Enhanced Data Mining Technique for Intrusion Detection using Data Mining Tool

Rohit Sahu*, Yukti Kesharwani#
*M.Tech Scholar,
Department of Computer Science and Engineering
#Assistant Professor,
Department of Computer Science and Engineering
Dr. C. V. Raman University

Abstract— As we know that security is one of the major concern because nowadays computer attack has become very common. Although there are many existing mechanisms for Intrusion detection, but the main issues is the security and accuracy of the system. In this paper we have used J48 decision tree data mining techniques as an intrusion detection mechanism using KDD dataset and ORNL (Oak Ridge National Laboratories) data set for experimental work. The J48 algorithm gives better result with higher accuracy which is compared with previous work which was performed on ORNL data set.

***Index Term* —Decision tree, Intrusion Detection and J48 algorithm.**

I. INTRODUCTION

Nowadays, many organizations and companies use Internet services as their communication and marketplace to do business such as at EBay and Amazon.com website. Together with the growth of computer network activities, the growing rate of network attacks has been advancing, impacting to the availability, confidentiality, and integrity of critical information data. Therefore a network system must use one or more security tools such as firewall, antivirus, IDS and Honey Pot to prevent important data from criminal enterprises. Due to increased number of internet users there is a problem due to intrusion which may damage data and information stored in computer server or data base server. So we need a filter which is able to filter malicious data and normal data.

Intrusion detection is the process of monitoring and analysing the events occurring in a computer system in order to detect signs of security problems. The intrusion detection and other security technologies such as cryptography, authentication and firewalls has gained in importance in last few years [3].

As network based computer systems play increasingly vital roles in modern society, they have become intrusion detection systems provide following three essential security functions- Data confidentiality, Data integrity, Data availability [12].

Confidentiality (or secrecy) means that information is disclosed only according to policy, integrity means that information is not destroyed or corrupted and that the system performs correctly, availability means that system services are available when they are needed [13].

There are two types of intrusion detection techniques: Misuse and Anomaly. Misuse detectors analyse system activity, looking for events or sets of events that match a predefined pattern of events that describe a known attack. As the patterns corresponding to known attacks are called signatures, misuse detection is sometimes called “signature-based detection.” Anomaly detectors identify abnormal unusual behaviour (anomalies) on a host or network. They function on the assumption that attacks are different from “normal” (legitimate) activity and can therefore be detected by systems that identify these differences [4]. Anomaly detection system monitors the behaviour of a system and flag significant deviations from the normal activity as an anomaly. Anomaly detection is used for identifying attacks in a computer networks, malicious activities in a computer systems, misuses in a Web-based systems [14].

In this paper we have used data mining approach to intrusion detection. This paper mainly focuses on the signature based intrusion detection systems and presents a way to identify patterns of harmful attacks by training the system on a database and testing the same. In order to support the training and testing the ORNL dataset is used, which consists of different types of network connections labelled with the category. A model with high accuracy will be tried to develop .Model will be trained and tested on the normal and known attacks. The rest of the papers consist following sections as followed. Section 2presents a review of related work. Section 3 deals our proposed work. Section 4 introduces the basic concept of methodology we used .Section 5 describes result. And last section concludes the paper.

II. RELATED WORK

Intrusion detection started in 1980's and since then a number of techniques have been introduced to built intrusion detection systems [2].

Currently building an effective ID is an enormous knowledge engineering task. System builders relay on their intuition and experience to select the statistical measures for anomaly detection. Experts first analyse and categorize attack scenarios and system vulnerabilities, and hand-code the corresponding rules and patterns for misuse detection. Because of the manual and Adhoc nature of the development process, current IDSs have limited extensibility and adaptability. Many IDSs only handle one particular audit data source, and their updates are expensive and slow [5][6].

