International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-7, May 2020

Anomaly Detection using Machine Learning: A Rule Based Classification and Ordered Regression Tree with Real Time Datasets

Jidiga Goverdhan Reddy, Sammulal Porika

Abstract: ***What we use the protection of system data and user credentials is still very dispensable presently in factual applications frequently used by common people. Also losing their assets and confidence level due to lack of knowledge about usage of applications and failure to grab the abnormal behavior. How system data and user credentials are helpful to creating clone by others causes of showing anomalous behavior and don't know to protect from the anomalies and how it is avoid. In this paper we are presenting short-lived discussion on anomaly detection and its nature of impact showing on original true datasets related to daily land transactions, medical and social networking. This paper shows the significant usage of machine learning approach applied in anomaly detection to know the fact anomalies in various datasets took from different sources. Here we are using an updated CART called Rule based Classification and Ordered Regression Tree (RBT-ORT). This method is new one with combination of Decision Tree; Rules of Random Tree giving a new adorned rule sets in classification and regression to ensure the improvement in results compare to other techniques. Our work carried out on three datasets, two are taken from UCI repository for machine learning and other one is real and original dataset Land sale data pertaining to land transactions noted in the year 2016-18. Finally the results of anomaly detection using Classification and Ordered Regression Tree compare with other machine learning methods such as ID3, C4.5, C-RT, PLSDA, CHAID, C4.5 Rule, I (Improved) - C4.5, K-Nearest neighbor and Neural Networks.***

Keywords: Anomaly, Anomaly detection, Classification, Regression;

I. INTRODUCTION

Today computer users are facing security problems associated with anomalies encountered through usage of malicious software and applications. So that, the protection of user data is an essential part of computer systems while data in flow or store stable in memories. In this case we need to create trained models of anomaly detection systems (ADS) to monitor the continuous occurrences of anomalies.

Revised Manuscript Received on May 30, 2020.

***** Correspondence Author

Jidiga Goverdhan Reddy* , Lecturer, Dept. of Technical Education ,Government of Telangana State and Research Scholar, JNTU University Hyderabad, India, jgreddymtech@gmail.com.

Sammulal Porika, Professor, Dept. computer engineering, JNTUH Nachupally, Jagitial, Affiliated to JNTU University, Hyderabad, TS State, India, sammulalporika@gmail.com

But it is a severe challenge to develop the flexible and robust methods to face the dynamic attacks. For this cause of situation, a contextual and continuous based ADS is required to update the models of security to protect the data stored and inflows in networks against fraudulent instances. Anomaly detection is playing a significant role in decisive applications to shield many anomalies including cyber attacks. But as on today many of the defender software's not able to detect the anomalies raise in coding exploits.

The ADS is a subset of traditional intrusion detection [1, 2] is defined as the process of finding the data patterns that may not matching the original behavior by monitoring the suspicious events raised in computer or network usage. The Anomaly is a kind of potential discrimination of data in the original data instances that may not showing as original as we expecting in the nature of normality as part of the input data collected from profiles. In anomaly detection [2, 6], the behavior was observed through normal runs of programs of applications. It is a program or approach to the intrusion detection gives the solution that examines whether security violation issues exist in programs and construct a standard model of normal behavior of users or systems and detects the behaviors that shows abnormal deviation from the model. The unauthorized access and usage of system files generally cause of identifying suspected behavior. The anomaly detection generally working towards to know the occurrences of outliers, anomalies, noise, malicious actions at system level execution data and application level. The main challenge is to be to face the anomalies in applications which involving providing a countermeasure though deploying machine learning approaches.

II. MACHINE LEARNING IN ANOMALY DETECTION

A. Why anomaly detection

The anomaly detection system is develop with use of machine learning's strong statistical background will improve the detection of latest and novel attacks, low false positive rate and substantial scope to enhance further applications. In this security field many intrusion detection systems [2,5,6,12] techniques have been developed based on machine learning for maximizing the results and performance. The aim of this work is to present a thorough investigation of both data mining [4, 14, 16, 17] and machine learning algorithms [5, 11] for these detection tasks.

Many are presented the algorithms categorized under various appropriate concepts like unsupervised model or semi supervised [18] models for static vs. dynamic data in present takes maximum resources. The dominant observations of ADS are anomaly behavior expectation, detecting zero-day exploits in regular behavior, structure of anomaly, detection of inside behavior, and awareness of anomalies, model selection and validation.

