INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)

ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online)
Volume 5, Issue 10, October (2014), pp. 21-31
© IAEME: www.iaeme.com/IJCET.asp
Journal Impact Factor (2014): 8.5328 (Calculated by GISI)
www.jifactor.com

IJCET

© I A E M E

DATA REDUCTION TECHNIQUES TO ANALYZE NSL-KDD DATASET

Shailesh Singh Panwar, Dr. Y. P. Raiwani

Department of Computer Science and Engineering, H N B Garhwal University, Srinagar, Uttarakhand - 246 174 India

ABSTRACT

In information field we have huge amount of data available that need to be turned into useful information. So we used Data reduction and its techniques. A process in which amount of data is minimized and that minimized data are stored in a data storage environment is known as data reduction. By this process of reducing data various advantages have been achieved in computer networks such as increasing storage efficiency and reduced computational costs. In this paper we have applied data reduction algorithms on NSL-KDD dataset. The output of each data reduction algorithm is given as an input to two classification algorithms i.e. J48 and Naïve Bayes. Our main is to find out which data reduction technique proves to be useful in enhancing the performance of the classification algorithm. Results are compared on the bases of accuracy, specificity and sensitivity.

Keywords: Data Reduction; Data Reduction Techniques; Classification; WEKA; NSL- KDD.

I. INTRODUCTION

With the tremendous growth of computer networks mostly computer system suffers from security vulnerabilities which are difficult to handle technically as well as economically by users [1]. Data reduction is a process in which amount of data is minimized and that data is stored in a data storage environment. By data reduction algorithms reduce massive data-set to a manageable size without significant loss of information represented by the original data. This process various advantages have achieved in computer networks such as increasing storage efficiency and reduce costs. There are two important motivating factors of data reduction, first is redundancy and second is reduction of complexity regarding live network acquisition [2].

Data reduction technique can be applied to obtain a reduce representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on

reduced data set should be more efficient yet produce the same (or almost the same) analytical results. For very large data sets, there is an increased likelihood that intermediate, additional steps, data reduction, should be performed prior to applying the data reduction techniques. Three basic operations in data reduction process are: delete a column, delete a row, and reduce the number of column.

The larger data sets have more useful information but it decreasing storage efficiency and increasing costs but larger data set have more useful information. Which techniques, methods or relative terms are used in larger data set may also used in smaller data set. So the benefit of data reduction techniques we propose increase as the data sets themselves increase size, complexity and reduce the costs. We have taken a broad view of large qualitative data sets, aiming to highlight trends, relationships, or associations for further analysis, without loss of any information [3].

The recent advantage data collection and storage capabilities have include information overhead in many applications sciences, e.g., on-line monitoring of spacecraft operations with time series data. In this we perform data reduction technique before storage or transmission of data. This can be some information loss, but not all features of data might be relevant. The motivating factor of this is that real system, which data we get after data reduction, that dimensionality is lower than the space, which is measure in. Reconstruct the lower dimensional samples which are needed and possible with varying degrees of accuracy [4].

