International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Filter Based Hybrid Decision Tree Construction Model For High Dimensional Anomaly Classification

Y.A.Siva Prasad, G.Rama Krishna

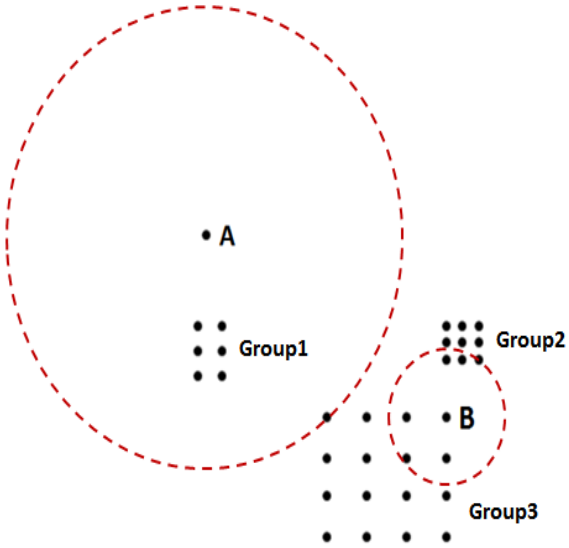
Abstract*: Anomaly discovery from the database is a process of filtering uncertain features , so that it can be used wide variety of applications. Anomaly detection on the complex data must take a long time due to the large number of features. In this proposed work, we extended the anomaly detection accuracy in distributed databases using multi-objective distributed decision tree algorithm. Proposed algorithm uses distributed entropy measure for selecting relevant attributes from the databases. Multi-Objective mechanism provides sensitiveness within the attributes as well as on the decision classes. Multi-Objective process introduces lower and upper bound mechanism for each node in the decision tree construction to preserve the data values in the decision rules. Experimental result performs well against different distributed datasets in terms of time and accuracy.

***Index Terms*: Data Mining, Patterns, Outliers.**

I. INTRODUCTION

Anomaly detection is categorized into two subcategories: supervised, or unsupervised. Supervised Anomaly detection determines the class of an observation from the classifiers. The classifiers are classified as the machine learning models whose parameters are learned from the training dataset. The primary challenge in constructing the classifiers is the skewness in the dataset between the normal class and anomaly. Due to the reason that the training data for anomalies is too small relatively based on the data for the normal class. . In the absence of training data for the Anomalies, the semi supervised methods may be used. In this instance, a machine learning model is chosen to capture the boundaries of the normal class. A new observation that falls outside of the boundaries is classified as an anomaly. Anomaly detection techniques effort to find the objects that might be different from the rest of the data objects in a given data set. Usually, anomalies are generated from certain misbehavior of the data points which get very different from the majority of data set. Random samples which are greatly deviates from its neighbors in relation to its local compactness is treated as outlier. The compactness is measured via the length of the k-nearest neighbor distances of its neighbors. Although a local outlier might not exactly differ from all other observations. Statistical approaches are the standard algorithms applied to outlier detection. The main aim of these approaches is that normal data objects follow a generating mechanism and abnormal objects deviate through generating mechanism. Given a certain type of statistical distribution , algorithms compute the parameters assuming all data points are generated by such a distribution (mean and standard deviation). Outliers are points that possess a low probability to be generated by the entire distribution (deviate greater than 3 times the standard deviation from the mean). These methods have the limitation that they will assume the data distribution . Another limitation of the statistical methods is that they don't scale well to large datasets or datasets of large attributes.

Density Based Anomaly Detection:



Red dash circles contain the k nearest neighborhood of 'A' and 'B' when k value is 7. Point outliers is defined as Separate data objects that are different with respect to the rest of dataset.Collective outliers is defined as "A set of data objects is considered as anomalous with respect to the entire data set, then members of the set are called collective outliers".The local anomalies have achieved much attention recently. The density based models can solve this issue well. And many density based outlier detection models have been proposed. These anomalies are frequently treated as noise that needs to be removed from a dataset. In the case of numerical features, models designed for categorical data often use discretization approaches to map intervals of continuous space into discrete values.

Revised Manuscript Received on 30 March 2019.
***** Correspondence Author
Mr.Y.A.Siva Prasad, Research Scholar, CSE Department, Koneru Lakshmaiah Education Foundation,Guntur,AndhraPradesh,India.
Dr.G.RamaKrishna, Professor,CSE Dept, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Guntur,AndhraPradesh,India.