Heba Ezzat Ibrahim et al.[7] proposed a multi-Layer intrusion detection. There experimental results showed that the proposed multi-layer model using C5 decision tree achieves higher classification rate accuracy, using feature selection by Gain Ratio, and less false alarm rate than MLP and naïve Bayes. Using Gain Ratio enhances the accuracy of U2R and R2L for the three machine learning techniques (C5, MLP and Naïve Bayes) significantly. MLP has high classification rate when using the whole 41 features in Dos and Probe layers.

Anusha Jayasimhan et al.[5] This paper shows the implementation by viewing intrusion detection as a data mining problem. One of the most common data mining approaches i.e. classification via decision trees has been adopted to detect intrusion detection patterns. There is a limitation that it cannot detect unknown attacks.

K.Nageswara rao et al.[8] evaluated the influence of attribute pre-selection using Statistical techniques on real-world kddcup99 data set. Experimental result shows that accuracy of the C4.5 classifier could be improved with the robust pre-selection approach when compare to traditional feature selection techniques but the only limitation in this research paper is implementing correct attribute selection measure in C4.5 decision tree algorithm.

Ala' Yaseen et al.[9] This paper concludes many clustering techniques that were previously proposed to solve the inherent IDS problems. Where, the clustering techniques involved in three general aspects namely: data pre-processing, anomaly detection, and data projection/alarm filtering. Eventually, recommendations for future researches followed by the conclusion are depicted at the end of this paper.

Mahmood[15] The goal of this paper is to provide a survey of some works that employ data mining techniques for intrusion detection and to address some technical issues. They proposed a new a idea in this paper that will view intrusion detection from a data warehouse perspective and integrate data mining and on-line analytical processing (OLAP) for intrusion detection purposes. One of the major limitations of the systems is that they lack adaptability to changing behavior patterns. Some technical issues were discussed which are critical in developing a true adaptive, real-time intrusion detection system

III. PROPOSED WORK

There are many existing mechanisms for Intrusion detection system, but the major issue is the security and accuracy of the system. To improve the problem of accuracy and the efficiency of the system a very common classification approach i.e. decision tree is used. Proposed research work introduces a framework to develop a classifier based on data mining techniques.

In this framework KDD and ORNL dataset is given to Pre-processing stages which classify in J48 algorithm and reduce irreverent features from the data set so that data with less number of features will require to feed to the classifier and will provide efficiency to the classifier. Machine learning tools WEKA are used to analyse the performance of datasets.

IV. EXPERIMENTAL SETUP

The experimental methodology followed in this research includes ORNL dataset and classification technique i.e. J48 decision tree algorithm. WEKA is an innovatory tool in the history of the data mining and machine learning research communities. By putting efforts since 1994 this tool was developed by WEKA team. WEKA contains many inbuilt algorithms for data mining and machine learning. It is open source and freely available platform-independent software. The people who are not having much knowledge of data mining can also use this software very easily as it provide flexible facilities for scripting experiments. As new algorithms appear in research International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014 15 literature, these are updated in software. WEKA has also became one of the favourite tool for data mining research and helped to progress it by making many powerful features available to all[1]. The steps performed for data mining in WEKA are:

1. Preprocess the datasets.
    - Load data
    - Pre-process data
    - Analyze attributes.
2. Classify the datasets.

- Select Test Options e.g:

– Use Training Set
– % Split,
– Cross Validation

- Run classifiers
- View results

V. DATA DESCRIPTION

KDD_cup_1999 data set consist of 494020 instances and 42 attribute. Oak Ridge National Laboratories (ORNL) have created 3 datasets which include measurements related to electric transmission system normal, disturbance, control, cyberattack behaviors. Measurements in the dataset include synchrophasor measurements and data logs from Snort, a simulated control panel, and relays.

The training dataset consists of 4966 instances and contains 129 attributes. The attacks types are broadly categories into four groups-

TABLE I
TYPES OF ATTACK

DoS	Denial of service
R2L	Remote to Local
U2R	User to Root
Probing	Surveillance, Port Scans, etc.