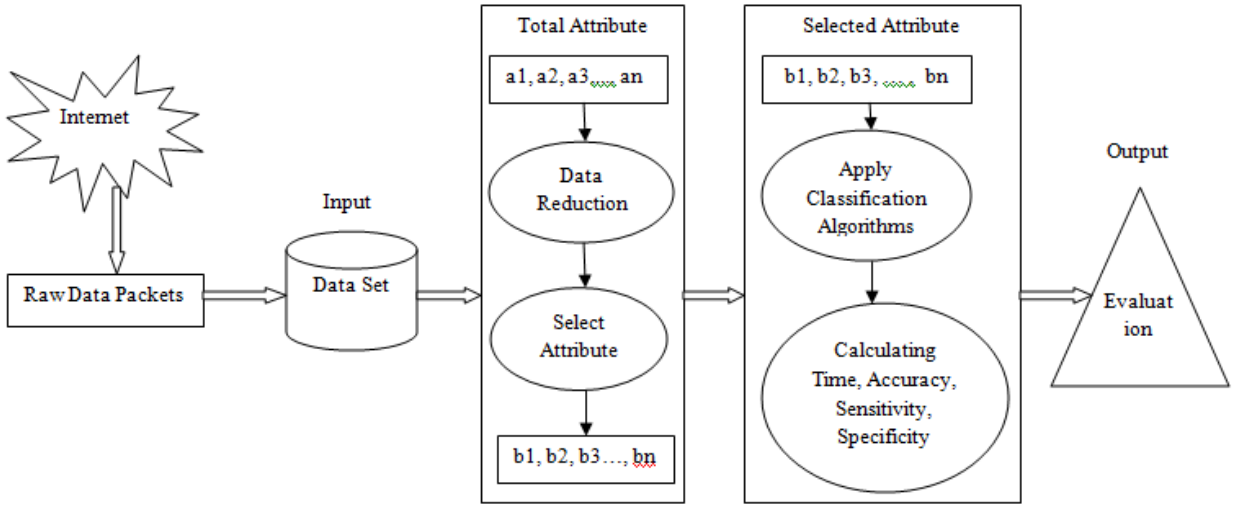

Fig I: Block diagram of data reduction [25]

The rest of this paper is organized as follows: Section II includes the literature review about different kinds of work done by the various authors related to data reduction and its techniques and clustering. In section III a brief introduction to data reduction and data reduction algorithms, which are used. Section IV is about the classification and its algorithms which we are used.The experiments and results are discussed in section V. In section VI we have discussed conclusion and future work in research field.

II. RELATED WORK

In network huge amount of data should be processed, and the data contains redundant and noisy features causing slow training and testing process, high resource consumption as well as poor detection rate. For this we reduce the data.

Tao et al [1] proposed the Manifold learning algorithm Based Network Forensic System. Manifold learning is a recent approach to nonlinear dimensionality reduction. The idea behind manifold learning is that dimensionality of many data sets is only artificially high. Each manifold learning algorithm used a different geometrical property of the underlying manifold. They are also called spectral methods. They are also called spectral methods, since the low dimensional embedding task is reduced to solving a sparse eigen value problem under the unit covariance constraint. However, due to this imposed constraint, the aspect ratio is lost and the global shape of the embedding data can not reflect the underlying manifold. For data reduction used following steps in this: - A Locally Linear Embedding Algorithm. B. Data Processing of the Data Set. C. Data Reduction with LLE.

Willetty et al [4] proposed Data Reduction Techniques. They have applied several data reduction technique in available datasets with the goal of comparing how an increasing level of compression affects the performance of SVM-type classifiers. Several data reduction techniques are applied to three datasets (WDBC, Ionosphere and PHM). The comparison of these techniques was based on how well the data can be classified by an SVM or PSVM (linear and nonlinear versions for each) at decreasing number of components retained. One dataset proved to be hard to classify, even in the case of no dimensionality reduction. Also in this most challenging dataset, performing PCA was considered to some advantages over the other compression techniques. Based on our assessment, data reduction appears a useful tool that can provide a significant reduction in signal processing load with acceptable loss in performance.

Fodo et al [9] proposed a survey of dimension reduction techniques. During the last ten year data collection and storage capabilities have include information overhead in sciences. Increase the number of observations traditional statistical methods break are down partly. The dimension of the data is that measured on each observation. High-dimensional datasets present many mathematical challenges as well as some opportunities and rise new theoretical developments. In high-dimensional datasets, for understanding the underlying phenomena of interest all measure variables are not important. Some computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data. Various applications are used to reduce the dimension of original data without loss of any information.

Robert et al [10] proposed Novel Data Reduction Technique. Large-scale networks generate enormous numbers of events that determine which are malicious attacks and which are not. In network analyst's first most severe attack are resolved in order to limit the potential for damage to the network as much as possible. There are many data reduction and event correlation technique for reducing the amount of data needing analysis; these techniques do not provide prioritization capabilities. In this, identifying and resolving the most critical events first. Impact assessment technique identifies the potential impact. Impact assessment improves the efficiency of the analysis process and reduces the amount of data needing to be transmitted over the network.

