International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Malicious Intrusion Detection Using Machine Learning Schemes

Madhavi Dhingra, S C Jain, Rakesh Singh Jadon

Abstract*: Wireless networks are continuously facing challenges in the field of Information Security. This leads to major researches in the area of Intrusion detection. The working of Intrusion detection is performed mainly by signature based detection and anomaly based detection. Anomaly based detection is based on the behavior of the network. One of the major challenge in this domain is to identify and detect the malicious node in wireless networks. The intrusion detection mechanism has to analyse the behavior of the node in the network by means of the several features possessed by each node. Intelligent schemes are the need of the hour in such scenario. This paper has taken a standard dataset for studying the features of the wireless node and reduced the features by applying the most efficient Correlation Attribute feature selection method. The machine learning algorithms are applied to obtain an effective training model which is then applied on the testing dataset to validate the model. The accuracy of the model is determined by the performance parameters such as true positive rate, false positive rate and ROC area. Neural network, bagging and decision tree algorithm RepTree are giving promising results in comparison with other classification algorithms.

***Keywords* : Data Mining, Intrusion Detection, Classifier, Malicious**

I. INTRODUCTION

Security in Manet is very challenging as these networks follow the dynamic topology and work without any central base station. Traditional security procedures like firewalls, encryption techniques don't work here because of its features. Thus an improved, efficient Intrusion detection and prevention system is needed so as to secure the underlying system. The paper focuses on identifying the node behaviour[1]. A node in the wireless network can behave normally or abnormally. Normal behaviour is determined as - when operations are satisfying the security principles in the network. Malicious behaviour is - when a node violates any of the security principles and either is under attack or performs attack by itself. The paper has focused on the identification of malicious attacks over nodes by analysing a standard UNSW-NB 15 data set.

Revised Manuscript Received on August 30, 2019.
* Correspondence Author
Madhavi Dhingra*, Assistant Professor in Department of Computer Science and Engineering at Amity University Madhya Pradesh, Gwalior, India.
Dr. S C Jain, working as Director in Amity School of Engineering & Technology at Amity University Madhya Pradesh, Gwalior, India.
Dr. Rakesh Singh Jadon, Professor & Head in Department of Computer Applications, Madhav Institute of Technology and Science, Gwalior, (M.P) , India.

The Dataset has large number of features, due to which it is difficult to analyse the training data, thus feature selection is applied over the dataset and then classification algorithms are performed to a better training model. The paper has been divided into multiple sections. The first section discusses about the intelligent approached that were implemented so far. The second section contains the proposed work involving feature reduction procedure using correlation attribute filter and description of the used classifier algorithms. The third section has the experimental results of the classifier algorithms over training and testing dataset. The last section performs the discussion over the obtained results by considering performance parameters.

II. RELATED WORK

From the past few years, several researches have been done using intelligent approaches in the field of Intrusion detection in wireless networks. Intelligent approaches make use of different types of agents namely static agents, mobile agents, and ant based agents. Static agents are again of two types - simple agents and multi-agents. Simple agents are able to analyse the environment and perform actions according to it. A new agent based approach based on simple agents were devised by Baker for intrusion detection[2]. It has used rough sets to handle imbalanced data and generated rules from the database. The limitation is that use of rough sets generated computational overhead.

Multi agent systems are mainly used in robotic applications where agents can perform individual tasks which are independent of each other but coordinate among themselves for security of the system. A multi agent based Intrusion detection system was developed by Xiaodong Zhu which has developed an adaptive learning module that learns from the network and host audit data and also used more than one data mining technique[3]. Mobile agents are dynamic in nature and can move from one node to another. A similar concept based IDS was proposed by Ghenima Bourkache for adhoc networks that works by using nearby mobile and reactive agents[4]. Its main purpose was to find the main source of the attack and to make it isolated. and has also given an intrusion detection framework that has used mobile agents[5]. Mobile agents have been used for developing integrated architecture to assist in designing network management system for security [6]and also used for implementation of secure error free attack resilient architecture[7]. Neural networks have played a key role in design and development of Intrusion Detection Systems. An efficient approach was used for classification by Verikas and Bacauskiene[8].