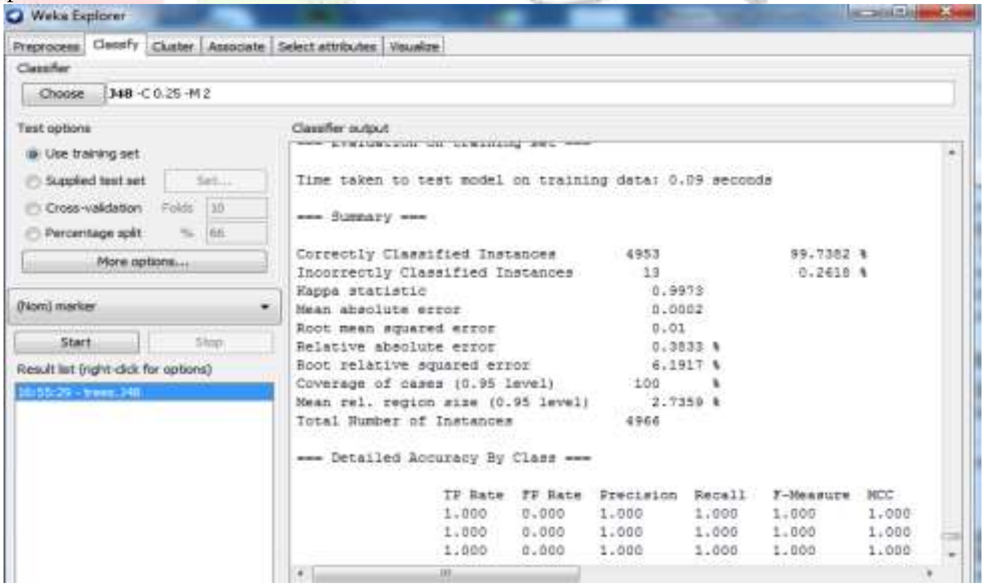
VI. J48 DECISION TREE

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found this algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable [1]

J48 is an extension of ID3. The additional features of J48 area counting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is generalization of a decision tree until it gains equilibrium off flexibility and accuracy.

VII. RESULTS

For training the system a part of the multiclass of ORNL DATASET is considered which consists of 4966 instances and contains 129 attributes and KDD_cup_1999 data set consist of 494020 instances and 42 attribute were trained and tested



J48 -C 0.25 -M 2

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) marker

10:55:29 - trees.J48

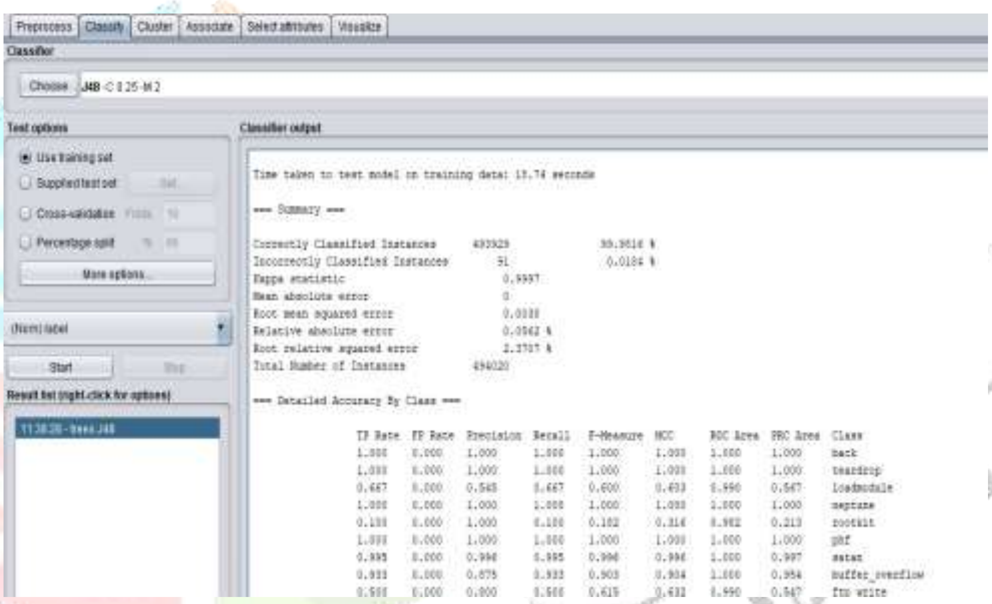
```
=== Evaluation on training set ===

Time taken to test model on training data: 0.09 seconds

=== Summary ===

Correctly Classified Instances        4953               99.7382 %
Incorrectly Classified Instances        13                0.2618 %
Kappa statistic                          0.9973
Mean absolute error                      0.0002
Root mean squared error                  0.01
Relative absolute error                  0.3833 %
Root relative squared error              6.1917 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)       2.7359 %
Total Number of Instances             4966

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC
                 1.000    0.000    1.000      1.000    1.000      1.000
                 1.000    0.000    1.000      1.000    1.000      1.000
                 1.000    0.000    1.000      1.000    1.000      1.000
```