Furtado et al [11] proposed Analysis of Accuracy of Data Reduction Techniques. Data warehouse is a growing interest in the analysis of data. Data warehouse can frequently take extremely large and typical queries. Obtain the best estimates which have smaller response times and storage need by data reduction techniques. In this apply the simple data reduction technique in several data sets to analysis the accuracy. Data cube density and distribution skew are important parameters and large range queries are approximated much more accurately then point or small range queries.

Joshua et al [12] proposed a global geometric framework for nonlinear dimensionality reduction. They have used some classical technique such as principal component analysis (PCA), multidimensional scaling (MDS) to solved dimensionality reduction problems that determine local information to learn with the basis of global geometric of a

data set. We analysis nonlinear degree of freedom that is lies in complex natural observation. Aim was to use nonlinear dimensionality reduction algorithms to determine a globally optimal solution for vital class of data main fold.

III. DATA REDUCTION AND ITS TECHNIQUES

Data reduction is the transformation of numerical or alphabetical digital information derived empirical or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts. By data reduction reduce massive data-set to a manageable size without significant loss of information represented by the original data.

The advantages of data reduction are results are shown in a compact form and easy to understand. The graphical or pictorial representations can be used. Overall patterns can be seen. In this comparisons can be made between different sets of data. The quantitative measures can be used. The disadvantages are original data are lost and the process is irreversible.

There are three data reduction strategies:-

1. Dimensionality Reduction: - Dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration and can be divided into feature selection and feature extraction i.e. Dimensionality Reduction is about converting data of very high dimensionality into data of much lower dimensionality such that each of the lower dimensions conveys much more information [5]. Feature selection (i.e., attribute subset selection) is selecting a minimum set of attributes (features) that is sufficient for the data mining task. Heuristic methods is step-wise forward selection and step-wise backward elimination. It is combining forward selection and backward elimination [6].

2. Clustering: - Clustering is the process of organizing objects into groups whose members are similar in some way. In other words Clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [7].Clustering is partition data set into clusters, and one can store cluster representation only. It can be very effective if data is clustered but not if data is “smeared”. There are many choices of clustering definitions and clustering algorithms.

3. Sampling: - The "big" data set to create a smaller one called sampling. It is used in conjunction with skewed data. Sampling obtaining a small sample to represent the whole data set. It Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data. Key principle of sampling chooses a representative subset of the data. Simple random sampling may have poor performance in the presence of skew. It develops adaptive sampling methods. Stratified sampling is approximate the percentage of each class (or subpopulation of interest) in the overall database [8].

Data reduction is the process of minimizing the amount of data that needs to be stored in a data storage environment. Data reduction can increase storage efficiency and reduce costs. We used WEKA tool for data reduction. In this, used different data reduction technique for reduce KDD 99 data set.

There are five data reduction technique as follows:-

1) Gain Ratio Attribute Evaluator: - Evaluates the worth of an attribute by measuring the gain ratio with respect to the class. Gain ratio is a modification of the information gain that reduces its bias on high-branch attributes. Gain ratio should be large when data is evenly spread and small when all data belong to one branch. It takes number and size of branches into account when choosing an attribute.

$$\text{Gain R (Class, Attribute)} = \frac{(\text{H(Class)} - \text{H(Class | Attribute)})}{\text{H(Attribute)}}$$

It corrects the information gain by taking the intrinsic information of a split into account (i.e. how much info do we need to tell which branch an instance belongs to). Importance of attribute decreases as intrinsic information gets larger [13, 14].

2) CfsSubset Attribute Evaluator: - CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficients are used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search [13].

3) Information gain Attribute Evaluator: **-** InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to the class.

$$\text{Info Gain (Class, Attribute)} = \text{H (Class)} - \text{H(Class)} - \text{H(Class|Attribute)}$$

where H is the information entropy. It is widely used standard feature selection method. Its disadvantage is that it does not take into account feature interaction. The information gain measure is used to select the test attribute at each node. The information gain measure prefers to select attributes having a large number of values [15].