B. Machine Learning: Rule based classification and Ordered regression tree (RBC-ORT)

The Machine learning (ML) applications are changing predominantly in development of adorned anomaly detection systems, because the ML is actually incorporate the properties such as advance statistical foundations, adaptability to dynamic changes , robustness in working, flexible to enhancement, performance based criteria involved, outcome effectiveness , multiple solutions and accurate results. The machine learning based anomaly detection performance improved by a use of result criteria and fast working models.

In this paper, we are working with an updated decision tree (DT) based technique called "Rule based Classification and Ordered Regression Tree" (RBC-ORT). The context of use the decision tree is mainly suitable for datasets may be large and accuracy of detection of anomalies is high. The DT is one good supervised non-parametric and influenced algorithm which is popular in machine learning techniques that includes classification and regression. The classification is based new rule sets on the observed data in discrete which classifies the instances generally in recursive based where node gives rules and leaves showing the class belongings. In this adorned procedure, the new rules are extracted from Random Tree algorithm [3,8,9] helping to DT classifier to classify the data instances effectively with high accuracy. Similarly the regression is for predicting of instances for target as small regions by means of training data. The DT is applied to classification, but not possible to extract much rules due to conventional and regression problems to classify the instances and prediction of instances. Sometimes ordered regression tree also called up-hilled decision tree [7].

III. LITERATURE REVIEW

In general, the anomaly detection falling into two broad categories based on previous work i.e. one is Statistical and Knowledge based anomaly detection and second most popular working platform is known as Machine Learning based anomaly detection. In this, proposed machine learning work is to be selected due to reasons like strong mathematical and functional background, low FPR and accuracy maximize, adorned approach, tracing zero-day intrusions,..etc. There was lot of work done in machine learning related to anomaly detection in security field. In the first type the authors developed ADS based on statistical and knowledge based formulas like gaussian random variables, multivariate models, information theory, time series, counters, time-related metrics, finite state machines, call-graph model (ndfa) [4], description languages-n-grams, expert systems and rule based. The other sides of ADS's are developed by using machine learning (ML) like broad supervised, semi-supervised, un-supervised algorithms including data mining techniques (DM). In this related work, we are only giving previous work done on anomaly detection using different machine learning schemes.

The majority machine learning algorithms are supervised applied over anomaly detection in real datasets to improve the accuracy of results and reduce the false positive rate. The conventional classifier tree used in anomaly detection and one is most popular ID3 [11] top down algorithm designed based on classification heuristics and greedy based. The CHAID [10] algorithm is an attractive DT, But pruning is poor. The C4.5 [12] is developed with maximizing features and constructs the tree on giving equal priorities to set of attributes. C4.5 also includes some drawbacks such as overfitting, null instances, and irrelevant attribute instances in some cases, hence boosting and bagging giving strength to this technique.

For C4.5 Rule [12, 20], the extension was developed based on adorned rules re-ordering will improve the rate of performance. For this, always make dynamic in rules updating for attribute is continuous. The I (improved)-C4.5 [18, 19] is an extension work of C4.5 also gets well in performance with changes made to entropy (β) and enhanced I_g ratio in place of standard I_g. In paper [20], active rule C4.5 developed for IDS in networks by classification rule sets extract from dataset KDD. The C4.5 in well for structured datasets is giving maximum outcome for boosting and bagging classifiers (Ensemble methods) for data don't contain any error data or to combine predictions of classifier under the assumptions of Gaussian naive bayes (GNB) process. Our background model CART [13] developed based on recursive classification, but takes cost additionally as usual for pruning process. Ada-Boost [6] algorithm is well explained and good for some cases than C4.5 with bagging in accuracy.

The random tree selection of RF tree is usually enhancing the performance outcomes better than all conventional tree classifiers including Ada-Boost [6, 12]. The RBDT-1[27] is new concept based on rules imposed dynamically and generate the tree, but this may not extends for large datasets. The other Machine learning also including the data mining [16] background approaches which generate rules from given data. The PCA used for anomaly or outlier detection on any data by dimension reduction, Support Vector Machine (SVM), Hidden Markov Model (HMM). Rule Learning, Neuro-Fuzzy (NF) computing, Multivariate Adaptive Regression Splines, Reinforcement learning, Ensembles of classifiers, Supervised, Unsupervised learning, and Linear Genetic Programming are the techniques widely used for anomaly detection[2, 5]. The NN [21] used in ADS with (MLP)

