ISSN: 2250-3676
Volume-4, Issue-2, 214-219

PERFORMANCE ANALYSIS OF DATA REDUCTION ALGORITHMS USING ATTRIBUTE SELECTION IN NSL-KDD DATASET

Meenu Choudhary[1], Prity[2], Vikas Choudhary[3]

[1]*Student, Computer Science & Engineering, Ajmer Institute of Technology, Rajasthan, India,* ***meenu.choudhary0@gmail.com***
[2]*Student, Computer Science & Engineering, Ajmer Institute of Technology, Rajasthan, India,* ***prity.only1@gmail.com***
[3]*Assistant Professor, Computer Science & Engineering, Ajmer Institute of Technology, Rajasthan, India,* ***vikasmca51@gmail.com***

Abstract

Abstract—Data sets – both structured & unstructured, which are so large and complex that processing it using database management tools or traditional applications to derive analytical insights becomes difficult. Capturing, storage, internal search, sharing, predictive analysis and visualization are some of the major tasks being performed on such data sets. With the increasing amount of data generated by social sharing platforms & apps, the process of data reduction has become inevitable. It involves compressing the data being generated & storing it in a data storage environment. In computer networks, these techniques have played a pivot role in increasing storage efficiency and reduced computational costs. In this paper, data reduction algorithms have been applied on NSL-KDD dataset. The output of data reduction algorithm is given as an input to two classification algorithms i.e. PART and Random Forest. The aim is to find out which data reduction technique proves to be useful in enhancing the performance of the classification algorithm. Performance is compared on factors like precision, sensitivity and accuracy.

Index Terms: *Data Reduction, NSL-KDD, Cfs Subset Eval, Gain Ratio Attribute Eva, Info Gain Attribute Eval, OneR Attribute Eval, Wrapper Subset Eval Classification etc.*

---------- *** ----------

1. INTRODUCTION

As the Web rapidly evolves, people are becoming increasingly enthusiastic about interacting, sharing and collaborating through social networks, mobile apps & collaborative media. Facebook, LinkedIn, Twitter have become an internet phenomenon causing the size of the social Web to expand exponentially. The distillation of knowledge from such a large amount of unstructured information, however, is an extremely difficult task, as the contents of today's Web are perfectly suitable for human consumption, but remain hardly accessible to machines. Big social data analysis grows out of this need and it includes disciplines such as social network analysis, multimedia management, social media analytics, trend discovery, and opinion mining. Data reduction algorithms reduce massive data-sets to a manageable size without significant loss of information. The main motivating factors for data reduction are removing redundancy and reducing complexity with respect to live network acquisition [2].

Threebasicoperationsindatareduction processare:deletea column,deletearow,andreducethe numberofcolumn. The benefitof datareduction techniquesIproposeincreasesasthedatasetsthemselves become increasingly large and complexand helps reducecosts.Ihavetaken abroadview oflargequalitative datasets,aiming tohighlight trends,relationshipsorassociations forfurtheranalysis, withoutlossofanyinformation[3].

Therecentadvances indatacollection andstoragecapabilities have proven to be very effective in conquering storage scalability challenges. For ex. the warehouse at Facebook stores upwards of 300 PB of Hive data, with an incoming daily rate of about 600 TB. Facebook uses Hive, a query engine based on Corona Map-Reduce used for processing and creating large tables. Data that is loaded into tables in the warehouse is primarily stored using Record-Columnar File Format (RCFile). RCFile is a hybrid columnar storage format that is designed to provide the compression efficiency of a columnar storage format, while still allowing for row-based query processing. The core idea is to partition Hive table data first horizontally into row groups, then vertically by columns so that the columns are written out one after the other as contiguous chunks. In addition,The HortonWorks engineering team designed and implemented ORCFile's on-disk representation and reader/writer &used different codecs and compression levels for different columns, storing additional statistics and exposing these to query engines.

2. DATA REDUCTION

Datareductionisthe reductionofmultitudinousamountsof datadowntothemeaningfulparts.

Therearethreedatareductionstrategies:

ISSN: 2250-3676
Volume-4, Issue-2, 214-219

2.1 Dimensionality Reduction

It involves convertingdataofveryhighdimensionalityinto that withlowerdimensionalitysuchthateachofthe lowerdimensions conveysmuchmoreinformation[5].
Heuristicmethodscombiningforwardselection and backwardelimination are also a part of this process. [6].

2.2 Clustering

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 others. It is used for exploratory data mining and stastical analysis in fields like retail, manufacturing, bioinformatics etc. Clustering is ineffective if data is "smeared"[7].

2.3 Sampling

It is used in conjunction with skewed data.It involves obtaining a small sample to represent the whole data set allowing a mining algorithm to run in complexity that is potentially sub-linear to the size of the data. Key principle of sampling is based on choosing a representative subset of the dataset. Simple random sampling may have poor performance in the presence of skew.

