resources in the node greater than or equal to the capacity of the node.

Overutilization is the condition in which utilizing the resources in the node less than the capacity of the node. Here, we have used Bayes' algorithm for efficient node selection process.

Node selection Algorithm

Node selection is important because the size of the data submitted by the user varies depending on the data collected by the sensor over time. The master selects the optimal number of nodes for performing classification by using Bayes' theorem. [2]

The master selects the required number of nodes (N) from 'n' number of nodes as,

$N= s_1,s_2,s_3,\ldots,s_k \quad k \in (1,n)$ (1)

where,

$s_1,\ldots,s_k$ = Slave nodes

N= Optimal number of nodes

Let, T be the number of tasks to be executed and n is the number of available nodes. Based on T, the number of slave nodes will vary. The probability of selecting node 'i' is given by,

$P(S_i) = \frac{1}{n} i \in (1,n)$ (2)

The prior probability for the node is calculated as,

$P(S_i/T) = (P(T/S_i) * P(S_i)) / P(T)\ i \in (1,n)$ (3)

Select the required nodes from the available nodes based on the requirement. The selection will be based on load balancing scheme. The prior probability of the node in available set of nodes is set as,

$P(T/S_i) = 1 - \max(R_i)/(A_j) \quad i \in (1,n) \text{ and } j \in (1,m)$ (4)

where,

R_i = Required nodes

A_j= Available nodes

The posterior probability is given as,

$P(S_i/T) = (P(T/S_i) * P(S_i)) / P(T) i \in (1,n)$ (5)

where,

$P(T) = \Sigma P(T/S_i) * P(S_i)\ i \in (1,n)$ (6)

The posterior probability in eqn(5) is calculated for each node to select the optimal nodes.

Begin
Initialize T= Number of tasks to be submitted, n=Number of nodes available
S_i = Selecting a node i from n
Compute,
$P(S_i) = \frac{1}{n}$
Calculate the prior probability for each node as,
$P(S_i/T) = (P(T/S_i) * P(S_i)) / P(T)$
Compute the value of $P(T/S_i)$ as,
$P(T/S_i) = 1 - \max(R_i)/(A_j)$
for each i
Calculate the posterior probability,
$P(S_i/T) = (P(T/S_i) * P(S_i)) / P(T)$
end for
End

Figure.3. Node Selection Algorithm

3.2. Anomaly detection by classification

The classification task is carried out by the selected slave nodes. The save nodes detect the anomaly in the data by performing RF classification algorithm.

In the original RF algorithm, the features are randomly selected by the algorithm to perform the classification task. In our proposed work to achieve effective performance, the features are selected by Chi-square algorithm [5].

Feature selection by Chi-square Algorithm

There are many features available in the dataset, out of which some features will be useful others may use less. The feature selection algorithm is used to select a useful and important feature for processing.

The selection of useful features increases the accuracy and reduce the computational cost required thereby higher performance will be achieved. The time required for training also reduced.

Chi-square is a numerical test that computes the deviation from expected value by considering the feature value is independent of the class value. The Chi-square is calculated by using the following measures, True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN), Probability count of positive values Pp and Probability count of negative values PN.

Chi-square measure = t(TP,(TP+FP) Pp) + t(TN,(TN+FN) PN) + t(FP,(TP+FP) PN) + t(FN,(TN+FN) Pp) (7)

where,

t(count, expect)= (count – expect)2/expect (8)

The Degrees of Freedom[3] and test statistic can be calculated as follows,

Degrees of Freedom: DF = (r-1)*(c-1) (9)

where,

r= number of levels of one categorical value

c= number of levels of other categorical value

Test statistics:

$\chi 2$ (f,c) = (N*(AD-CB)^2)/((A+C)(B+D)(A+B)(C+D)) (10)

where,

N = Total number of data records

A=Number of times feature ‘t’ and class ‘c’ co-occurs

B= Number of times feature ‘t’ appears without class ‘c’

C= Number of times class ‘c’ appears without feature ‘t’

D = Number of times neither class ‘c’ nor feature ‘t’ appears

The important features are selected by Chi-square algorithm and the less important features are removed from the dataset.