Retrieval Number: F4578049620/2020©BEIESP
DOI: 10.35940/ijitee.F4578.059720
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

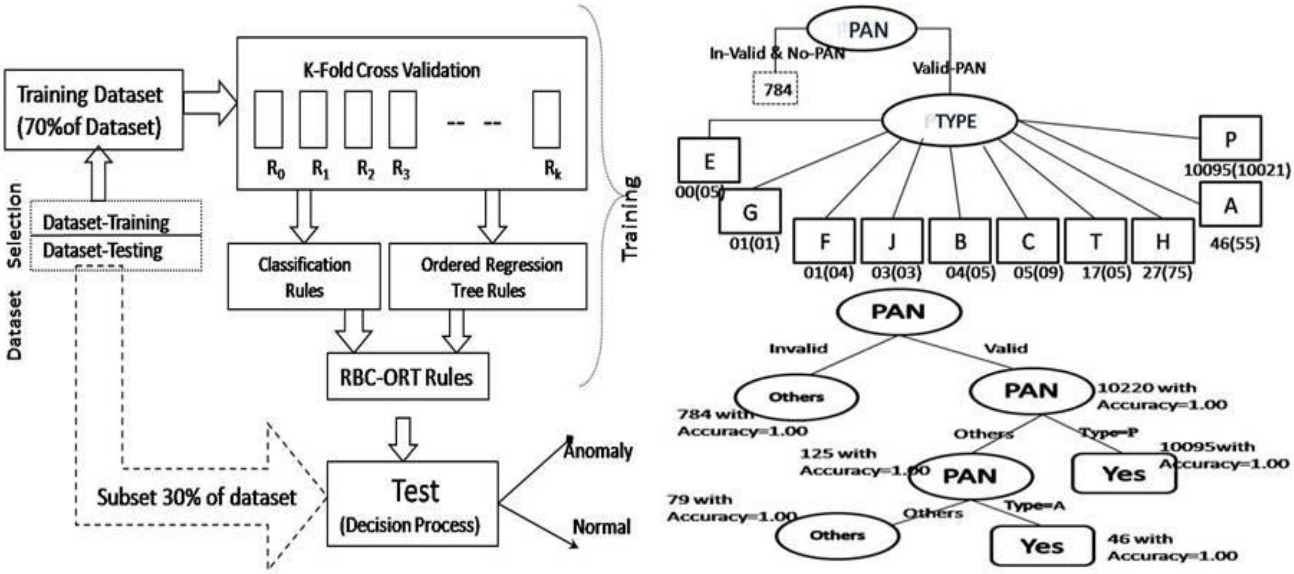


Fig.1. Proposed Model Framework (Left Figure shows RBC-ORT model design view, Right Figure show general classification tree and regression model of dataset-1.

-multi layer perceptrons is also popular in data instances classification and may be improving the DR and accuracy with changing of NN parameters. Naïve bayes (NB) [24] is based on GNB also unadorned and better than ID3 in performance and also NB giving good classification results after rational updating of rules.

IV. FRAME WORK OF PROPOSED SYSTEM

The frame work of RBC-ORT, a rule based classification and regression model is shown in Fig.1 (left), it is a simple process begins with a selected datasets as input and divides them into training and testing datasets according to no. of instances in the dataset. For training purpose generally select the samples more than 50% of original dataset. The ratio of training and testing is 50:50 or 70:30 and in this paper we have selected the 70:30 ratios for all datasets. The training dataset is use the K-Fold cross validation to get accurate outcome as results. The generation of regions R = {R_0, R_1, R_2 …R_n} uses k-fold cross validation in training.

In the frame work of RBC-ORT (Rule based classification and Ordered regression tree) Rules (Left) and the making of decision tree with rules derived from random tree (RT). In Fig (Right-top) shows the simple DT fabricated on high Ig attribute called permanent account number [PAN] for dataset (Real time Land sale data) and Regression tree shown partially in (Right-bottom). In this paper we are giving only common framework model for all datasets, but explained more about real time land sale data. From the Fig.1 (left), the RBC-ORT (Rule based classification and Ordered regression tree) is combine the rules of classification and regression. The rules are generated separately from dataset data instances generally grouping into α-rules during training and β-rules for entire dataset during testing.