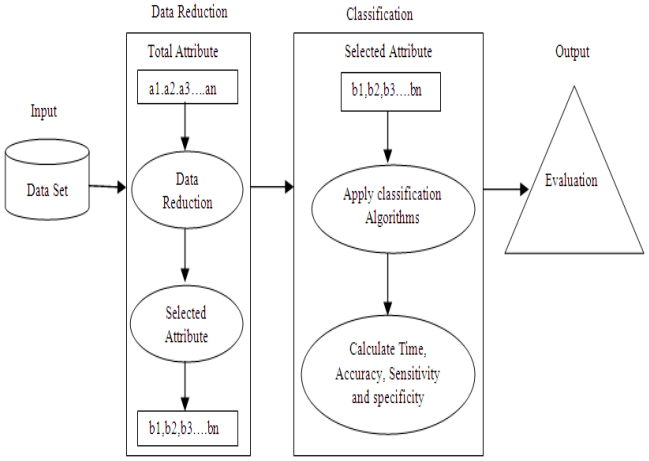


Fig 1: DataReduction and ClassificationProcess

3. DATA REDUCTION TECHNIQUES

Therearefive main datareductiontechnique which are as follows:

3.1 Gain Ratio Attribute Evaluator:

It 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 belongs to one branch. It takes the number and size of branches into account when choosing an 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].

3.2 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].

3.3 Information Gain Attribute Evaluator:

InfoGainAttributeEval evaluates the worth of an attribute by measuring the information gain with respect to the class 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].

3.4 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].

3.5 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

ISSN: 2250-3676
Volume-4, Issue-2, 214-219

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].

4. RELATED WORK

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.
Willetty et al [4] proposed Data Reduction Techniques 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..
Fodo et al [9] proposed a survey of dimension reduction techniques. He said that some computationally expensive novel methods can construct predictive models with high accuracy from high-dimensional data.

5. 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:

5.1 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.

5.1 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.

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. There are many types of classifiers are available like tree, bayes, function rule. Basic aim of classifier is predict the appropriate class [19].

a) 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.

Table1.NSL-KDD DatasetAttributes

Total Attributes		
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
loggedin	srv_serror_rate	dst_host_rerror_rate
num_compromised	rerror_rate	dst_host_srv_rerror_rate
root_shell	srv_rerror_rate	class

ISSN: 2250-3676
Volume-4, Issue-2, 214-219

Table II. Attributes after applying different data reduction algorithms

	FULL DATASET	Cfs Subset Eval	Gain Ratio Attribute Eval	Info Gain Attribute Eval	One RAttribute Eval	Wrapper Subset Eval
Volume (Mb)	20.4	10.8	10	11.8	14.7	16.9
No. of Selected Attributes	42	06	17	19	28	36
NO. of Unpotential Attributes	0	36	25	23	14	6

b) 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.

5. EXPERIMENTS AND RESULTS

WEKA tool is used for experiment, formally called Waikato Environment for Knowledge Learning. 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. KDD cup-99 dataset is used for experimental evaluation, Most pattern classification methods are not able to process