4) Wrapper Subset Evaluator: - The wrapper approach depends on the classifier that should be used with the resulting attribute subset. Wrapper methods evaluate subsets by running the classifier on the training data, using only the attributes of the subset. The better the classifier performs, usually based on cross-validation, the better is the selected attribute set. One normally uses the classification-accuracy as the score for the subset. Though this technique has a long history in pattern recognition, introduced the term wrapper that is now commonly used [16].

5) OneR attribute evaluator: - Rule based algorithms provide ways to generate compact, easy-to-interpret, and accurate rules by concentrating on a specific class at a time Class for Evaluating attributes individually by using the OneR classifier. OneR is a simple and very effective data reduction algorithm which is frequently used in data mining applications. OneR is the simplest approach to finding a classification rule as it generates one level decision tree. OneR constructs rules and tests a single attribute at a time and branch for every value of that attribute. For every branch, the class with the best classification is the one occurring most often in the training data [17].

IV. CLASSIFICATION

Classification is a data mining task that maps the data into predefined groups & classes. It is also called as supervised learning. It consists of two steps:

1. Model construction: It consists of set of predetermined classes. Each tuples/sample is assumed to belong to a predefined class. The set of tuples used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formula.

2. Model usage: This model is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur.

The classification accuracy of most existing data mining algorithms needs to be improved, because it is very difficult to detect several new attacks, as the attackers are continuously changing their attack patterns. The classifier will separate the input dataset to two classes: Normal and Anomaly [18].

Classification data mining technique Classification maps a data item into one of several pre-defined categories. These algorithms normally output "classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient "normal" and "abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class. There are many types of classifiers are available like tree, bayes, function rule. Basic aim of classifier is predict the appropriate class [19]. The classification accuracy of the most existing data mining algorithms needs to be improved, because it is very difficult to detect several new attack, as the attack are continuously changing their attack patterns. Classification algorithm to analysis the accuracy, sensitivity, specificity and Kappa statistics for analyze and compare the performance. The accuracy, sensitivity and specificity were calculated by True Positive, False Positive, False Negative and True Negative.

1. Random Forest:-Random forest is a powerful approach to data exploration, data analysis and predictive modeling. It is an ensemble method which uses recursive partitioning to generate many trees and then aggregate the results. Using a bagging technique, each tree is independently constructed using a bootstrap sample of the data. Also, using the random subspace selection, randomly selected numbers of predictors are considered in the splitting criterion during optimization at each node. In addition to prediction for new data, importance measures of each variable may be calculated from the data generated during this ensemble method. This may be useful for model reduction, especially when there are a large number of predictors [20].

2. PART (Projective Adaptive Resonance Theory): PART is an Instance-based learner using an entropic distance measure [21]. The PART algorithm developed in is based on the assumptions that the model equations of PART (a large scale and singularly perturbed system of differential equations coupled with a reset mechanism) have quite regular computational performance described by the following dynamical behaviors, during each learning trial when a constant input is imposed.

V. EXPERIMENTS AND RESULTS

In order to reduce the data variety of data reduction technique mentioned above, the NSL-KDD dataset is selected as the experiment object. Now it is the most widely used dataset for the evaluation for data reduction in network forensic domain.

KDD cup-99 dataset is used for experimental evaluation, as discussed earlier. It is very hard to execute the proposed technique on the NSL KDD cup- 99 dataset for estimating the performance, because it is a large scale and in order to compare the performance of techniques considered, we used a 10% subset of original NSL KDD Cup-99 data set. Attributes in the NSL KDD Cup data sets had all forms of data like continuous, discrete, and symbolic, with significantly varying resolution and ranges. Most pattern classification methods are not able to process data in such a format. Hence pre-processing was required. Pre-processing consisted of two steps: first step involved mapping symbolic-valued attributes to numeric-valued attributes and second step implemented scaling [22]. KDD have 42 attributes. These are shown in table I.