Random Forest Algorithm

To predict the data the RF classifier uses a set of classification and regression trees. Standard divide and conquer approach is used by RF to increase the performance. The classifier comprises of a mix of tree classifiers where each classifier is created by utilizing an arbitrary vector, and each tree makes a unit choice for the most popular class to classify the input vector.

The Random Forest experimented for their investigation has chosen features at every node to grow a tree. Hence the RF classifier becomes more robust against data noise and overtraining them any other classifier based on boosting. Random forests are an ensemble learning method for classification, which

works by building multiple decision trees [32]. The multiple decision trees are constructed from the feature selected by the Chi-square algorithm. To classify a new object the decision from each tree in the forest is needed. The class which gains the maximum votes for the object is selected by the forest.

Steps used in RF algorithm

Select the useful features using Chi-square algorithm.
Initialize n-tree value, which is the number of decision trees in the forest.
Specify the split parameter and select the root node.
Initialize the maximum height of the tree.
Define the minimum number of data points for the leaf node.
Predict the outcome value from each decision trees.
Calculate vote for each target value predicted by each decision tree.
The target value with highest vote is measured as the final prediction of RF algorithm.

4. Results and Discussions

For experimental validation, we have used the four publicly available datasets from Intel Berkley Research Lab. Intel lab has deployed 54 sensors in the lab and these sensors collected the data such as temperature, humidity, voltage and light values once every 31 seconds. The dataset containing features such as Date, Time, epoch, moteid, temperature, humidity, voltage and light values. Table 1 gives the dataset format Intel dataset. We have used Python for the implementation of our proposed approach.

Table 1: The format of Intel Dataset

Features	Data Type
Date	Date
Time	Date
Epoch	int
Moteid	int
Temperature	real
Humidity	real
Light	real
Voltage	real

Evaluation Metrics

The Evaluation metrics used are accuracy, sensitivity, specificity and processing time. These evaluation metrics are computed with the four datasets available from Intel Lab such as Temperature, Humidity, Light and Voltage. The metrics are compared with RF with Feature Selection (FS) and RF without FS. RF with FS shows more accuracy than RF without FS.

Accuracy

Classification accuracy is the measure of how accurately the model predicts the data. Accuracy reveals classifiers ability to classify the dataset exactly.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

The Accuracy value obtained for the four datasets are as in Fig 3.

The resultant measures show that the RF with FS is having the highest classification accuracy than RF without FS across the four datasets. Our proposed model of RF with FS is having accuracy rate is above 95% for all the four datasets.

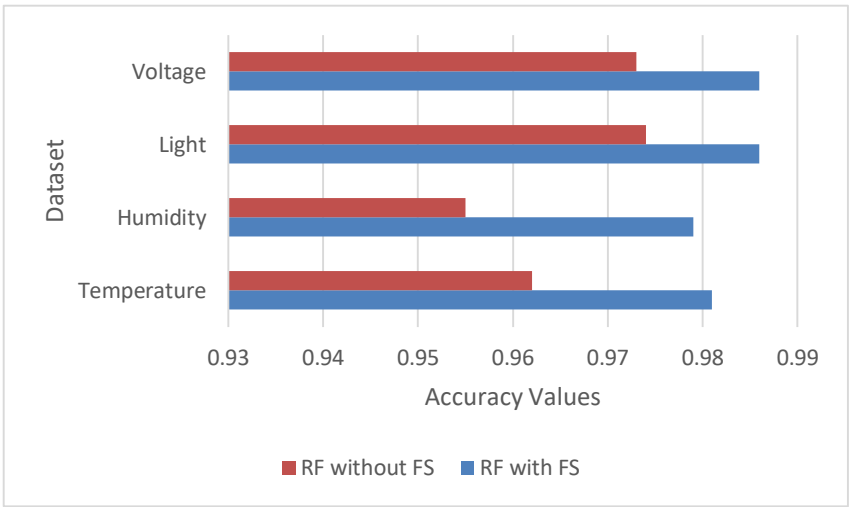

Figure. 3. Comparison of classification Accuracy values among the four datasets

Sensitivity

Sensitivity is the measure of true positive rate. It is the ratio of number of anomalies detected by the model to the total number of anomalies in the dataset.