Among unsupervised models, those that do not use class information, efficient approaches include equal frequency or equal width binning are used. The second challenge has become the focus in the literature with large and distributed nature of the datasets. Most of the anomaly detection models are designed to tackle continuous dataset. Since they need efforts to transform numerical to binary and categorical to number data. Also, most of the existing models are not directly applicable to categorical datasets. Different approaches based on similar concepts have been proposed to detect outlier transactions.

Anomaly detection may refer to an unsupervised models that produces a data mining mdoel for identifying instances that deviate from the normal in a data. An anomaly detection model is a one-class classification which is used to describe the feature relationship in the distributed data. The basic anomaly detection models on the whole follow the following steps:

• Identify the number of feature vectors to classify the instances.

• Determine the metric to compute the degree of deviation from the instance set.

• Set some threshold measure which exceeds the metric computation is considered as anomaly.

The application of k-means clustering model along with the outlier detection technique has very low true positive rate. Later, k-means with PCA model was implemented to find the feature based anomaly detection. As the volume of data and features increases, traditional data mining models fail to detect the anomalies for boosting the accuracy and efficiency [6].

Problems in anomaly detection models:

The key challenges of traditional anomaly detection models on various applications include: medical, creditcard, stock market and other complex realtime applications include:

1. Type of the anomaly:It indicates the variation in a value or context anomaly when a value is normal or abnormal.
2. Data type: Data can be uni-variate or multi-variate according to the number of features and its types.
3. Training data: Type of input training data and its size.

II. LITERATURE SURVEY

Knorr et al. [3] proposed the DB(pt, dt) Outlier detection scheme, wherein an object obj is said to be to get an outlier if at the very least fraction pt of the total objects have greater than dt distance to obj. They defined several techniques to find such objects. For instance the index based approach computes distance range using spatial index structure and excludes an object if its dt-neighbourhood contains greater than 1−pt fraction of total objects. They proposed nested loop algorithm to avoid the cost of building an index. They additionally proposed growing a grid so that any two objects beginning with the same grid cell have a distance of the most dt to one another. In this way objects ought to be in relation to those from neighboring cells to examine if they're outliers.

Deviations from the normal indicate anomalies that are then assumed to be an intrusion or attack. Different modeling approaches have included statistical methods [2], rule based systems, neural networks [4] and other soft-computing techniques [3].

Anomaly detection approaches can be classified into three main domain areas as shown in Figure 2. Knowledge based intrusion detection systems and statistical based intrusion detection systems.

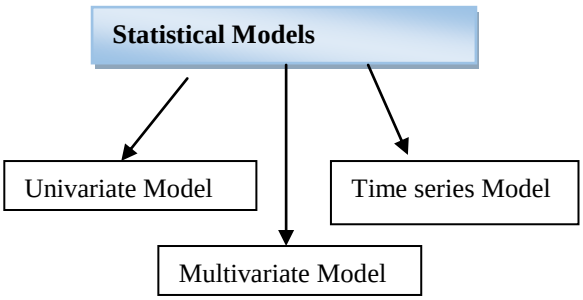


Fig 1: Statistical anomaly detection

Some of the major issues in these models should be pointed out: First, each model should be trained by an expert in such a way that the network packets generated during the attack are considered as normal type. Second, initializing the different values or metrics is a highly difficult task, especially in the case of false positive and true negative patterns are detected.

Association Rule Mining Models

Association rule mining is one of the most widely used approaches in data mining technology and also used for network intrusion detection system. Association patterns define the relationship between the anomaly and normal features using the statistical support and confidence measures.There are two main phases to detect the network attacks using association rule mining techniques. Initially, it detects the candidate patterns using the credit card data and minimum support measure. Then, it constructs the frequent patterns using the minimum support and minimum confidence thresholds. These frequent patterns are evaluated to find the interesting relationships among the network features. Traditional models such as Apriori, FPgrowth, CPTree ,etc are used as network intrusion detection systems[5].

Sequential Pattern Analysis

Similar to association rule mining models , sequential models are designed for the purpose of mining the source of credit frauds and its associated root access with respect to time. There are two important features in the sequential mining process, such as the time gap between the transaction events and duration of the transaction patterns[6-7].

Clustering Models

The basic idea of clustering is to group the similar type of network patterns into meaningful subclasses so that the objects within the same cluster are most similar and the objects from different clusters are quite different from each other. Basically, clustering models are classified into four categories they are : hierarchical techniques, partitioning techniques ,grid based techniques and density based techniques[8-9]. Entropy based clustering measure is used to detect the key feature ranking for attack detection process.