Figure 1: Shows result of dataset after training using ORNL Dataset

Once the system has been trained, it can be tested for its performance. The data sets include whole training set itself, 10 cross validation is applied on the training set, splitting the training dataset and providing a completely different test dataset. Based on the records of the different datasets results are obtained separately for the system as shown in the Table.

TABLE 2 TRAINING THE SYSTEM BY CROSS VALIDATION ON DATASETS
This table shows highest accuracy with k=10

Datasets used for Testing	Correctly classified instances	Incorrectly classified instances	TP Rate	FP Rate	Precision	Recall	F-measure	ROC
7	92.509	7.409	0.075	0.056	0.313	0.075	0.071	0.579
9	92.586	7.4041	0.074	0.056	0.283	0.074	0.038	0.514
10	97.338	3.261	0.997	0	0.997	0	0	1



J48 -C 0.25 -M 2

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) label

```
Time taken to test model on training data: 13.74 seconds

=== Summary ===

Correctly Classified Instances      493929              99.9816 %
Incorrectly Classified Instances        91               0.0184 %
Kappa statistic                          0.9997
Mean absolute error                      0
Root mean squared error                  0.0038
Relative absolute error                  0.0562 %
Root relative squared error              2.3707 %
Total Number of Instances           494020

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     back
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     teardrop
                 0.667    0.000    0.545      0.667    0.600      0.603    0.990     0.567     loadmodule
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     neptune
                 0.100    0.000    1.000      0.100    0.182      0.316    0.982     0.213     rootkit
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     phf
                 0.995    0.000    0.998      0.995    0.996      0.996    1.000     0.997     satan
                 0.933    0.000    0.875      0.933    0.903      0.904    1.000     0.954     buffer_overflow
                 0.500    0.000    0.800      0.500    0.615      0.632    0.990     0.547     ftp_write
```

Figure 2: Shows result of dataset after training using KDD_cup_1999 Dataset

TABLE 3 TESTING THE SYSTEM BY SPLITTING DATASETS ON DIFFERENT PERCENTAGE

Percentage split on training datasets	Correctly classified instances	Incorrectly classified instances	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
50%	99.4443%	0.5557%	0.994	0.006	0.994	0.994	0.997	0.997
60% 66%	99.4443% 99.533%	0.5557% 0.467%	0.994 0.995	0.006 0.005	0.994 0.995	0.994 0.995	0.994 0.995	0.996 0.996
70% 80%	99.484% 99.6626%	0.516% 0.3374%	0.995 0.997	0.005 0.004	0.995 0.997	0.995 0.997	0.995 0.997	0.998 0.999

Table 2 and 3 shows the decision tree that is constructed after the system is trained. The number of leaves used to build the tree is 4848, and the size of the tree is 4877.