They have used training of neural network with an attached error function. A multi agent intrusion detection system was proposed by Chi-Ho Tsang that has used ant based agents for anomaly detection[9]. This approach has reduced the percentage of false positives in the system. Neural networks have also been used in the process of feature selection[10], this approach has determined the correlation between the features and selected them for building neural network architectures.

An adaptive new fuzzy IDS was also developed by Jeich Mar to reduce the detection time in MAC layer of wireless network[11]. Genetic algorithm-based IDSs simplifies the analysis of real time data[12].Genetic algorithms are also being used for feature selection as they give more promising outcomes with the heuristic approach[13,14].

III. PROPOSED WORK

UNSW-NB 15 data set[15] was generated by IXIA perfect storm toll in the Cyber range lab of Australian Centre for Cyber Security (ACCS). This dataset has been created after the application of 12 algorithms and tools. TCP dump tool was used to generate 100 GB of traffic data, which contains collection of normal and abnormal activities. The dataset has total of 49 features which also contain a class feature identifying the normal or malicious activity[16,17]. 49 Features are categorized into five groups: Flow, Basic, Content, Time, and Additionally Generated.

The attributes of the dataset are categorized into 6 broad groups, the details of which are given in Table 1.

Table - I: UNSW-NB15 dataset feature categorization[16,17]

S.No.	Name of the category	Description
1	Flow features	Includes the identifier related attributes between hosts such as client-to-serve or server to-client.
2	Basic features	It containes the attributes regarding the connections of protocols.
3	Content features	It contains the attributes of TCP/IP and also contain some attributes of http services.
4	Time features	It includes the attributes of time such as round trip time of TCP protocol start/end packet time arrival time between packets *etc.*
5	Additional generated features General purpose features(from number 36 - 40)	Additional features for specific purpose
6	Connection features (from number 41- 47)	Conatins information regarding chronological order of the last time feature
7	Labelled Features	Label related features like normal or anomaly.

The target class i.e. Malicious activity is categorised in nine kinds of attacks including one normal activity class. The attacks are categorized as Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms.The work on the dataset is done on the WEKA Platform, a standard tool for data mining.

The following steps were followed in order to develop an optimal model of machine learning for identification of malicious activity on the nodes.

A. Feature Selection

Due to large number of features, the feature selection algorithm is applied to select the most relevant and promising features. The Correlation Attribute Evaluation is applied over the dataset to reduce the features. Correlation Attribute Evaluation consider the subset of features by assessing its ability to identify attacks. The CFS uses the heuristic approach and identify the best subset which can lead to best results. The correlation between the output and the attributes are computed and only this attributes are chosen that have level of correlation between moderate and high. All those attributes which are having low correlation are discarded. Weka tool performs the feature selection with Ranker search method.

B. Training and Testing of Dataset

After experimenting with the several classification approaches, the following six classification approaches are used for training and testing of dataset[18].