Entropy measure is computed within the cluster to evaluate the attack's cluster consistency. So, the higher the cluster consistency the smaller the cluster entropy within the cluster. Further, the more the anomaly data are, the higher the inter cluster entropy measure.Center based partitioning techniques such as k-means and k-medoids are the basic clustering methods due to their balancing and partitioning mechanism. K-medoids approach is more robust than traditional K-means based clustering algorithm.

III. PRO PROPOSED SYSTEM

When mining the different real-time data using pattern mining algorithms, it needs to preprocess the data, and then receives training dataset by analyzing the features of data used in anomaly detection If we have good algorithms, and not high-quality training data, the detective result will be not good. Different from other application fields, anomaly detection usually uses some artificial intelligence methods, which analyze data by choosing a model. However choosing models always depends on instinct and expert knowledge, and there isn't an objective method to evaluate the data.
In this proposed approach a statistical control chart algorithm is used in order to find the anomalies in different continuous datasets. We proposed dynamic 3 sigma based control chart to detect anomalies in the dataset. Basic flow structure of the proposed algorithm is shown below:

Databases:

In this part, different datasets are taken as input to preserve anomaly patterns against decision rules.

Let $D_1, D_2, D_3 \ldots D_n$ be the distributed user data. Each dataset has different combinations of numerical, nominal, interval or ratio scaled attributes. Each database has one trained data along with or without test data. Each dataset is represented as D= { $\tau_1 \tau_2 \tau_3 \ldots \tau_n$ } with attributes A={ $at_1 at_2 \ldots at_n$ }.

Data Preprocessing algorithm

Database D,
For each data record in D
Do
For each feature F in the record
Do
If(F!=NULL)
Then
Continue;
Else
F_type=check_type(F);
If(F_type==numerical)
Then

Miss_Value= $\dfrac{Max(F) * \sigma_F^2 - Min(F) * \mu^2_F}{N(N-1) * [\mathrm{Max(F)} - \mathrm{Min(F)}]}$

Value(F)=Miss_Value;
End if
If(type==Categorical)
Then
Freq[]=frequency(F);// each category of class attribute.
Probability of each instance value per class.

Prob[]= $\sum_{i=1}^{m} \Pr ob(x_j / C_i)$;
i=1,2,3…m classes
j=1,2…n instances
rank=Max{freq[]}/Max{Prob[]};
Fill the value with the max ranked class value.
End if
Done
Done

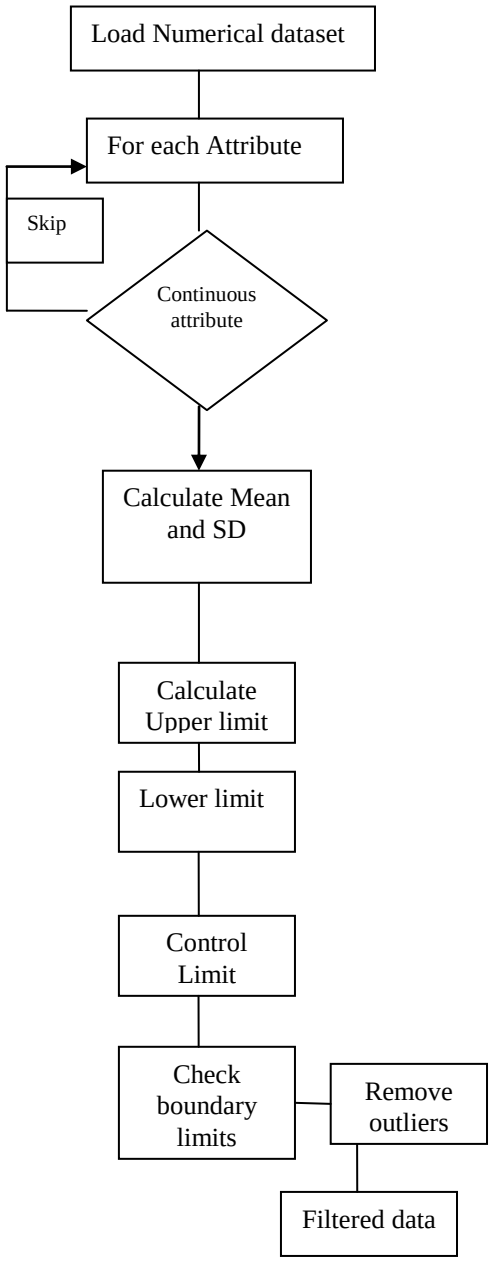
In this algorithm, missing values or inconsistent values are replaced with the computed value. If the attribute is numerical then all the missing values are replaced with the computed Max-Min value. If the attribute is categorical , then all the missing values are replaced with probabilistic ranked value.