The classification algorithm is choice to select here and we have considered Random Tree (RT) to form a rule set {R_α , R_β}. The β-rules are updated to α-rules after testing of dataset is completed. The dataset named as (X) contains possible data instances{X_0, X_1, X_2 ,X_3 ,… …X_n} divide into regions R_i by calculating average variance, means.. etc. In this paper, we are creating a new rule based classification and ordered regression tree model to stabilize the results compared other models developed on DT and traditional supervised algorithms. The anomaly detection with new model is creating flexibility in working and easy to observe the results.

The rule based classification algorithm is depends on the generalized DT is fabricated for real land sale dataset has an attribute set {PAN, Type, DATE} among the many attributes in original dataset. The partial regression tree is fabricated is shown in Fig.1 (right-bottom) with PAN split attribute for Land Sale dataset. In this dataset the missing values are filled with some related values and the experimental work done on only dataset contains PAN attribute. The rule set collection is not shown exclusively in algorithm, but used in implementation such as our previous work [26].

V. ALGORITHM

The algorithm shows the classification model and ordered regression tree, the function derived on dataset contains rules for generalized classification as shown in our previous work [26]. The algorithm has parameters such as |T| denoting no. of terminal nodes, α is a cost complexity parameter during training and pruning also.

Algorithm: Rule based Classification and Ordered Regression Tree

(1).Select Dataset (D)
For Classification:
(2).Derive Ig for A_i i-th attribute;
(3).Compare [I (A_i),I (A_j)] select any attribute A with maximum I as root ;
(4).Dataset D (D_0,D_2,D_3,…D_n) Where each D_iЄ(A_n U C_n) and D_i=(X_i ,Y_i), X_iЄA_n and Y_iЄC_n
(5).Divide the Attribute Domain Space into Regions R_1 ,R_2 ,….R_n Based on High I_g;
For Ordered Regression
(6).Use Recursive Splitting Based on

$$\hat{f}(x) = \sum_{i=1}^{n} C_m \quad I(X_1, X_2) \in R_m \qquad (1)$$

Minimize the RSS (Residual sum of Difference) or SSE (Sum of Square Error)

$$RSS = \sum_{i=1}^{m} \sum_{i \in R_m} \cdot \ (Y_i - \hat{Y}R_m)^2 \qquad (2)$$

(7).Apply Cost Complexity Pruning by selecting a Cost Complexity parameter α;

$$\text{Minimize}\ \{RSS + \alpha\ |T|\ \} \qquad (3)$$

From the algorithm, we can observe that the recursive binary splitting is applied in training phase to develop maximum tree in depth to attain more clarity in observation and stop if pruning is required to develop best tree by constructing optimal sub trees at each level. The information gain (I_g) is calculated for selecting splitting attribute to maximize the classification and build the complete tree.

$$I(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - EH(A) \qquad (4)$$

Where, all parameters are calculated as shown in [26].

VI. DATASETS FOR CASE STUDY

To carry out the experimental work, we selected three datasets 1) Real time dataset of Land sale data noted in the year 2016-18 2) Heart-male decease dataset from UCI machine learning repository and 3) Facebook Live Sellers in Thailand Data Set from UCI machine learning repository [25].

Land Sale data: it is pertaining to land transactions of one of the state made during the year 2016-18. This dataset contains 14 attributes and 10348 records about transactions made with respect to selling and purchasing lands. In this we have focused on Classification of records based on attribute which is having maximum information gain (I).

Heart decease: This dataset is taken from source [25], a ML repository from UCI and it is contains 8 attributes and 208 observations (instances).

Facebook Live Sellers in Thailand: This Data Set is from source [25], a ML repository from UCI and it is contains contains 12 attributes and 7051 observations (instances) pertaining to year 2019. The dataset shows facebook data and messages happen between different Thai people for business purpose.

VII. EXPERIMENTAL WORK AND RESULTS

Our practical work done on the datasets of some real time data, other sources and considered some benchmark comparisons. The three case studies (UCI ML source and Land dataset) are programmed on system with 3GHz (2GB Memory), XP-OS, and simulation tool: Matlab [22]. The performance benchmark for case studies, some data mining tools are considered [23].

The experimental work done on all three datasets are shown in Table.1. The datasets are trained with 70% of the sample data from original and test cases done on remaining samples.

Table- I: Test results giving the evaluation of ML algorithms performance (per 100%) of all three datasets after training of 70% of samples. As per Confusion matrix performance measures are ACC (accuracy), ER (error rate), FPR (false positive rate), and DR (detection rate).