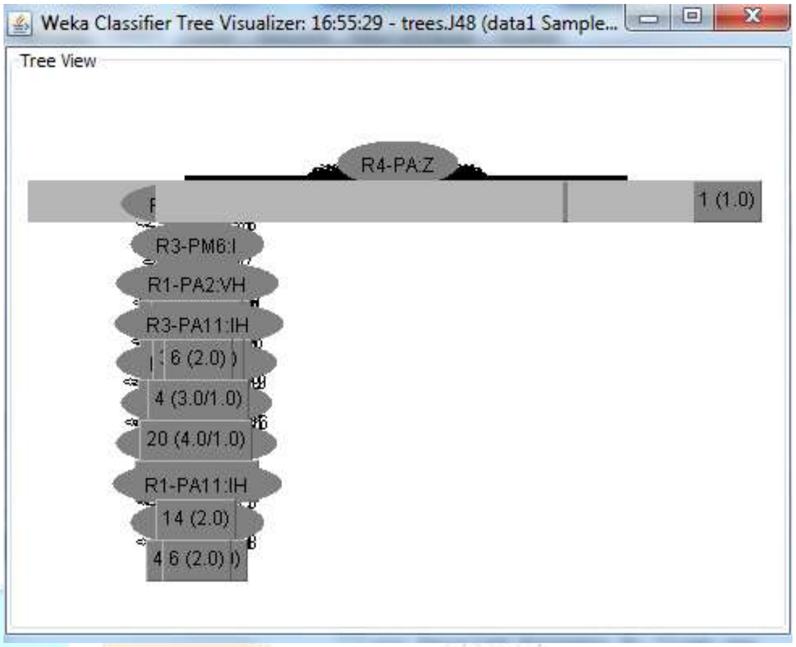


Figure 3: Visualization of Decision Tree

CONCLUSIONS

In this research we have implemented techniques for intrusion detection which gives better performance. In this research we have investigated in signature based intrusion detection which detects only known attacks. The future enhancement of this system is, it removes its drawback by implementing a system that detects both unknown and known attack. This J48 algorithm gave higher accuracy over NB and SVM. This algorithm shows 99.73% of accuracy and 99.98 for ORNL and KDD_cup_1999 dataset respectively.

ACKNOWLEDGMENT

The author would like to thank Mrs. Yukti Kesharwani, Assistant Professor, Department of Computer Science and Engineering for valuable suggestions. Appreciation also goes to the Department of Computer Science and the Dr. C.V. Raman University of Science and Technology for the support.

REFERENCES

[1]Gaganjot Kaur,Amit Chhabra,"Improved J48 Classification Algorithm for the Prediction of Diabetes",International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014.

[2]Yogendra kumar jain and Upendra,"An efficient Intrusion Detection Based on Decesion Tree Classsifier using Feature Reduction. International Journal of Scientific and Research Publication, Volume 2, Issue1, January 2012

[3]E.Kesavulu Reddy, Member IAENG, V.Naveen Reddy, P.Govinda Rajulu,"A Study of Intrusion Detection in Data Mining", Proceedings of the World Congress on Engineering 2011 Vol III WCE 2011, July 6 - 8, 2011, London, U.K.

[4]Rebecca Bace and Peter Mell,"Intrusion Detection Systems", NIST Special Publication on Intrusion Detection Systems.

[5]Anusha Jayasimhan,Jayant Gadge," Identifying Intrusion Patterns using a Decision Tree", International Journal of Computer Applications (0975 – 8887) Volume 45– No.9, May 2012.

[6] Lee,Salvatore J. Stolfo," A framework for constructing features and models for intrusion detection systems," ACM Transactions on Information and System Security, Vol. 3, No. 4, November 2000, Pages 227–261.

[7]Heba Ezzat Ibrahim,Sherif M. Badr, Mohamed A. Shaheen," Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems", International Journal of Computer Applications (0975 – 8887),Volume 56– No.7, October 2012.