Table I: NSL-KDD Dataset Attributes

Toatal Attribute		
Duration	su_attempted	same_srv_rate
protocol_type	num_root	diff_srv_rate
service	num_file_creation	srv_diff_host_rate
flag	num_shells	dst_host_count
src_byte	num_access_file	dst_host_srv_count
dst_byte	num_outbound cmds	dst_host_same_srv_rate
land	is_host_login	dst_host_diff_srv_rate
wrong_fragment	is_gust_login	dst_host_same_src_port_rate
urgent	count	dst_host_srv_diff_host_rate
hot	srv_count	dst_host_serror_rate
num_failed_login	serror_rate	dst_host_srv_serro_rate
logged in	srv_serror_rate	dst_host_rerror_rate
num_compromised	rerror_rate	dst_host_srv_rerror_rate
root_shell	srv_rerror_rate	class

WEKA tool is used for experiment, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. WEKA operates on the predication that the user data is available as a flat file or relation; this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values [23].

Table II consists of the selected attributes after applying data reduction algorithms over complete data set and the volume the dataset is reduced to. Each algorithm have different no of attributes based on their evaluation criteria. Now the main task is to find out which classification algorithm gives better results for data reduction over NSL-KDD dataset. For this purpose we have implemented two well known classification algorithms PART and Random Forest and tried to find out over which data reduction algorithm output these two algorithms gives better results in terms of accuracy, sensitivity and specificity.

Fig II contains the output of data reduction algorithms. Number of selected attributes is then given as an input to classification algorithm.

Table II: Attributes after applying different data reduction algorithms

	Volume(Mb)	No. of Selected Attributes	No. of Unpotential Attributes
KDD FULL DATASET	20.4	42	0
Cfs Subset Eval	10.8	6	36
Gain Ratio Attribute Eval	10	17	25
Info Gain Attribute Eval	11.8	19	23
OneR Attribute Eval	14.7	28	14
Wrapper Subset Eval	16.9	36	6
Symmetrical uncertAttribute Eval	10.6	19	23