$$Sensitivity = \frac{TP}{TP+FN} \quad (12)$$

The obtained Sensitivity values for the four datasets are as in Fig 4.The obtained result for sensitivity shows that RF with FS is having higher true positive rate than RF without FS. Hence, our proposed model is having high sensitivity rate. We obtained the sensitivity rate of above 97% for all the datasets by using RF with FS.

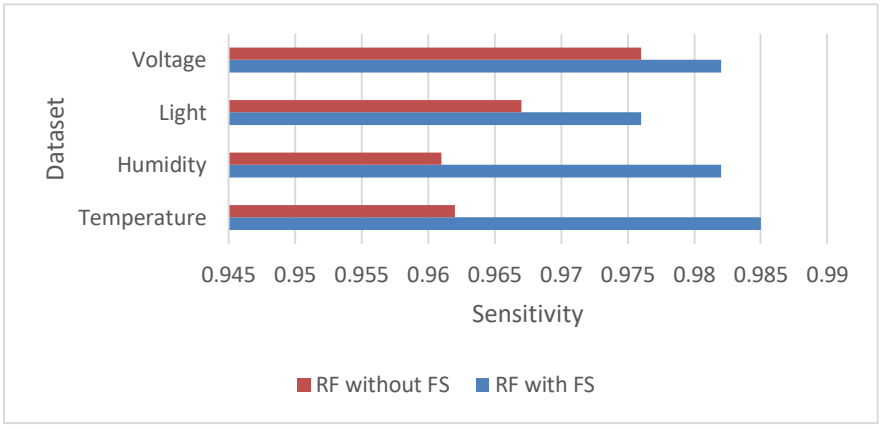

Figure.4. Comparison of Sensitivity values among the four datasets

Specificity

Specificity is the measure of true negative rate. It is the percentage of normal data classified as anomaly data.

$$Specificity=TN/(TN + FP) \quad (13)$$

The obtained Specificity values for the four datasets are as in Fig 5.The obtained result for specificity shows that RF with FS is having higher true negative rate than RF without FS. Our proposed model of RF with FS gives the specificity rate of over 98% for all the four datasets.

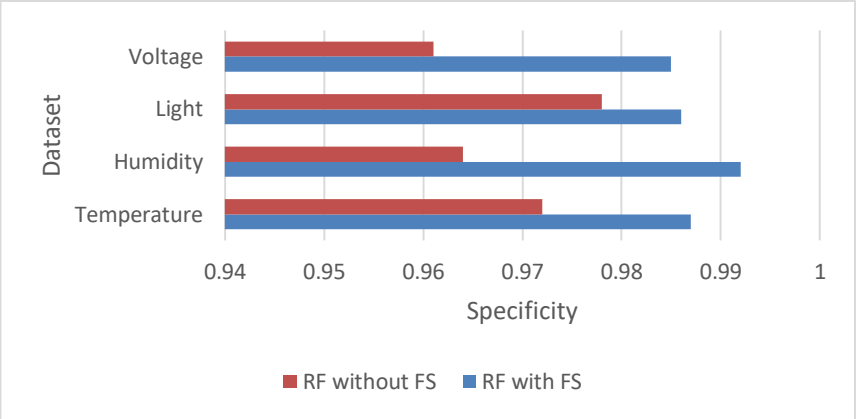

Figure.5.Comparison of Specificity values among the four datasets

Processing Time

The processing time is the total computational time required by the systems in the cloud to complete its execution. Since we have used the node selection algorithm to select the optimal number of nodes required for processing, the processing time is considerably reduced.