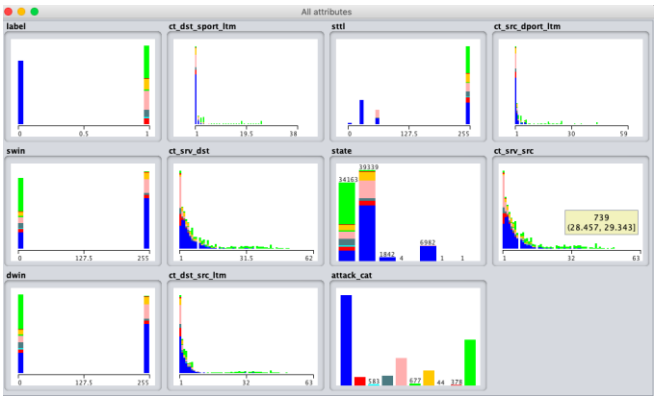
- MultilayerPerceptron - Multilayer Perceptrons is a form of neural network in which the data is provided to the input layer, where it passes from one or more hidden lanes and results are made from the output layer. Such method is suitable for the Classification problems where class is specified in the input. Data is given in the form of tables.
- SMO - SMO belongs to the Support Vector Machines that was designed for solving classification problems.It works for numeric variables and convert automatically the categorical values into numeric values. This also normalises the input data. Its basic concept is to separate the input data into two categories by a line. It uses support vector instances from the training set that are nearest to the line.
- LazyIbk - The k-nearest neighbors is called LazyIbk in Weka. It is used for clarification and regression problems both. It sorts the training dataset first and then find out the similar behavior for making prediction. It considers the difference between the training instances and gives good performance. For classification problem, the algorithm computes the mode of k similar instances from the training dataset to give predictions.
- Bagging - Bootstrap Aggregation (or Bagging for short), is a simple and very powerful ensemble method. Its basic principle is to work with multiple machine learning algorithms and combine them in order to have more better predictions in comparison with a single model. It also reduces the variance of the algorithms giving high variance. It performs bootstrap method to high variance algorithm mainly on decision trees.
- RepTree - Decision trees have several variations. RepTree is one of them. The process of work attars from the root and moving towards the leaves to evaluate the data instance.

It follows the greedy approach to select the best point that separate the data in two parts and make predictions for the suitable class. After making the tree, it is improved so as to make more efficient model. RepTree forms more than one tree following many iterations by using regression. It selects the best among all the trees which is further improved by evaluating with the mean square error. It belongs to the category of fast decision tree learner.

- RandomTree - RandomTree is a supervised classification machine learning technique which works by developing many learners. Its basic principle is to construct a decision tree which is best in terms of prediction. The random tree make use of forest which is collection of tree predictors. It takes the input, checks whether it belongs to the one of the tree of the forest and give result of the class which receives highest number of matches.

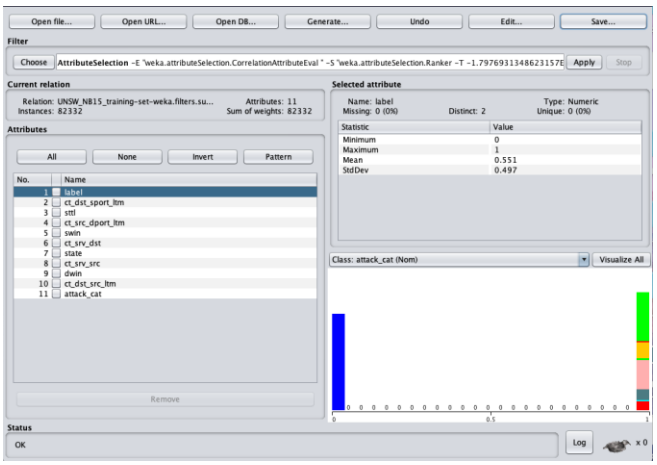
IV. EXPERIMENTAL RESULTS

The experiment is performed in the following steps -



1. Reduced Training Dataset with 7 features

Step 1 - The correlation attribute evaluation is applied on the UNSW-NB 15 data set training dataset and the features are reduced from 45 to 11 features including class feature.These features are sbytes, id, smean, sload, label, bytes, service, mean, ct_dst_sport_ltm, proto, and class



attack_cat. The total instances are 82332.

2. Reduced Training Dataset features Visualisation

Step -2 Reduced Training dataset has been classified by using the following algorithms-

a. MLP

This classifier has given 84.93% Accuracy rate with training time of 0.8 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly	69925	84.9305 %
The Instances classified incorrectly	12407	15.0695 %
The value of Kappa statistic error		0.7882
The value of Mean absolute error		0.0387
The value of Root mean squared error		0.1403

b. SMO

This classifier has given 82.37% Accuracy rate with training time of 0.56 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly	67819	82.3726 %
The Instances classified incorrectly	14513	17.6274 %
The value of Kappa statistic		0.7524
The value of Mean absolute error		0.1617
The value of Root mean squared error		0.2753