	Land Sale Data				Heart-Male				Facebook			
Algorithm	**FPR**	**DR**	**ACC**	**ER**	**FPR**	**DR**	**ACC**	**ER**	**FPR**	**DR**	**ACC**	**ER**
ID3	0.11	92.65	93.53	5.46	28	67.46	70.19	29.81	33.04	78.92	79.54	18.84
CHAID	1.21	92.62	92.35	6.65	34.06	72.62	70.92	27.08	21.35	77.64	80.12	19
C4.5	0.51	95.2	96.54	3.45	31.06	78.62	76.92	23.08	16.42	82.38	83.34	16.08
C-RT	1.33	93.34	94.87	5.12	28	67.46	70.19	29.81	18.74	82.64	80.12	19.07
I-C4.5	0.25	96.31	96.63	3.35	31	78.68	77.01	22.8	12.18	83.64	81.54	19
Rule C4.5	0.25	96.86	97.71	2.29	27.32	81.65	80.94	19.18	8	89.64	90.12	9.67
NN (MLP)	2.14	91.35	92.51	7.48	27.43	64.25	69.48	31.13	15.9	80.12	75.29	24.27
K-NN	5.36	91.32	91.73	8.21	25.38	74.65	74.51	25.48	18.37	85.05	86.41	13.3
PLSDA	4.32	90.65	91.21	8.61	32.06	63.63	66.34	33.65	12	85.37	67.65	31.19
RBC-ORT	0.13	97.68	98.02	1.07	10.22	90.92	92.5	7.48	5.7	92.35	95.34	3.92

International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-7, May 2020

Table- II: Describe the overall classification of transactions (Records) types' dataset-1 (Land Sale

RECORD/PAN TYPE	TOTAL NO.OF RECORDS WITH PAN=10220 (YES)										TOTAL NO.OF RECORDS WITHOUT PAN=784 (NO)
STATUS	E	G	F	J	B	C	T	H	A	P	NONE TYPE
BEFORE TRAINING	0	1	1	3	4	5	17	27	46	10095	784
AFTER CLASSIFICATION	5	1	4	3	5	9	5	75	55	10021	784
NO.MISS CLASSIFIED	5	0	3	0	1	4	12	48	9	74	0
NO.OF FALSE POSITIVES	95										0
OVERALL ACCURACY	99% FOR PAN YES										0

P-Personal F-Firm C-Company H-HUF A-AOP T-Trust J-Judicial E-Limited Authority B-Board Of Individuals G-Government

The table.1 showing only the performance parameters of all algorithms, where as setting up of parameters for different algorithms is not common. The time factors of training and testing also considered and it is depending on algorithm selected, parameters, and sample selection, no of attributes, data type and size of data. The ID3 take less time contrast to remaining, due to its simplicity it will take less time and PLSDA, CHAID, C-RT, C4.5 and Neural Networks consumes time for additional overhead for setting of intermediate parameters and dynamic change of rules necessary in some cases. The ID3 algorithm use the splitting criteria based on size, tree depth, leaves and nodes. The CHAID also similar and showing equal performance equal to ID3, The C4.5, I (improved)-C4.5, other rule based C4.5 are based on intermediate parameters such as confidence level (or p-value), splitting criteria, pruning criteria, and dynamic rules. The NN uses MLP technique with intermediate parameters such as no. of iterations to be required, learning rate, and no. nodes per each layer, validation criteria, error rate and necessary of hidden layer (HL).