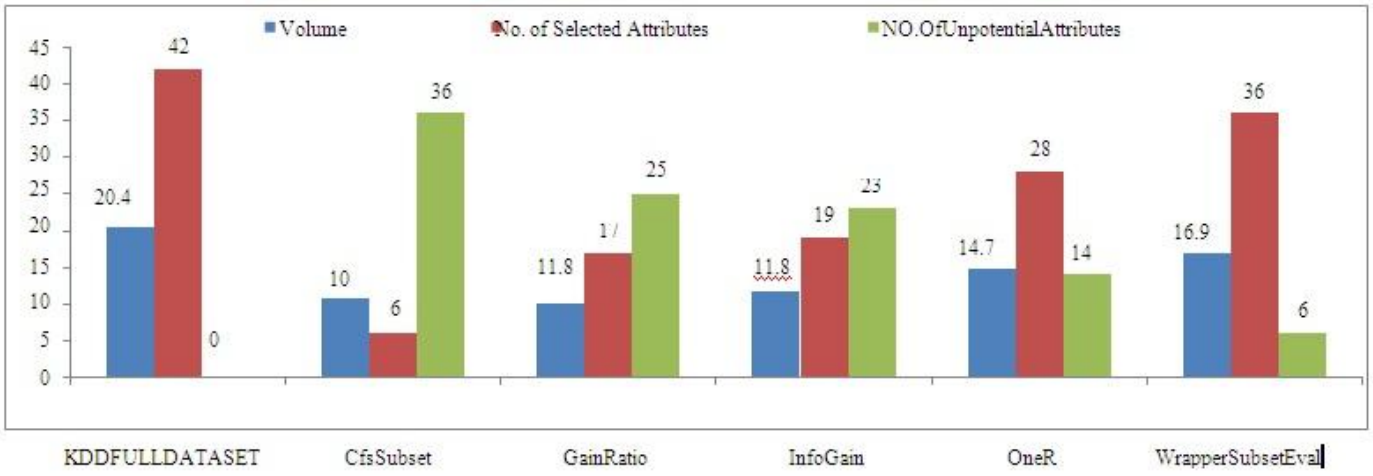


Fig 2. No. of Selected attributes after applying data reductional algorithm

data in such a format. Hence pre-processing was required.

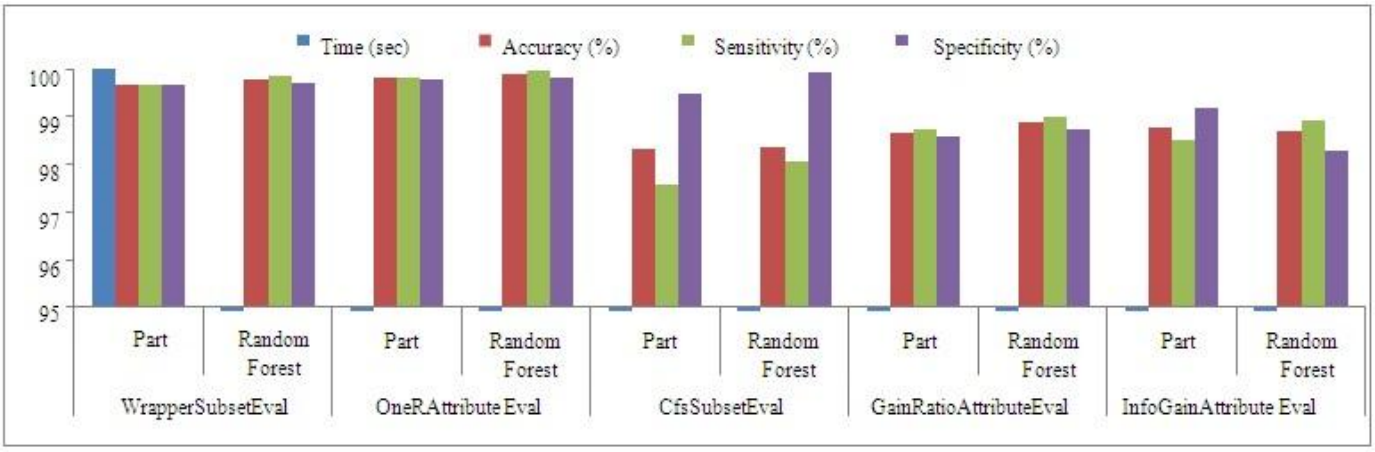
Pre-processing consisted of two steps: first step involved mapping symbolic-valued attributes to numeric-valued attributes and second step implemented scaling [22]. KDD have 42 attributes. These are shown in table 1. Table 2 consists of the selected attributes after applying data reductional algorithms over complete dataset and the volume the dataset is reduced to. Each algorithm has 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 classifational algorithms PART and Random Forest and tried to find out over which data reductional 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 classifational algorithm.

ISSN: 2250-3676
Volume-4, Issue-2, 214-219

Table 3. Results for Different Classification Algorithm studyontheNSL-KDD datasetshowsthatOneRattribute

Data Reduction Algorithms	Total Attribute	Selected Attribute	Classification Algorithms	Time(sec)	Accuracy	Sensitivity	Specificity
Cfs SubsetEval	42	6	Part	6.35	98.32	97.57	99.47
			Random Forest	13.03	98.37	98.06	98.92
GainRatioAttributeEval	42	17	Part	27.25	98.67	98.72	98.59
			Random Forest	18.70	98.89	98.99	98.74
Info Gain AttributeEval	42	19	Part	36.19	98.78	98.50	99.19
			Random Forest	2.32	98.70	98.93	98.29
OneR AttributeEval	42	28	Part	82.03	99.81	99.82	99.77
			Random Forest	20.86	99.89	99.96	99.82
WrapperSubset Eval	42	36	Part	104.74	99.68	99.68	99.66
			Random Forest	27.91	99.79	99.87	99.70

Fig 3. Analyzing best Data Reduction Algorithm



6. CONCLUSION

Data reductionalgorithmsreduce massivedata-setto a manageable sizewithoutsignificantlossofinformation represented by theoriginaldata. Theattributeselection methods ofdatareduction techniqueshelptoidentifysomeof the important attributes, thusreducing the memory requirement aswellasincreasethespeedofexecution.The purposeofthisexperimental workistofindoutwhichdata reduction algorithmgivesbetterresults in big datasets.Theworkwas carried outintwophases.Firstphasewastoreducethedataset andselectpotential attributesthatwillbegivenasaninputto second phase. In second phase the two classification algorithms,PART andRandomForest were introduced. Thenonthebasisofoutputfrom classification algorithms, comparison was done to find out which data

reduction algorithm outperformed intheexperiment. The evaluation provestobethebestamongallthedatareduction techniques.