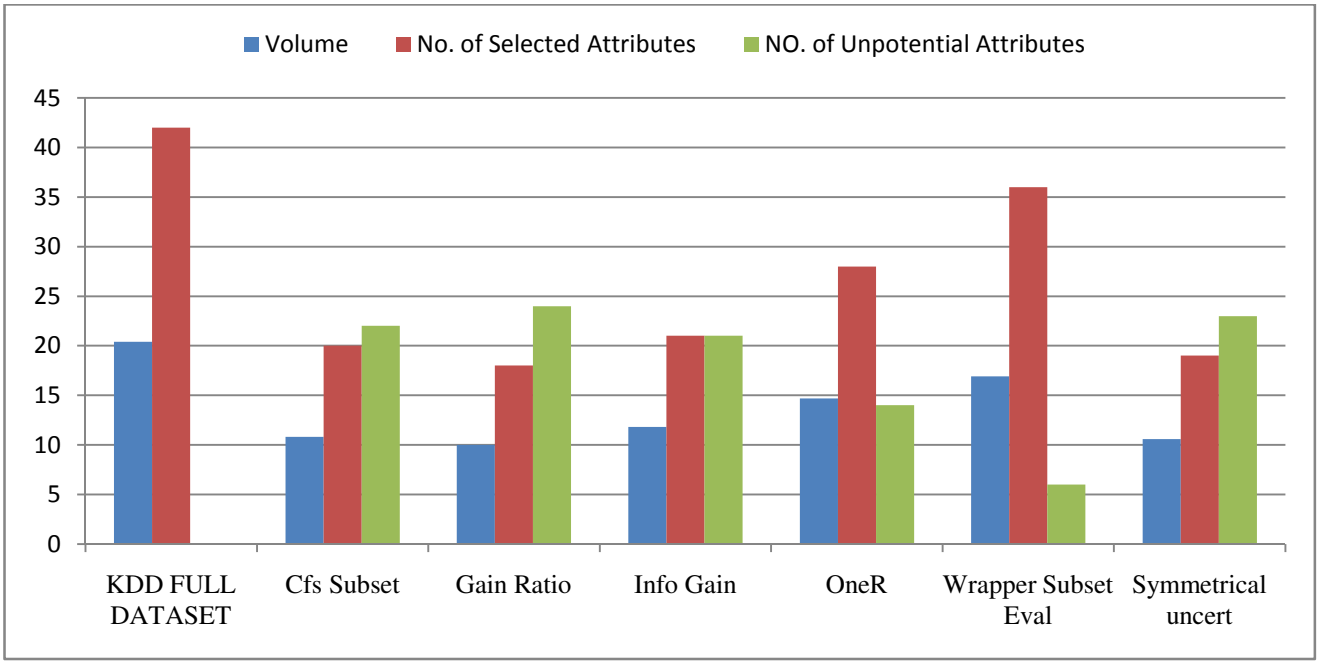

Fig II: No. of Selected Attributes after applying data reduction algorithm

Table III: Result for different classification algorithms

Data Reduction Algorithms	Total Attribute	Selected Attribute	Classification Algorithms	Time(sec)	Accuracy(%)	Sensitivity(%)	Specificity(%)
Cfs Subset Eval	42	6	j48	10.56	98.40	97.68	99.50
			NaïveBayes	1.39	69.79	51.67	90.00
Gain Ratio Attribute Eval	42	17	j48	57.09	98.76	98.72	98.77
			NaïveBayes	4.88	81.76	73.20	94.81
Info Gain Attribute Eval	42	19	j48	56.44	98.87	98.78	99.01
			NaïveBayes	6.3	87.49	82.45	95.19
OneR Attribute Eval	42	28	j48	49.84	99.68	99.73	99.67
			NaïveBayes	8.28	90.70	94.25	99.67
Wrapper Subset Eval	42	36	j48	19.84	99.88	99.83	99.97
			NaïveBayes	10.05	90.51	90.33	90.72
Symmetrical uncertAttribute Eval	42	19	j48	57.39	98.87	98.78	99.01
			NaïveBayes	6.47	87.49	82.45	95.18