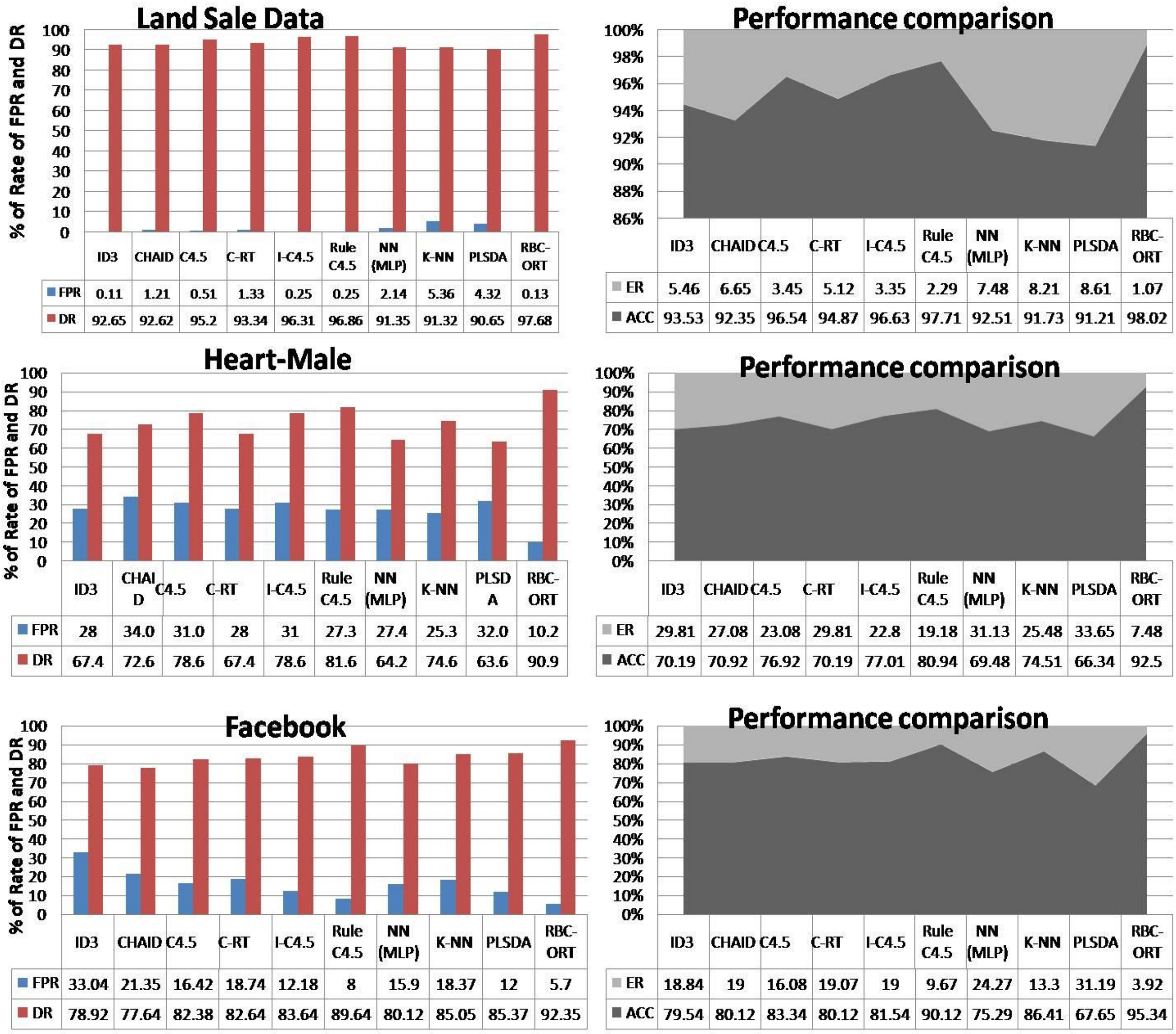
The Table.2 is showing the real dataset information based on basic classification model and simulated results are shown in Table.1. The practical outcomes are datasets visualized in Fig.2, showing the average comparison of ingredients of performance related to confusion matrix of ADS and each ML algorithm ROC is plotted with DR and FPR with bar charts.

We observe that our algorithm rule based decision tree and ordered regression tree (RBC-ORT) is showing maximum AUC also due to classification of the instances accurately and assumed classified correctly for first land dataset one, on the other side of coin , the performance is not much equal and little bit down for dataset-2 and dataset-3. Our latest rule based algorithm is -showing good performance like low FPR compare to all and accuracy is good with respect to all.

A. Discussion on Results

From above Table.1 and Fig.2, all the experimental results are listed and compared to related classification algorithms. The proposed new RBC-ORT learning algorithm is showing much more accurate results due to clear data without many anomalies. The results of DT based algorithms are much better than non-DT algorithms. The comparison of RBC-ORT is very close to Rule C4.5 results, because of its strong rule set formation and it is close to its predecessors such as ID3, C4.5. So, the combination of traditional rules and Random Tree (RT) always giving accurate and efficient performance in terms of high accuracy and low FPR. The C4.5 rule based one is also performing well expectations for dataset-1 at C.L=25, 96.86% is recorded and almost equal to our algorithm.