The performance efficiency of our proposed work is evaluated with the evaluation metrics and the obtained results are given in fig.3, 4, and 5.Our system gives a high accuracy measure in detecting the anomalies present in the data. The computational complexity of RF algorithm is reduced and the accuracy rate is increased with the Chi-square feature selection algorithm. The processing time is reduced by selecting the optimal number of nodes required with the use of node selection algorithm. The node selection depends on the size of the task. For our model, two optimal nodes are selected for processing the dataset.Hence, the performance efficiency of our model in terms of accuracy rate and processing time is more when compared with existing models. The processing time of our proposed model and existing models are compared and the result obtained is shown in fig.6. Fig. 6 depicts the processing time of the conventional techniques used in Alrashdiet et al[7], Kumarage et al.[8], Alwadiet al.[36] and Genuer et al.[37] and proposed cloud based anomaly detection technique using RF classifier with Chi-square feature selection algorithm and optimal node selection through node selection algorithm. The overall execution time or the processing time considerably low when comparing with these existing techniques.

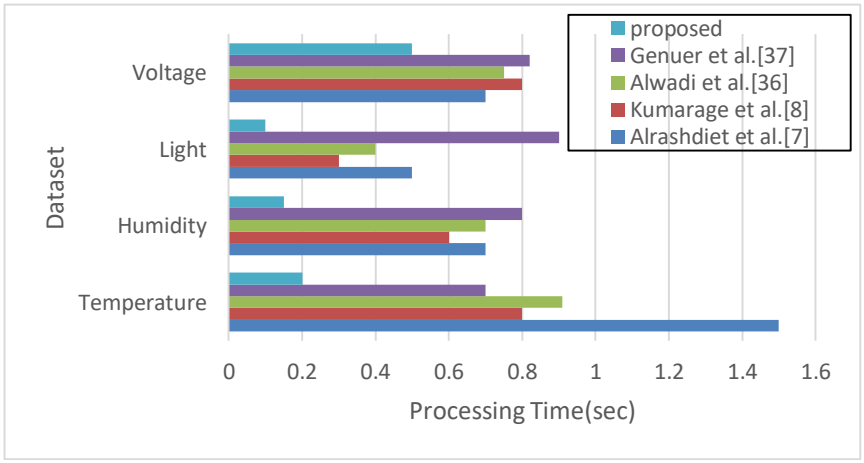

Figure. 6. Processing time of proposed and existing techniques.

The overall performance of our proposed work is excellent with Intel dataset in the cloud environment. We have obtained good accuracy and processing time values.

5. Conclusion

In this paper, we have proposed cloud based anomaly detection model by using Node selection algorithm and RF classifier with Chi-square feature selection algorithm. Thus, our proposed cloud based anomaly detection model can perform secure anomaly detection based on the Node selection algorithm and RF classifier with Chi-square feature selection algorithm within the cloud computing environment comprising the Trusted Source. This Trusted Source belongs to the cloud environment and the analytics service is performed by the slave nodes in the same cloud. The computational burden is reduced by selecting the optimal number of slave nodes. The anomaly detection operations are performed by the selected slave nodes. The experimental validation is performed on four publicly available sensor datasets of Intel Lab to analyze the performance of our proposed work on anomaly detection process. The proposed model illustrated high accuracy in detecting data anomalies.

In future, other machine learning algorithms with different datasets would be evaluated in the same scenario to compare the performance with the proposed one. Also, a case study to further enhance the performance of the proposed technique in comparison to other anomaly detection techniques would also be investigated.