c. LazyIBK

This classifier has given 86.74% Accuracy rate with training time of 391.31 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly	71419	86.7451 %
The Instances classified incorrectly	10913	13.2549 %
The value of Kappa statistic		0.8146
The value of Mean absolute error		0.0332
The value of Root mean squared error		0.1289

d. Bagging

This classifier has given 86.45% Accuracy rate with training time of 0.42 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly	71183	86.4585 %
The Instances classified incorrectly	11149	13.5415 %
The value of Kappa statistic		0.8107
The value of Mean absolute error		0.0349
The value of Root mean squared error		0.1315

e. RepTree

This classifier has given 86.30% Accuracy rate with training time of 0.14 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly	71058	86.3067 %
The Instances classified incorrectly	11274	13.6933 %
The value of Kappa statistic		0.8084
The value of Mean absolute error		0.0351
The value of Root mean squared error		0.1324

f. RandomTree

This classifier has given 86.74% Accuracy rate with training time of 0.45 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly	71419	86.7451 %
The Instances classified incorrectly	10913	13.2549 %
The value of Kappa statistic		0.8146
The value of Mean absolute error		0.0332
The value of Root mean squared error		0.1289

Table - II: Summary of Classifier Algorithms on Training Dataset

S.No.	Classifier Algorithm	Testing Time (in seconds)	Accuracy (%)
1	MLP	0.8	84.93
2	SMO	0.56	82.37
3	LazyIBK	391.31	86.74
4	Bagging	0.42	86.45
5	RepTree	0.14	86.30
6	RandomTree	0.45	86.74

Comparing the Accuracy rate of all these algorithms, it has been observed that LazyIBK, Bagging and RepTree are performing better than other classifier algorithms.

Step -3 To perform testing, the testing dataset must have the same features as of training dataset. Thus, the testing dataset from is taken and reduced to 7 features directly. The total instances in the testing dataset are 175341.In the weka tool, the test dataset is supplied to the classifier algorithm and the results achieved are as below-

The Instances classified correctly 134501 76.7082 %
The Instances classified incorrectly 40840 23.2918 %
The value of Kappa statistic 0.7044
The value of Mean absolute error 0.0538
The value of Root mean squared error 0.1694

b. SMO

This classifier has given 75.79% Accuracy rate with training time of 23.89 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly 132898 75.794 %
The Instances classified incorrectly 42443 24.206 %
The value of Kappa statistic 0.693
The value of Mean absolute error 0.1624
The value of Root mean squared error 0.2765

c. LazyIBK

This classifier has given 77.40% Accuracy rate with training time of 1792.85 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly 135715 77.4006 %
The Instances classified incorrectly 39626 22.5994 %
The value of Kappa statistic 0.7131
The value of Mean absolute error 0.0518
The value of Root mean squared error 0.1756

d. Bagging

This classifier has given 78.03% Accuracy rate with training time of 22.38 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly 136835 78.0394 %
The Instances classified incorrectly 38506 21.9606 %
The value of Kappa statistic 0.721
The value of Mean absolute error 0.0519
The value of Root mean squared error 0.1682

e. RepTree

This classifier has given 78.03% Accuracy rate with training time of 19.12 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly 136827 78.0348 %
The Instances classified incorrectly 38514 21.9652 %
The value of Kappa statistic 0.7211
The value of Mean absolute error 0.0515
The value of Root mean squared error 0.1688

f. RandomTree

This classifier has given 77.19% Accuracy rate with training time of 19.14 seconds. The detailed results with confusion matrix are as follows -

The Instances classified correctly 135355 77.1953 %
The Instances classified incorrectly 39986 22.8047 %
The value of Kappa statistic 0.7104
The value of Mean absolute error 0.0521
The value of Root mean squared error 0.1765