Let coming to dataset-2 (Heart), all the ML algorithms are not well and giving unfortunate performance due its nature of dataset and mixed combinations. The FPR and ER are very high for this set, CHAID has highest FPR with 34.06 and PLSDA is recorded highest error rate compare to all. But our algorithm has scored maximum performance by classifying all samples with adorned rule sets. With this DR=90.92, ACC=90.92 is recorded and also see the low FPR and ER. The C4.5 and its successors also recorded good results for dataset-2. The third dataset is facebook, it is complex dataset, but due to its nature of variables set and values are help to score good results compare to dataset-2. For this one, the ID3 is showing poor performance, FPR=33.04 it is very compare to all, where as the ER is high for PLSDA algorithm. The accuracy is 95.34 and DR=92.35 are scored. The final observation on results analysis is showing that the Dataset-1 is showing better results than remaining datasets.



Land Sale Data	ID3	CHAID	C4.5	C-RT	I-C4.5	Rule C4.5	NN (MLP)	K-NN	PLSDA	RBC-ORT
FPR	0.11	1.21	0.51	1.33	0.25	0.25	2.14	5.36	4.32	0.13
DR	92.65	92.62	95.2	93.34	96.31	96.86	91.35	91.32	90.65	97.68

Performance comparison	ID3	CHAID	C4.5	C-RT	I-C4.5	Rule C4.5	NN (MLP)	K-NN	PLSDA	RBC-ORT
ER	5.46	6.65	3.45	5.12	3.35	2.29	7.48	8.21	8.61	1.07
ACC	93.53	92.35	96.54	94.87	96.63	97.71	92.51	91.73	91.21	98.02

Heart-Male	ID3	CHAID	C4.5	C-RT	I-C4.5	Rule C4.5	NN (MLP)	K-NN	PLSDA	RBC-ORT
FPR	28	34.0	31.0	28	31	27.3	27.4	25.3	32.0	10.2
DR	67.4	72.6	78.6	67.4	78.6	81.6	64.2	74.6	63.6	90.9

Performance comparison	ID3	CHAID	C4.5	C-RT	I-C4.5	Rule C4.5	NN (MLP)	K-NN	PLSDA	RBC-ORT
ER	29.81	27.08	23.08	29.81	22.8	19.18	31.13	25.48	33.65	7.48
ACC	70.19	70.92	76.92	70.19	77.01	80.94	69.48	74.51	66.34	92.5

Facebook	ID3	CHAID	C4.5	C-RT	I-C4.5	Rule C4.5	NN (MLP)	K-NN	PLSDA	RBC-ORT
FPR	33.04	21.35	16.42	18.74	12.18	8	15.9	18.37	12	5.7
DR	78.92	77.64	82.38	82.64	83.64	89.64	80.12	85.05	85.37	92.35

Performance comparison	ID3	CHAID	C4.5	C-RT	I-C4.5	Rule C4.5	NN (MLP)	K-NN	PLSDA	RBC-ORT
ER	18.84	19	16.08	19.07	19	9.67	24.27	13.3	31.19	3.92
ACC	79.54	80.12	83.34	80.12	81.54	90.12	75.29	86.41	67.65	95.34

Fig.2. Shows the performance related parameters and comparison for three datasets and ML algorithm on X-axis along with value and rate of performance on Y-axis to be taken in % (as per 100%) : All left figures shows the FPR and DR for datasets and all right shows the performance comparison of all datasets by Accuracy and Error Rate.

VIII. CONCLUSION

The Results analysis indicating that the comparison of performance is based on benchmark algorithms in machine learning. In this paper, the adorned approaches are giving maximum outcomes of performance related to anomalous behavior in terms of accuracy almost above 90% and low false positive rate approximately below 10%. The method we adapted in this paper is fully novel approach and helpful to real time datasets to classify. The datasets features may affect the results originally. The results are shown in our paper in terms of %. Finally our method used in this paper is little bit complex in time factor, by observations on work showing delay in large datasets. The renovated model of ordered regression tree with ensemble combinations extract maximum outcomes in performance in future work.

.

REFERENCES

1. Denning DE, An IDS Model, In IEEE Transactions on Software Engineering, CA, 1987.
2. Stefan Axelsson, IDS : A Survey and Taxonomy, Chalmers University, Technical Report 99-15, March 2000.
3. Liu, F. T., Ting, K. M., & Zhou, Z. H. “Isolation forest In Data Mining”, 2008. ICDM’08. Eighth IEEE International Conference (pp. 413–422) 2008.
4. W. Lee and S. J. Stolfo, Data mining approaches for intrusion detection, 7th USENIX Security Symposium, Berkeley, CA, USA, 1998, pp. 79-94.
5. T. Lane and C. E. Brodley, An Application of Machine Learning to Anomaly Detection, Proceedings of the 20th National Information Systems Security Conference, pp 366-377, Baltimore, Oct. 1997.
6. Breiman L, Random Forests, Machine Learning vol. 45, pp. 5-32, 2001.