References

1. Kolias C, Kambourakis G, Stavrou A, et al. Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset[J]. IEEE Communications Surveys & Tutorials, 2016, 18(1): 184-208.
2. Naidila Sadashiv and Dilip Kumar S M, “A Baye's theorem based node selection for Load balancing in cloud environment”, International Journal on Cloud Computing: Services and Architecture (IJCCSA) Vol. 7, No. 1, February 2017.
3. Sumaiya Thaseen and Ch. Aswani Kumar,” Intrusion Detection Model Using fusion of Chi-square feature selection and multi class SVM”,Journal of King Saud University - Computer and Information Sciences,DOI:http://dx.doi.org/10.1016/j.jksuci.2015.12.004, 2015.
4. Mingjian Cui, Jianhui Wang and Meng Yue “Machine Learning Based Anomaly Detection for Load Forecasting Under Cyberattacks”, IEEE Transactions on Smart Grid,2019.
5. Bahassine, S. Kissi M. Madani, A. New Stemming For Arabic Text Classification Using Feature Selection and Decision Trees, in: Proceedings of the 5th InternationalConference on Arabic Language Processing (CITALA) Oujda, Morocco, (2014) pp.200–205.
6. A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” IEEE Commun.Surv. Tutor., vol. 18, no. 2, pp. 1153–1176, 2016.
7. Ibrahim Alrashdi, Ali Alqazzaz,EsamAloufi, RaedAlharthi, Mohamed Zohdy, Hua Ming, “AD-IoT: Anomaly Detection of IoT Cyberattacks Smart City Using Machine Leaming”,IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). doi:10.1109/ccwc.2019.8666450,2019.
8. H. Kumarage, I. Khalil, Z. Tari, A. Y. Zomaya, Distributed anomaly detection for industrial wireless sensor networks based on fuzzy data modelling, J. Parallel Distrib. Comput. 73 (6) (2013) 790–806.
9. P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, R. Thibaux, Intel lab data, 2004 (db.csail.mit.edu/labdata/labdata.html).

10. D. Djenouri, L. Khelladi, A. Badache, A survey of security issues in mobile ad hoc and sensor networks, IEEE Communications Surveys & Tutorials 7 (2005) 2–28.

11. E. Shi, A. Perrig, Designing secure sensor networks, IEEE Wireless Communications 11 (2004) 38–43.

12. Haowen Chan, B. Przydatek, D. Song, and A. Perrig, "SIA: Secure Information Aggregation in Sensor Networks," Proc. 1st ACM Int'l. Conf. Embedded Networked Sensor Sys., Nov. 2003, pp. 255–65.

13. RiyazAhamedAriyaluranHabeeb , FarizaNasaruddin, , Abdullah Gani , Ibrahim AbakerTargio Hashem , Ejaz Ahmed and Muhammad Imran, "Real-time big data processing for anomaly detection: A Survey", International Journal of Information Management, 2018.

14. Laura Rettig, MouradKhayati , Philippe Cudre-Mauroux and MichałPiorkowski," Online Anomaly Detection over Big Data Streams", Applied Data Science – Springer,2019.

15. J. Zhang, M. Lou, T. W. Ling, and H. Wang, "HOS-Miner: a system for detecting outlytingsubspaces of high-dimensional data," in Proceedings of the 30th International Conference on Very Large Databases, Toronto, Canada, 2004, pp. 1265– 1268.

16. Donghwoon Kwon, Hyunjoo Kim, Jinoh Kim, Sang C Suh, Ikkyun Kim, and Kuinam J Kim. A survey of deep learning-based network anomaly detection. Cluster Computing, pages 1–13, 2017.

17. Mehdi Mohammadi, Ala Al-Fuqaha, SamehSorour, and Mohsen Guizani. Deep learning for iot big data and streaming analytics: A survey. arXiv preprint arXiv:1712.04301, 2017.

18. John E Ball, Derek T Anderson, and Chee Seng Chan. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. Journal of Applied Remote Sensing, 11(4):042609, 2017.