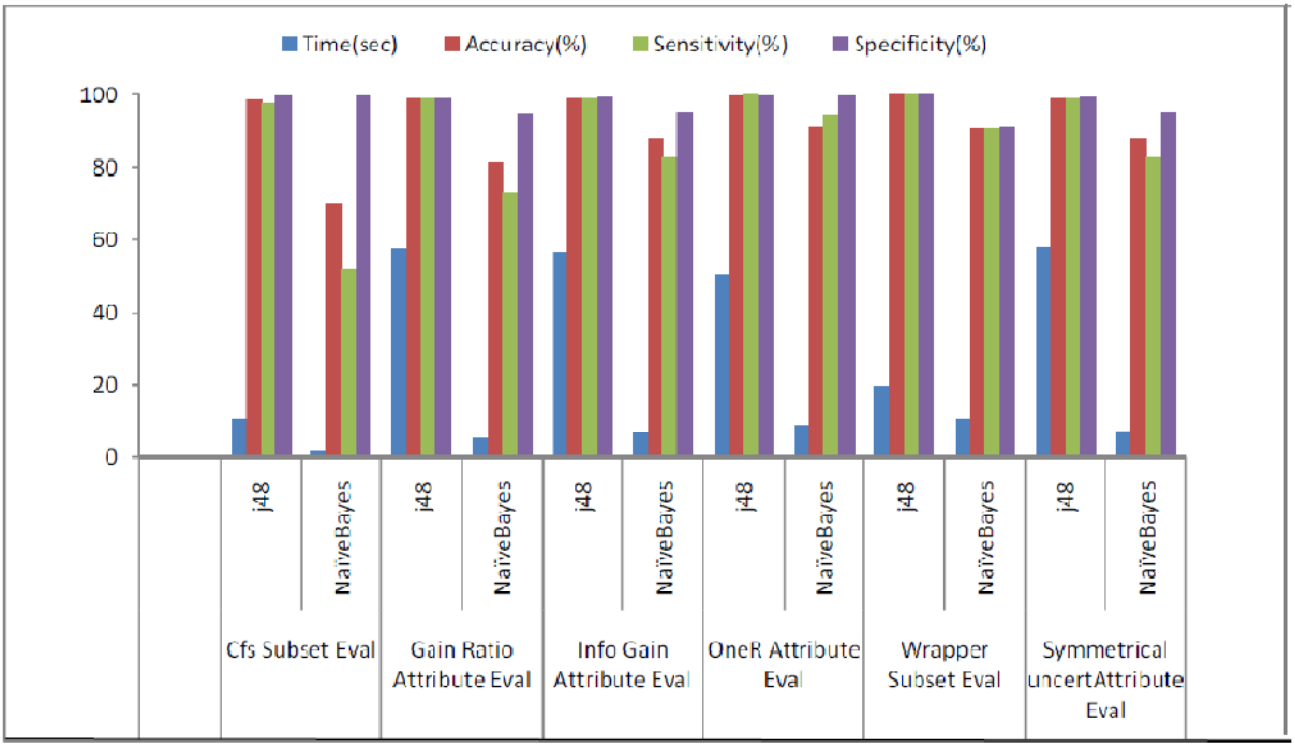

Fig III: Analyzing best data reduction algorithm

1. Results:- We have applied data reduction algorithms on NSL-KDD dataset. The output of each data reduction algorithm is given as an input to two classification algorithms i.e. J48 and Naïve Bayes. Results are shown in Table III. Fig III represents the output of both the classifiers on the reduced dataset. We have computed the performance on the basis of three parameters i.e. accuracy, sensitivity and specificity.

2. Analysis:- Result shows that the OneR Attribute evaluation data reduction technique has outperformed in experiment. Between the two classifiers we have selected the combination of Wrapper Subset evaluation and Random Forest classifier gives the best result with 99.88% accuracy, 99.83% specificity 99.97% sensitivity and 20.86 seconds of computation time. And the Naïve Bayes evaluation and Cfs Subset Classifiers gives the worse result with 69.79% accuracy, 51.67% specificity 90.00% sensitivity and 1.39 seconds of computation time.

VI. CONCLUSION

Data reduction algorithms reduce massive data-set to a manageable size without significant loss of information represented by the original data. The attribute selection methods of data reduction techniques help to identify some of the important attributes, thus reducing the memory requirement as well as increase the speed of execution. The purpose of this experimental work is to find out which data reduction algorithm gives better results. The work is implemented in two phases. First phase is to reduce the dataset and select potential attributes that will be given as an input to second phase. In second phase we have implemented two classification algorithms, PART and Random Forest, and find out the accuracy, sensitivity and specificity on the reduced dataset. Then on the basis of output from classification algorithms comparison is done to find out which data reduction algorithm has outperformed in the experiment. The study on the NSL-KDD dataset shows that OneR attribute evaluation proves to be the best among all the data reduction techniques.

In future we can perform this work on large size dataset and can compare large no of data reduction techniques. We can also combine the data reduction with clustering to enhance the efficiency of the algorithms.

REFERENCES

[1] P. Tao, C. Xiaosu, L. Huiyu and C. Kai, "Data Reduction for Network Forensics Using Manifold Learning", *Sch. of Computer Sci. & Technol., Huazhong Univ. of Sci. & Technology*. Wuhan, China, pp. 1-5, 2010.

[2] M. Rouse, "Data Reduction", Last Updated on August 10, [Available Online] http://searchdatabackup.techtarget.com/definition/data-reduction.

[3] E. Namey, G. guest, L. Thairu, L. Johnson, "Data Reduction Techniques for Large Qualitative Data Sets", 2007, pp137-162 [Available Online] http://www.stanford.edu/~thairu/07_184. Guest.1sts .pdf.

[4] R. Georgescu, C. R. Berger, P. Willett, M. Azam_, and S. Ghoshal, "Comparison of Data Reduction Techniques Based on the Performance of SVM-type Classifiers". *Dept. of Electr. and Comp. Engineering, University of Connecticut, Storrs, CT 06269, Qualtech Systems Inc., Wetherseld, USA*, 2010.