7. Adam Tauman Kalai, "Learning nested halfspaces and Uphill decision tree". ICCLT-2007 pp.378-392 LNCS Vol.4539.
8. Zhiruo Zhao, Kishan G. Mehrotra, Chilukuri K. Mohan "Online Anomaly Detection Using Random Forest"International Conference (IEA/AIE) 2018: Recent Trends and Future Technology in Applied Intelligence pp 135-147, LNCS book series , volume 10868.
9. Sudipto Guha,. Nina Mishra, , Palo Alto, Schrijvers, "Robust Random Cut Forest Based Anomaly Detection" On Streams Proceedings of the 33 rd International Conference on Machine Learning, NY, USA, 2016. JMLR: W&CP volume 48.
10. Kass, Gordon V, An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics, Vol. 29, No. 2 (1980), pp. 119–127.
11. Quinlan J.R., Induction of decision trees, Machine Learning 1, 81-106, Kluwer Publishers 1986.
12. Quinlan J. R., C4.5: Programs for Machine Learning, Morgan Kaufmann, LosAltos, 1993.
13. Breiman L., Friedman J., Olshen R., and Stone C, Class'n and Regression Trees, Wadsworth Int. Group, 1984.
14. S. Jiang and W. Yu, a Combination Class'n Algorithm Based on Outlier Detection and C4. 5, Springer Publications, 2009.
15. W. W. Cohen, Fast effective rule induction, In Proceedings of the Twelfth International Conference on Machine Learning Chambery, France. 1993, pp. 115–123.
16. J.A. Michael and S.L. Gordon, Data mining technique for marketing, sales and customer support, Wiley, NY-USA, 1997.
17. W. Y. Loh and Y. S. Shih, Split selection methods for classification trees, Statistica Sinica 7, pp. 815–840, 1997.
18. R. Rakotomalala and S. Lallich, Handling noise with generalized entropy of type beta in induction graphs algorithm, Proceedings of ICCSI, pp. 25-27, 1998.
19. R. Rakotomalala, S. Lallich and S. Di Palma, Studying the behavior of generalized entropy in induction trees using a m-of-n concept , Proceedings of 3rd European Conference on KDD, pp. 510-517, 1999.
20. Prema Rajeswari and Kannan A, An active rule approach for network intrusion detection with enhanced C4.5 Algorithm, IJCNSS , 285–385, 2008.
21. Ghosh A and Schwartzbard A, A study using NN for anomaly detection and misuse detection, Reliable Software Technologies(http://www.docshow.net/ids/useni x_sec99.zip).
22. www.mathworks.in/products/matlab/.
23. eric.univ-lyon2.fr/~ricco/sipina.html and http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html.
24. Amor S. Benferhat, and Z. Elouedi, Naive Bayes vs Decision Trees in Intrusion Detection Systems, (SAC '04)-Proc. ACM Symp.Applied Computing, pp. 420-424, 2004.
25. Bache K. and Lichman, M. UCI Machine Learning Repository, CA: University of California, School of Information and Computer Science, 2013 [http://archive.ics.uci.edu/ml].
26. Goverdhan Reddy Jidiga and P.Sammulal "RBDT: The Cascading of Machine Learning Classifiers for Anomaly Detection with Case Study of Two Datasets" , Springer-AISC Journal Series. 320, 978-3-319-11218- pp 309-324, 2015.
27. A. Abdelhalim and Issa Traore, The RBDT-1 method for rule-based decision tree generation, (ECE-09-1), University of Victoria, STN CSC, Victoria, BC, Canada, 2009.

AUTHORS PROFILE

Jidiga Goverdhan Reddy, Presently working as a Lecturer (Sr. Scale) in Department of Technical Education, Hyderabad, Government of Telangana State. Also a Research Scholar at JNT University and presently pursuing Ph.D as External at JNTU, Hyderabad, India. I did my B.Tech and M.Tech in Computer Science and Engineering from JNTU University, Hyderabad.

Sammulal Porika, Professor, Presently working at CSE Dept, JNTUH College of Engineering, Nachpally, Jagtial, affiliated to JNTU Hyderabad, Telangana State , India. He did his Ph.D from OU Hyderabad, Telangana State, sammulalporika@gmail.com