19. B Ravi Kiran, Dilip Mathew Thomas, and RanjithParakkal. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. arXiv preprint arXiv:1801.03149, 2018.

20. S. Li and J. Wen, "A model-based fault detection and diagnostic methodology based on pca method and wavelet transform," Energy and Buildings, vol. 68, pp. 63–71, 2014.

21. A. Fabrizio and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in In European Conference on Principles of Data Mining and Knowledge Discovery, Berlin, Heidelberg, 2002. Springer, 2002, pp. 15–27.

22. Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski,

23. Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and MateiZaharia, "A view of cloud computing", Communication ACM, vol. 53, no. 4, pp. 50–58, April 2010.

24. Q. Wei, G. Xu, and Y. Li, "Research on cluster and load balance based on Linux virtual server", in Proc. International Computing Applications, vol. 105, pp. 169–176, 2011.

25. W. Chen, Y. Zhang, and Z. Xiong, "Research and realization of the load balancing algorithm for heterogeneous cluster with dynamic feedback", Journal of Chongqing University, vol. 33, no. 2, pp.2–14, 2010.

26. S. Song, T. Lv, and X. Chen, "Load balancing for future internet: An approach based on game theory", Journal of Applied Mathematics, vol. 2014, no. 2014, Article ID 959782, Feb. 2014.

27. A. M. d. A. MESLEH, "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System " Journal of Computer Science, vol. 3, p. 6, 2007.

28. Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai, "A Survey of Single-Scene

29. Video Anomaly Detection", IEEE, Computer Vision and Pattern Recognition, 2020, arXiv preprint arXiv:2004.05993, 2020 - arxiv.org.

30. Ming Zhao, Jingchao Chen," A Review of Methods for Detecting Point Anomalies on Numerical Dataset", IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2020.

31. Jinan Fan, Qianru Zhang, Jialei Zhu, Meng Zhang, Zhou Yang, Hanxiang Cao,"Robust deep auto-encoding Gaussian process regression for unsupervised anomaly detection, Neurocomputing, 2020.

32. T.T. Nguyen , G. Armitage ," A survey of techniques for internet traffic classifica- tion using machine learning", IEEE Commun. Surv. Tutor. 10 (4) (2008) 56–76 .

33. A.L. Buczak , E. Guven , "A survey of data mining and machine learning methods for cyber security intrusion detection", IEEE Commun. Surv. Tutor. 18 (2) (2016) 1153–1176 .

34. M.S. Mahdavinejad , M. Rezvan , M. Barekatain , P. Adibi , P. Barnaghi , A.P. Sheth , "Machine learning for internet of things data analysis: a survey", Digital Com- mun. Netw. 4 (3) (2017) 161–175 .

35. Sahil Garg, Kuljeet Kaur, ShaliniBatra, Gagangeet Singh Aujla, Graham Morgan, Neeraj Kumar, Albert Y. Zomaya , Rajiv Ranjan,"En-ABC: An ensemble artificial bee colony basedanomaly detection scheme for cloud environment",Journal of Parallel and Distributed Computing, Volume:135, Pages:219–233,2020.

36. AmarMeryem and BouabidELOuahidi, "Hybrid intrusion detection system using machine learning", Network Security, Volume 2020, Issue 5, Pages 8-19,2020.

37. S. Krishnaveni, PalaniVigneshwar, S. Kishore, B. Jothi and S. Sivamohan,"Anomaly-Based Intrusion Detection System Using Support Vector Machine", Artificial Intelligence and Evolutionary Computations in Engineering Systems, pp 723-731, 2020.

38. Mohammad AbdulazizAlwadi, GirijaChetty," Sensor Reduction Method for Intel Berekely Data Set Based on Machine Learning",International Journal of Information Engineering and Applications, 2018.

39. Genuer, R., Poggi, J.-M., Tuleau-Malot, C., & Villa-Vialaneix, N. "Random forestsfor big data. Big Data Research", 9, 28–46, 2017.