[8] K.Nageswara rao, D.RajyaLakshmi, T.Venkateswara Rao," Robust Statistical Outlier based Feature Selection Technique for Network Intrusion Detection", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[9] Ala' Yaseen Ibrahim Shakhatreh , Kamalrulnizam Abu Bakar ,"A Review of Clustering Techniques Based on Machine learning Approach in Intrusion Detection Systems", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3, September 2011.

[10] "NSL-KDD data set for network-based intrusion detection systems ", Available on: http://nsl.cs.unb.ca/NSL-KDD.

[11]AdetunmbiA.Olusola,Adeola S.Oladele.,Daramola O.Abosede,"Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I WCECS 2010, October 20-22, 2010, San Francisco, USA.

[10] Reema Patel, Amit Thakkar, Amit Ganatra,"A Survey and Comparative Analysis of Data Mining Techniques for Network Intrusion Detection Systems", International Journal of Soft Computing and Engineering (IJSCE)ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[13] Sandhya Peddabachigari, Ajith Abraham*, Johnson Thomas,"Intrusion Detection Systems Using Decision Trees and Support Vector Machines".

[14] M. Sathya Narayana1, B. V. V. S. Prasad2, A. Srividhya3 and K. Pandu Ranga Reddy", Data Mining Machine Learning Techniques – A Study on Abnormal Anomaly Detection System",International Journal of Computer Science and Telecommunications [Volume 2, Issue 6, September 2011].

[15] Mahmood Hossain," Data Mining Approaches For Intrusion Detection: Issues And Research Directions",Department of Computer Science, Mississippi State University, MS 39762, USA.

[16] S Pandey, R Miri, S R Tandan "Diagnosis and Classification of Hypothyroid Disease Using Data Mining Techniques" International Journal of Engineering Research and Technology, (2013)

[17]P Gupta, S R Tandan, R Miri "Decision Tree Applied For Detecting Intrusion "International Journal of Engineering Research and Technology (IJERT) (2013)

[18]Khushboo Sharma, S R Tandan "An Optimized Parallel Confidence Measures Algorithm on Web Log Data" International Journal of Engineering Research and Technology (IJERT) (2013)

[19]Asha Miri, S.R.Tandan, Rohit Miri "Pseudo Code to Eliminate Unwanted Data Sets for Fuzzy Mining Association Rule" International Journal for Research in Applied Science & Engineering Technology (IJRASET) (2015).

[20]Rohit Miri, Priyanka Tripathi, S R Tandan" Exploration of Novel Algorithm for Reduced Computational Time by Using Fuzzy Classification Technique in Data Mining" International Journal for Research in Applied Science & Engineering Technology (IJRASET) (2015)

[21]Rohit Miri, Priyanka Tripathi, S R Tandan "Novel Algorithm For Finding The Range Of Fuzzy Values For Quantitative Data Sets By Data Mining And Fuzzy Technique" Journal Of Advanced Database Management & Systems Journal Of Advanced Database Management & Systems, (2015)

[22]S R Tandan, Rohit Miri, ,Dr Priyanka Tripathi "A Bird Eye Review on Reduced Time Complexity by Using Data Mining and Fuzzy Techniques" international journal for research in applied science and engineering technology (ijraset), (2014)

[23]S R Tandan Rohit Miri, Priyanka Tripathi "TRApriori Classification Based Algorithm by Fuzzy Techniques to Reduced Time Complexity" International Journal of Computer Science & Information Technology, International Journal of Computer Science & Information Technology, (2014)

[24]Rohit Miri, Priyanka Tripathi, Keshri Verma, S.R. Tandan "Novel Algorithm to Reduced Computational Data Sets for Fuzzy Association Rule" International Journal For Research In Applied Science And Engineering Technology (IJRASET), (2015).

[24]R Miri, P Tripathi, S R Tandan "Novel Algorithm for Reduced Computational Data by Using Fuzzy Classification and Data Mining Techniques" Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies" ACM, (2014)