Table - III: Summary of Classifier Algorithms on Testing Dataset

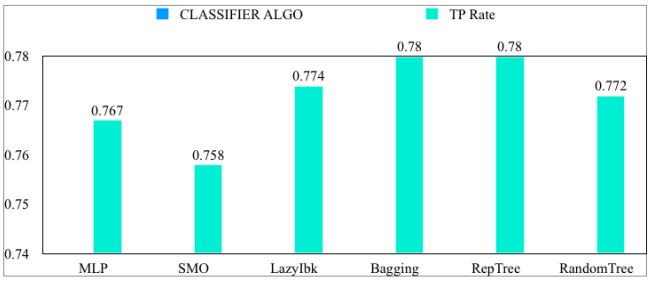
S.No.	Classifier Algorithm	Testing Time (in seconds)	Accuracy (%)
1	MLP	21.64	76.70
2	SMO	23.89	75.79
3	LazyIBK	1792.85	77
4	Bagging	22.38	78
5	RepTree	19.12	78
6	RandomTree	19.14	77.19

The results have clearly shown that the classifiers Bagging and RepTree and RandomTree have performed better with testing dataset also. The Overfitting does not exist for all the cases as testing accuracy is less than training accuracy with respect to each classifier.

V. DISCUSSION

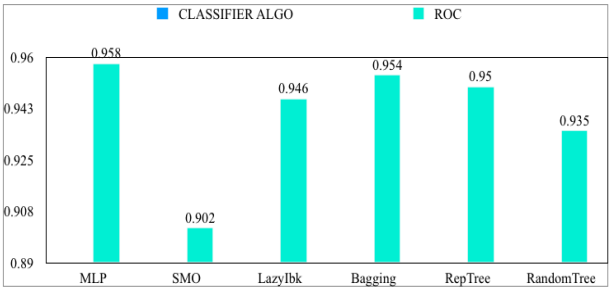
The results of the classifier algorithms are analysed on the dataset by evaluating its performance. The Performance of the Classifier are evaluated on the basis of following parameters -

1. True Positive Rate - True Positive Rate is computed by the number of correct positive predictions divided by the total number of positives. The best value is 1.0 while the worst value is 0.0.

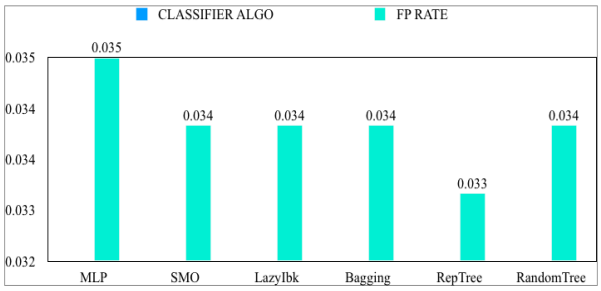


3. Comparison of True Positive Rate

2. False Positive Rate - False positive rate is computed by the number of incorrect positive predictions divided by the total number of negatives. The best value is 0.0 while the worst value is 1.0.



4. Comparison of False Positive Rate



5. Comparison of ROC

3. ROC Area - Accuracy is evaluated by the area under the ROC (Receiver Operating Characteristics) curve. An area of 1 represents a perfect result while less than that is not considered as good result.

VI. CONCLUSION

The proposed research work has implemented the Intelligent machine learning classification algorithms on UNSW-NB 15 data set after transforming the dataset by using Correlation feature selection algorithm. The results indicate that the classifier Multilayer Perceptron is giving best True positive rate and ROC Area, whereas Bagging and RepTree are giving remarkable results on the reduced dataset. The accuracy is decided on the basis of performance parameters of the classifier. The work can be utilised for obtaining more clear results on the real time dataset regarding malicious attacks on the nodes of the wireless network.