Outlier_removal(List)

- For each attribute in List
- Do
- Statistical Control limits to eliminate outliers
- Lower Control limit: $\mu_X - \lambda\sigma_X$
- Upper Control Limit: $\mu_X + \lambda\sigma_X$
  - Control Limit: μ_X
- Done

Procedure:

Input: Continuous dataset
Output: Dataset without anomalies.
Procedure:
Step 1:Load dataset with continuous attributes.
Step 2: Check each attribute in the dataset as real attribute or not.
Step 3: Calculate mean and standard deviation of each attribute.
Step 4: Calculate upper control limit of each attribute (2)
Step 5: Calculate control limit of each attribute (3).
Step 6: Calculate lower control limit of each attribute (1)
Step 7: Check whether each object in the dataset falls within three categories i.e lower, upper or control limits.
Step 8: If the object is out of bound then it is removed from the dataset instances.
Step 9: This process is repeated until all data points are completed.
Step 10:Finally dataset without outliers are stored in file.



Distributed anomaly detection model such as C4.5 is an improved version of decision trees over C4.5 from the training data, using the concept of information entropy. The training data is a set d1, d2,d3 .... dn data objects in the dataset D. Each di = xl,x2,... is a sample values where xl,x2,... represents features or attributes of the sample. The training data associated with a vector C = cl,c2,... where cl,c2,...cn represents the class to which each sample belongs to dataset. Every node of the decision tree, chooses one attribute of the data the most efficiently splits its range of samples into subsets in a single class. The attribute with the highest calculated information gain is selected to get the decision attribute. In this proposed approach the most relevant split attribute is given input to sensitivity to preserve patterns. Each most relevant split attribute is checked against multi-objective model to preserve the patterns.

Most relevant attribute is calculated by using following distributed entropy measure:

DEM(DistributedEntropyMeasure)

(D)= $-D_i \sum_{i=1}^{m} l \log \sqrt[\alpha]{D_i}$,m different classes

DEM(DistributedEntropyMeasure) (D)= $-D_i \sum_{i=1}^{m} l \log \sqrt[\alpha]{D_i}$

$= -D_1 \log \sqrt[\alpha]{D_1} + D_2 \log \sqrt[\alpha]{D_2} \ldots\ldots D_n \log \sqrt[\alpha]{D_n}$

Where D_1 indicates set of samples which belongs to target class C_1, D_2 indicates set of samples which belongs to target class C_2 and so on. α is the sensitive factor of the attributes. Distributed Entropy to each attribute is calculated using

$$DistributedInfo_A(D) = \sum_{i=1}^{v} |D_i| / |D| \times DEM(D_i)$$

The term Di is the ith partition data. DEM(D) is the expected information required to classify a tuple from D based on the partitioning by A.

IV. EXPERIMENTAL RESULTS

All experiments were performed with the configurations Intel(R) Core(TM)2 CPU 2.13GHz, 2 GB RAM, and the operating system platform is Microsoft Windows XP Professional (SP2).

Numerical Data : Realtime Network Capture Data

```
Sequence != 651862017 AND SIP != /34.217.184.213 AND Sequence != 278444200  ->  Class != attack
Sequence != 278422440 AND Window != 56780 AND Window != 44200 AND SIP != /54.240.227.37  ->  Class != attack
Sequence != 651862017 AND SIP != /34.217.184.213 AND Sequence != 278444200  ->  Class != attack
Sequence != 234205687 AND SIP != /54.240.227.37 AND Window != 35700  ->  Sync != FALSE
Window != 64189 AND SIP != /54.240.227.37 AND Length != 221  ->  Sync != FALSE
Window != 16468 AND Sequence != 278524440 AND Sequence != 1582698569  ->  Sync != FALSE
Window != 64189 AND SIP != /54.240.227.37 AND Sequence != 234205687  ->  Sync != FALSE
SIP != /54.240.227.37 AND Length != 1030 AND Sequence != 458784894  ->  Sync != FALSE
Sequence != 675694783 AND Window != 16468 AND Sequence != 278767880  ->  Sync != FALSE
SIP != /54.240.227.37 AND Length != 1030 AND Window != 64189  ->  Sync != FALSE
Sequence != 675694783 AND Window != 16468 AND Length != 291  ->  Sync != FALSE
Sequence != 1980722091 AND Sequence != 279095640  ->  Window != 65280
Window != 64189 AND Sync != FALSE  ->  DIP != /151.101.65.69
```

Sequence != 1509162484 AND Sequence != 1258573967 -> Window != 51680
Length != 221 AND Class != attack AND Window != 13674 -> Sequence != 1582698569
Window != 16358 -> Length != 61
Window != 16358 -> Length != 61
Sequence != 1509162484 AND Sequence != 279095640 -> Window != 65280
DIP != /202.58.9.200 AND Length != 720 -> Window != 5877
SIP != /54.240.227.37 -> DIP != /151.101.65.69
Sequence != 1582698569 AND Sequence != 2038372474 -> Length != 221
Length != 720 AND Sequence != 1645719254 -> Window != 16363
Window != 16363 AND Window != 65280 -> Sequence != 1582698569
Sync != FALSE AND Window != 74 AND Window != 13674 -> Sequence != 1582698569
Window != 5877 AND Sequence != 2038372474 -> Length != 221
Window != 14661 AND Window != 65166 -> Length != 285
DIP != /151.101.65.69 AND Sequence != 279095640 -> Window != 65280
Window != 16358 AND Class != attack AND Sync != FALSE -> Sequence != 1582698569
Sync != FALSE AND Length != 777 AND Window != 13674 -> Sequence != 1582698569
SIP != /54.240.227.37 AND Sync != FALSE -> Window != 16363
Window != 16511 AND Window != 74 -> Sequence != 1582698569
Sequence != 675694783 AND Window != 16468 AND Window != 16505 -> Class != attack
DIP != /103.231.98.196 -> Sync != FALSE
SIP != /103.102.166.224 AND Window != 15989 -> Class != attack
Sequence != 1582698569 AND SIP != /123.201.147.16 -> Class != attack
SIP != /123.201.147.16 AND Window != 65166 -> Sync != FALSE
DIP != /103.231.98.196 -> Class != attack
Window != 14280 AND Sequence != 278524440 AND SIP != /123.201.147.16 -> Sync != FALSE
SIP != /123.201.147.16 AND SIP != /54.240.227.37 -> Class != attack
Sequence != 234205687 AND Class != attack AND DIP != /103.231.98.196 -> Sync != FALSE
Window != 65280 AND SIP != /23.20.239.12 -> Sync != FALSE
Window != 65280 AND Sequence != 278524440 AND DIP != /103.231.98.196 -> Sync != FALSE
DIP != /151.101.154.133 AND Sequence != 278524440 AND Sequence != 1582698569 -> Sync != FALSE
DIP != /151.101.154.133 AND Sequence != 278524440 AND Sequence != 1582698569 -> Sync != FALSE
Class != attack AND SIP != /54.240.227.37 AND DIP != /103.231.98.196 -> Sync != FALSE
Sequence != 858769696 AND Sequence != 278524440 AND DIP != /103.231.98.196 -> Sync != FALSE
SIP != /23.20.239.12 AND DIP != /151.101.65.69 -> Class != attack
DIP != /103.231.98.196 AND Window != 5877 AND Length != 777 -> Sync != FALSE
Sequence != 1676069360 AND Length != 1030 AND DIP != /103.231.98.196 -> Sync != FALSE
DIP != /151.101.65.69 AND Length != 1030 AND DIP != /103.231.98.196 -> Sync != FALSE
DIP != /151.101.65.69 AND Length != 1030 AND DIP != /103.231.98.196 -> Sync != FALSE
DIP != /151.101.65.69 AND Length != 1030 AND DIP != /103.231.98.196 -> Sync != FALSE
DIP != /151.101.65.69 AND Length != 1030 AND DIP != /103.231.98.196 -> Sync != FALSE
Sequence != 234205687 AND SIP != /54.240.227.37 AND DIP != /103.231.98.196 -> Sync != FALSE
Window != 10540 AND Length != 1030 AND DIP != /103.231.98.196 -> Sync != FALSE
Class != attack AND Window != 16468 AND DIP != /103.231.98.196 -> Sync != FALSE
Window != 64189 AND SIP != /54.240.227.37 AND DIP != /103.231.98.196 -> Sync != FALSE
Sequence != 278422440 AND Window != 56780 AND Window != 44200 AND DIP != /64.233.161.120 -> Class != attack
Sequence != 278422440 AND Window != 56780 AND Window != 44200 AND SIP != /107.178.254.65 -> Class != attack
Class != normal AND Window != 15904 AND Length != 594 -> Sync != FALSE
Sequence != 858769696 AND Sequence != 278524440 AND Class != normal -> Sync != FALSE
Length != 777 AND Window != 56780 AND Class != normal -> Sync != FALSE
Class != normal AND Length != 1030 AND DIP != /103.231.98.196 -> Sync != FALSE
Window != 65280 AND Class = normal -> Sync != FALSE
Window != 65280 AND Class = normal -> Sync != FALSE