[5] A. Ghodsi, "Dimensionality Reduction", Technical Report, 2006-14, Department of Statistics and Actuarial Science, University of Waterloo, pp. 5-6, 2006.

[6] Ricardo Gutierrez, "Dimensionality reduction", *Lecture Notes on Intelligent Sensor Systems,* Wright State University [Available Online] http://research.cs.tamu.edu/prism/lectures/is s/ iss_ l10.pdf .

[7] Cluster Analysis, [Available Online], http://en.wikipedia.org/wiki/Cluster_analysis.

[8] Data Preprocessing, [Available Online], www.cs.uiuc.edu/homes/hanj/cs412/bk3/03 Preprocessing.ppt .

[9] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction" *Science,* 290(5500), 2000.

[10] Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram, "Comparative Study of Attribute Selection Using Ratio and Correlation Based Feature Selection", *International Journal of Information Technology and Knowledge Management,* vol. 2, no. 2, pp. 271-277, July-December 2010.

[11] Decision Trees an Introduction, [Available Online], www.liacs.nl/~knobbe/intro _ dec_ tree.ppt .

[12] J. Novakovic, "Using Information Gain Attribute Evaluation to Classify Sonar Targets", *17th Telecommunications forum (Telfor),* Serbia, Belgrade, pp. 1351-1354, 2009.

[13] S. B. Aher, Mr. LOBO, "Data Mining in Educational System using WEKA". *International Conference on Emerging Technology Trends (ICETT'11), pp*.20-25, 2011.

[14] I. K. Fodor, "A survey of dimension reduction techniques", Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, 2002.

[15] E. Robert, B. Anupama, and C. Trent, "A novel data reduction technique", *Seventh Annual Workshop on Cyber Security and Information Intelligence Research,* Oak Ridge, Tennessee, ACM, 2011.

[16] P. Furtado and H. Madeira, "Analysis of Accuracy of Data Reduction Techniques", *University of Coimbra*. Portugal, 1999.

[17] S. H. Vege, "Ensemble of Feature Selection Techniques for High Dimensional Data", *Masters Theses & Specialist Projects,* [Available Online] http://digitalcommons.wku.edu/theses/1164

[18] *B. Neethu,* "Classification of Intrusion Detection Dataset using machine learning Approaches". *International Journal of Electronics and Computer Science Engineering,* pp. 1044-1051, 2012.

[19] N. S. Chandolikar and V. D. Nanadavdekar, "Comparative Analysis of Two Algorithms for Intrusion Attack Classification Using KDD CUP Dataset", *International Journal of Computer Science and Engineering (IJCSE)*, pp. 81-88, Aug 2012.

[20] I. Maglogiannis, K. Karpouzis, M. Bramer, and S. Kotsiantis, "Local Ordinal Classification," in *Artificial Intelligence Applications and Innovations*. vol. 204: Springer US, pp. 1-8, 2006.

[21] G. Kalyani, A. J. Lakshmi, "'Performance Assessment of Different Classification Techniques for Intrusion Detection", *Journal of Computer Engineering*, vol. 7, no. 5, pp 25-29, Nov-Dec. 2012

[22] NSL-KDD dataset, [Available Online] http://iscx.ca/NSL-KDD/

[23] Weka User Manual, [Available Online],www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf

[24] S.S.Panwar, "Performance analysis of Data Reduction Techniques", *2nd International*

[25] *Conference on Recent Trends in Computing*", pp 301-307, 4-5 oct 2013.

[26] Dr. Naveeta Mehta and Shilpa Dang, "Identification of Important Stock Investment Attributes using Data Reduction Technique", International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 2, 2012, pp. 188 - 195, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

[27] R. Vijaya Prakash, Dr. A. Govardhan and Prof. S.S.V.N. Sarma, "Mining Non-Redundant Frequent Patterns in Multi-Level Datasets using Min Max Approximate Rules", International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 2, 2012, pp. 271 - 279, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.

[28] Nitin Mohan Sharma and Kunwar Pal, "Implementation of Decision Tree Algorithm After Clustering Through Weka", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 1, 2013, pp. 358 - 363, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.