REFERENCES

1. Jangra1,A. Goel,N. Priyanka and Bhati,K. - Security Aspects in Mobile Ad Hoc Networks (MANETs): A Big Picture, International Journal of Electronics Engineering, pp. 189- 196, 2010.
2. Bakar AA, Othman ZA, Hamdan AR, Yusof R, Ismail R: An Agent Based Rough Classifier for Data Mining. Eighth International Conference on Intelligent Systems Design and Applications , vol 1. IEEE Computer Society, Washington; 2008:145-151.
3. Zhu X, Huang Z, Zhoul H: Design of a Multi-agent Based Intelligent Intrusion Detection System. IEEE International Symposium on Pervasive Computing and Applications. IEEE, Amsterdam; 2006:290-295.
4. Bourkache G, Mezghiche M, Tamine K: A Distributed Intrusion Detection Model Based on a Society of Intelligent Mobile Agents for Ad Hoc Network. In the 2011 Sixth IEEE International Conference on Availability, Reliability and Security, Vienna, August 2011. IEEE, Amsterdam; 2011:569-572.
5. Wang Y, Behera S, Wang J, Helmer G, Honavar V, Miller L, Lutz R, Slagell M: Towards the automatic generation of mobile agents for distributed intrusion detection system. J. Syst. Softw. 2006, 1(34):1-14. Elsevier.
6. Fonk C-h, Parr GP, Morrow PJ: Security schemes for Mobile Agent based Network and System Management Framework. J. Networks Syst. Manag. Springer 2011, 19: 232-256.
7. Mell P, Marks D, McLarnon M: A Denial of service resistant intrusion detection architecture. Compute Networks J Elsevier, Amsterdam, 2000.
8. Verikas A, Bacauskiene M: Feature selection with neural networks. Pattern Recognition Letters, Elsevier 2002, 23: 1323-1335. 10.1016/S0167-8655(02)00081-8.
9. Tsang C-H, Kwong S: Multi-Agent Intrusion Detection System in Industrial Network using Ant Colony Clustering Approach and Unsupervised Feature Extraction. In the IEEE Conf. Proc. on Industrial Technology. IEEE, Amsterdam; 2005:51-56.
10. Kabir MM, Islam MM, Murase K: A New Wrapper Feature selection approach using Neural Network. Neuro Computing 2010, 73: 3273-3283. Elsevier.
11. Mar J, Yeh Y-C, Hsiao I-F: An ANFIS-IDS against Deauthentication DOS Attacks for a WLAN Taichung, 17–20 October 2010. IEEE, Amsterdam; 2010:548-553.
12. Me L: GASSATA, a genetic algorithm as an alternative tool for security audit trials analysis. In Proceedings of 1st International workshop on Recent Advances in Intrusion Detection. Belgium; 1998.
13. Goldberg DE: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Boston; 1989.
14. Stein G, Chen B, Wu AS, Hua KA: Decision tree classifier for network intrusion detection with GA-based feature selection, Proceedings of the 43rd Annual Southeast Regional Conference. Volume 2. ACM, Georgia; 2005:136-141.
15. "UNSW-NB15 dataset," Available: http://www.unsw.adfa.edu.au/australian- centre- for- cyber- security/cybersecurity/ ADFA-NB15-Datasets, 2015, retrieved December 15, 2016.
16. Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.
17. Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset." Information Security Journal: A Global Perspective (2016): 1-14.
18. Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News, Sushil kumar Kalmegh, IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015. ISSN 2348 – 7968, pp. 438-446a

AUTHORS PROFILE

Madhavi Dhingra is working as Assistant Professor in Department of Computer Science and Engineering at Amity University Madhya Pradesh, Gwalior. She has done M.Tech in Computer Science from UPTU and also qualified UGC-NET and Gate in year 2012. She is currently pursuing Ph.D. from Amity University Madhya Pradesh, Gwalior.

Dr. S C Jain is working as Director in Amity School of Engineering & Technology at Amity University Madhya Pradesh, Gwalior. He has done Ph D from Barkatullah University Bhopal-2011 on Mgmt of Tech Diffusion with special ref to National Security. He has been an An alumnus of BITS Pilani, IIT Kharagpur, College of Defence Management and has been awarded with the award of prestigious 'Vishisht Seva Medal' twice.

Dr. Rakesh Singh Jadon is Professor & Head in Department of Computer Applications, Madhav Institute of Technology andScience,Gwalior (M.P.). He has done his Ph.D. in 2002 from IIT Delhi Computer Science & Engineering. On "A Fuzzy Theoretic Approach for Characterizing Video Sequences".