=== Evaluation ===

Elapsed time: 157.993s

Number of Iterations :33
F-Measure: 0.96360
Recall : 0.96553
TP rate : 0.9552
FP rate : 0.04479999999999995
Classification Accuracy 0.95627
Error rate 0.10186

Result 2: Mixed Attributes : Somatic Cancer data

fre >= 0.01 AND pattern != TG AND polyphen != possibly -> inExAct != false
CNT <= 78.0 -> isSomatic != false
mutAss != stoploss AND pattern != TG -> dbSNP != false
CNT <= 78.0 -> pattern != CA
fre >= 0.01 AND pattern != CA AND dbSNP != false AND SeqContent != CTA -> isSomatic != false
inExAct != false AND inExAct != false -> dbSNP != false
inExAct != false AND inExAct != false -> SeqContent != CTA
VAF >= 8.7 -> isSomatic != false
SeqContent != CGC AND mutAss != stoploss AND polyphen != possibly -> inExAct != false
CNT <= 78.0 -> VAF >= 8.7
mutAss != stoploss -> SeqContent != ACG
fre >= 0.01 -> inExAct != false
pattern != CA AND SeqContent != ACG -> dbSNP != false
CNT <= 78.0 -> isSomatic != false
inExAct != false -> pattern != CA
fre >= 0.01 -> VAF >= 8.7
fre >= 0.01 -> CNT <= 78.0
dbSNP != false -> CNT <= 78.0
fre >= 0.01 AND pattern != CA -> SeqContent != CTA
inExAct != false AND mutAss != stoploss -> dbSNP != false
CNT <= 78.0 AND pattern != TG AND polyphen != possibly -> inExAct != false
pattern != CA AND inExAct != false -> dbSNP != false
dbSNP != false -> isSomatic != false
dbSNP != false -> mutAss != stoploss
mutAss != stoploss AND SeqContent != CTA -> isSomatic != false
CNT <= 78.0 AND SeqContent != CTA -> isSomatic != false
CNT <= 78.0 -> fre <= 0.96
pattern != CA AND pattern != TG AND dbSNP != false -> inExAct != false
CNT <= 78.0 -> inExAct != false
CNT <= 78.0 AND pattern != TG AND polyphen != possibly -> inExAct != false
polyphen != possibly AND pattern != CA -> dbSNP != false
CNT <= 78.0 AND polyphen != possibly -> SeqContent != CTA
VAF >= 8.7 -> fre >= 0.01
mutAss != stoploss AND CNT <= 78.0 -> dbSNP != false
polyphen != possibly -> inExAct != false
mutAss != stoploss -> dbSNP != false
fre <= 0.96 AND VAF >= 8.7 -> mutAss != stoploss
fre >= 0.01 -> SeqContent != ACG
pattern != CA -> dbSNP != false
fre >= 0.01 AND mutAss != stoploss -> inExAct != false
CNT <= 78.0 AND pattern != TG AND dbSNP != false -> inExAct != false
dbSNP != false AND SeqContent != CTA -> isSomatic != false
SeqContent != CGC AND mutAss != stoploss AND polyphen != possibly -> inExAct != false
SeqContent != CTA AND isSomatic != false -> mutAss != stoploss
isSomatic != false -> SeqContent != ACG
inExAct != false -> mutAss != stoploss
isSomatic != false AND pattern != TG -> dbSNP != false
VAF >= 8.7 -> pattern != CA
SeqContent != CTA AND inExAct != false -> dbSNP != false
polyphen != possibly AND mutAss != stoploss -> SeqContent != CTA
pattern != CA -> CNT <= 78.0
mutAss != stoploss -> CNT <= 78.0
VAF >= 8.7 -> inExAct != false
CNT <= 78.0 -> pattern != CA
VAF >= 8.7 -> inExAct != false
pattern != CA -> dbSNP != false

=== Evaluation ===

Elapsed time: 98.265s

Number of Iterations :35
F-Measure: 0.96440
Recall : 0.96382
TP rate : 0.95869
FP rate : 0.04130999999999996
Classification Accuracy 0.96841
Error rate 0.12838

Table 1: Comparison of proposed model to traditional models on mixed attributes.

Models	Precision	TruePositive	MSE
ID3	0.864	0.856	0.264
C45	0.893	0.873	0.215
CART	0.914	0.9024	0.164
CART	0.934	0.9253	0.089
ProposedModel	0.986	0.9793	0.0354

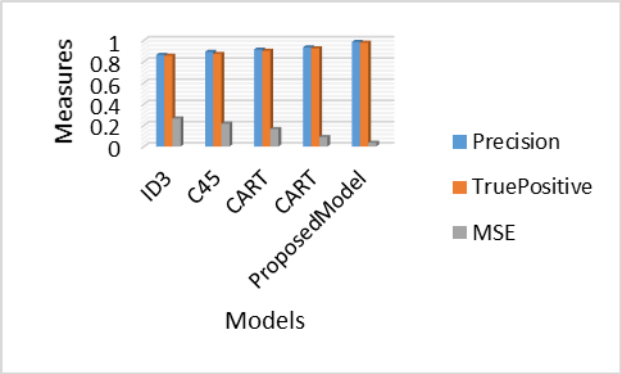


Figure 2: Comparison of proposed model to traditional models on mixed attributes.

Table 1: Comparison of proposed classifier to existing classifiers with different datasets

Accuracy					
Datasets	ID3	C45	CART	RETree	Proposed Model
Glass	0.924	0.942	0.9343	0.946	0.973
Diabetes	0.956	0.947	0.963	0.968	0.984
Credit	0.957	0.974	0.972	0.967	0.988
Census	0.974	0.964	0.9756	0.972	0.985

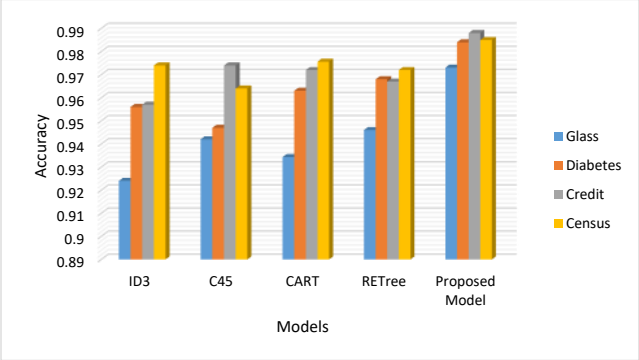


Figure 3: Comparison of proposed classifier to existing classifiers with different datasets

I. CONCLUSION

Anomaly detection play a vital role in the large number of applications with different types of feature space. In this work, a statistical control limits based decision tree construction is performed on the numerical dataset and mixed attribute datasets. Multi-Objective mechanism provides sensitiveness within the attributes as well as on the decision classes. Multi-Objective process introduces lower and upper bound mechanism for each node in the decision tree construction to preserve the data values in the decision rules. Experimental result performs well against different distributed datasets in terms of time and accuracy.

REFERENCES

1. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: Identifying densitybased local outliers,” In Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM Press, pp. 93–104, 2000.
2. Y. Tao and D. Pi, “Unifying density-based clustering and outlier detection,” 2009 Second International Workshop on Knowledge Discovery and Data Mining, Paris, France, pp. 644–647, 2009.
3. E. M. Knorr and R. T. Ng. Algorithms for mining distance-based outliers in large datasets. In *VLDB*, 1998.
4. Pavel Berkhin, “Survey of Clustering Data Mining
5. Techniques”,http://citeseer.ist.psu.edu/berkhin02survey.html
6. K. P. Chan and A.W. C. Fu, “Efficient time series matching by wavelets,” In Proceeding ICDE ’99 Proceedings of the 15th International Conference on Data Engineering, Sydney, Austrialia, March 23-26, 1999, p. 126, 1999..
7. V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, p. ARTICLE 15, July 2009.
8. R. Fujimaki, T. Yairi, and K. Machida, “An anomaly detection method for spacecraft using relevance vector,” in Learning, The Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). Springer, 2005, pp. 785–790.
9. Sweeney L. 2002 K-anonymity: A model for protecting Journal on Uncertainty, fuzziness and Knowledge based systems.
10. Weiwei Fang, Bingru Yang, “Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data,” In Proceedings of the International Conference on Computer Science & Software Engineering, 2008.
11. [10] S. Russell and N. Peter, Artificial Intelligence. A Modern Approach 2/E. Prentice-Hall, 2002.

12. F. Emekci* , O.D. Sahin, D. Agrawal, A. El Abbadi, “Privacy preserving decision tree learning over multiple parties
13. Y.A.Siva Prasad, Dr.G.Rama Krishna, “"Distributed Differential Privacy Preserving Mechanism on Real Time Datasets", International Journal of Applied Engineering Research(IJAER),(ISSN 0973-4562)Vol 10,number4-2015.
14. M. Xue, C. Zhu, "Applied Research on Data Mining Algorithm in Network Intrusion Detection," jcai, pp.275-277, 2009 International Joint Conference on Artificial Intelligence, 2009.
15. D. E. Denning, “An intrusion detection model,” IEEE Transaction on Software Engineering, 1987.
16. T. Bhavani et al., “Data Mining for Security Applications,” Proceedings of the 2008 IEEE/IFIP International Conference on Embedded and Ubiquitous Computing - Volume 02, IEEE Computer Society, 2008.
17. T. Lappas and K. P. ,"Data Mining Techniques for (Network) Intrusion Detection System," January 2007
18. M. Davy, F. Desobry, A. Gretton, and C. Doncarli. An online support vector machine for abnormal events detection. Signal Processing, 86:2009–2025, 2005.
19. P. Domingos and G. Hulten. Mining high-speed data streams. In Proceedings of the 6th ACM SIGKDD, 2000.
20. D.J. Hand and R.J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. Machine Learning, 45:171–186, 2001.
21. W. Duch, T. Winiarski, J. Biesiada, J, and A. Kachel, “Feature Ranking, Selection and Discretization,” Int. Conf. on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), pp. 251 – 254, 2003.
22. W. R. Veitch, “Identifying Characterstics of High School Dropouts: Data Mining with a Decision Tree Model,” Paper Presented at Annual Metting of the American Educational Research Association, San Diego, CA, 2004 (ERIC Document No. ED490086).
23. Y.A.Siva Prasad, and G. Ramakrishna” Distributed differential privacy preserving mechanism on real time datasets”, International Journal of Applied Engineering Research,(2015).
24. Y.A.Siva Prasad, and G. Ramakrishna. "A Novel Probabilistic Based Feature Selection Model For Credit Card Anomaly Detection." Journal of Theoretical & Applied Information Technology 94.2 (2016).
25. Sathish, T., Periyasamy, P., "Modelling of HCHS system for optimal E-O-L combination section and disassembly in reverse logistics", Applied Mathematics and Information Sciences, vol. 13, no. 1, pp. 57-62, 2019.
26. Sathish, T., Muthukumar, K., Palani Kumar, B., "A study on making of compact manual paper recycling plant for domestic purpose", International Journal of Mechanical and Production Engineering Research and Development, vol. 8, no. Special Issue 7, pp. 1515-1535, 2018.
27. Sathish, T., "Experimental investigation on degradation of heat transfer properties of a black chromium-coated aluminium surface solar collector tube", International Journal of Ambient Energy, vol. 1, no. 1, pp. 1-5, 2018.

AUTHORS PROFILE

Mr.Y.A.Siva Prasad, Research Scholar in CSE Department,, KLEF University, Vaddeswaram, Andhra Pradesh, and Life member of CSI, IAENG, Having 13 years of Teaching Experience, and presently working as an Associate Professor at SV College of Engineering, Tirupati, Andhra Pradesh

Dr. G.Rama Krishna, worked as a Prof., CSE dept KLEF University, carried out his research at Saha Institute, Calcutta for about five and half years from 1966-1971 in theoretical physics using computers at ISI, Calcutta, IIT Khargapur, IIT Kanpur, & IIT Madras extensively for solving problems in nuclear models and Obtained Ph.D. in 1975. Worked as a lecturer in NIT,Warangal in Physics Dept.From 1971-1975 teaching B.Tech and M.Sc.(tech) students .Taught computer programming for M.Phil (Computer Methods)students at University of Hyderabad ,Hyderabad from 1975 to 1980.Taught Computer programming and application analysis On leak detection in Oil and Gas Pipelines to the customers Of ONGC, IOC, OIL India, HPCL and BPCL at ECIL and Customer sites across India from 1975 to 2003