REFERENCES

1. P.Tao,C.Xiaosu,L.HuiyuandC.Kai,"'DataReductionforNetwork ForensicsUsingManifoldLearning",*Sch.ofComputerSci.&Technol., HuazhongUniv.of Sci.&Technology*.Wuhan,China,pp.1-5,2010.
2. M. Rouse, "Data Reduction", Last UpdatedonAugust 10,[Available Online] http://searchdatabackup.techtarget.com/definition/data-reduction.
3. E.Namey,G.guest,L.Thairu,L.Johnson, "Data ReductionTechniques forLargeQualitativeDataSets",2007, pp137-162[AvailableOnline]http://www.stanford.edu/~thairu/07_184.Guest.1sts.pdf..
4. R. Georgescu, C. R. Berger, P. Willett, M. Azam_, and S.Ghoshal,"ComparisonofData ReductionTechniquesBasedon the Performance ofSVM-type Classifiers".*Dept. ofElectr.andComp. Engineering,UniversityofConnecticut, Storrs, CT 06269,QualtechSystems Inc.,Wetherseld,USA*, 2010.

ISSN: 2250-3676
Volume-4, Issue-2, 214-219

5. A. Ghodsi, "Dimensionality Reduction", Technical Report, 2006-14, DepartmentofStatistics andActuarialScience,UniversityofWaterloo, pp.5-6, 2006.6. Ricardo Gutierrez, "Dimensionality reduction", *Lecture Notes on Intelligent SensorSystems*, WrightState University [AvailableOnline] http://research.cs.tamu.edu/prism/lectures/iss/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.../03Preprocessing.ppt.
9. J.B.Tenenbaum,V.d.Silva,andJ.C.Langford,"AGlobalGeometric Framework for Nonlinear Dimensionality Reduction" *Science*, 290(5500),2000.
10. AshaGowdaKaregowda, A. S. Manjunath&M.A.Jayaram, "ComparativeStudy ofAttributeSelectionUsing Ratio and Correlation BasedFeature Selection",*InternationalJournalofInformation TechnologyandKnowledgeManagement*,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,"UsingInformationGainAttributeEvaluationto Classify SonarTargets",*17thTelecommunications forum(Telfor)*,Serbia, Belgrade, pp.1351-1354, 2009.
13. S.B.Aher,Mr.LOBO,"DataMininginEducational Systemusing WEKA".*International Conference onEmergingTechnologyTrends (ICETT'11), pp*.20-25, 2011.
14. I.K.Fodor,"Asurveyofdimensionreductiontechniques",Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory, 2002.
15. E. Robert, B. Anupama, andC. Trent, "A novel data reduction technique", *Seventh Annual Workshop on Cyber Security and Information Intelligence Research*,OakRidge,Tennessee,ACM,2011.
16. P.FurtadoandH.Madeira,"AnalysisofAccuracyofDataReduction Techniques",*University ofCoimbra*.Portugal,1999.
17. S.H.Vege,"Ensemble ofFeature SelectionTechniquesforHigh DimensionalData",*MastersTheses&SpecialistProjects*,[Available Online]http://digitalcommons.wku.edu/theses/1164
18. *B.Neethu*,"ClassificationofIntrusionDetectionDatasetusingmachine learning Approaches". *International Journal of Electronics and ComputerScienceEngineering*,pp. 1044-1051,2012.
19. N.S.ChandolikarandV.D.Nanadavdekar,"ComparativeAnalysisof TwoAlgorithmsforIntrusionAttackClassificationUsing KDD CUP Dataset",*InternationalJournal of ComputerScience andEngineering (IJCSE)*,pp.81-88,Aug 2012.
20. I.Maglogiannis,K.Karpouzis,M.Bramer,andS.Kotsiantis,"Local Ordinal Classification,"in*ArtificialIntelligenceApplications and Innovations*.vol.204:Springer US,pp.1-8,2006.
21. G.Kalyani,A.J. Lakshmi,"Performance AssessmentofDifferent ClassificationTechniquesforIntrusionDetection",*Journal of Computer Engineering*,vol.7,no.5, pp25-29, Nov-Dec.2012
22. NSL-KDD dataset,[Available Online]http://iscx.ca/NSL-KDD/
23. WekaUser Manual,[Available Online], www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf