3rd International Conference on New Horizons in Science, Engineering and Management and Humanities
IIMT College of Engineering, Greater Noida (India)
19th April 2019 www.conferenceworld.in
ISBN : 978-93-87793-82-8

A Review on Data Partition Method with Feature Selection Method for Big Data

Prof. Patel Falguni N[1], Prof. Swati D. Bendale[2], Prof. Sneha A. Gaywala[3]

PhD Scholar, GTU & Assistant Professor, Information Technology Department, S.V.I.T., Vasad, Gujarat, India[1], Assistant Professor, Information Technology Department, S.V.I.T., Vasad, Gujarat, India[2,3]

Abstract: we *present the essentials of data mining, data reduction methods and data partitioning methods. We also give view on feature selection as a part of data reduction method with big data. Big data has volume, velocity, variety etc. but here, we concentrates on volume of data which are used for data mining process. So, such big data are partitioned based on different approach, their partition types are discussed here. big data partitioning approach and its effect with feature selection method are discussed as a comparative literature survey which is helpful for researcher to work ahead in data partition and data reduction methods together for enhance performance of big data processing.*

Keywords: *Big Data, Data Reduction, Data Partition, Feature Selection, Horizontal Vertical partition*

I. Introduction:

In recent days, As digital technology and its application increase, large data also increase day by day. The important question is-how to utilize this data for decision making, prediction trend or situation, support to industry, organization etc [1][5]. data analysis or discovering knowledge called data mining[1]. Data mining methods like association rule mining, outlier detection, clustering, classification, regression and summarization. Data mining process include data collection, pre-processing, apply data mining method and result evaluation. Data pre-processing include data filtering, selection, reduction, transformation, normalization etc. sub task to prepare data such that data mining method accept it as input and so get proper output from method. The paper's center topic is data reduction for high dimensional dataset and data partition. Handling large high dimensional dataset is toughest job as traditional data mining methods cannot perform better. Now days, distributed parallel data mining environment can handle data which has following properties: Volume in terms of samples/features, Velocity like arrival/processing speed of data, Variety of data like structured, unstructured data (text, images, videos etc.). Other V's also included as big data definition but above 3V's is generally used to define properties of data. Big data environment provide processing of data in parallel distributed processing, which is helpful to reduce processing time [23].

ISBN : 978-93-87793-82-8

1.1 Data Reduction:

Data reduction is categorized into three types like feature reduction, instance reduction and data compression as seen in fig.- 1 . Some data have high dimensional data need to be handling with feature reduction technique [12]s. Feature reduction has been performed with feature selection, feature extraction or construction. Instance reduction has been handled with instance selection and instance construction [9-11][20][25]. Data compression and decompression is used to reduce size of data but analysis is difficult with compressed data. So, time and processing cost increase to do decompression and compression. High dimensional dataset means more number of features in data and dataset examples are text related articles, documents etc., medical data like genome data etc. Large dataset means more number of samples in data. In this paper, we are considering large high dimensional dataset which process on big data environment.

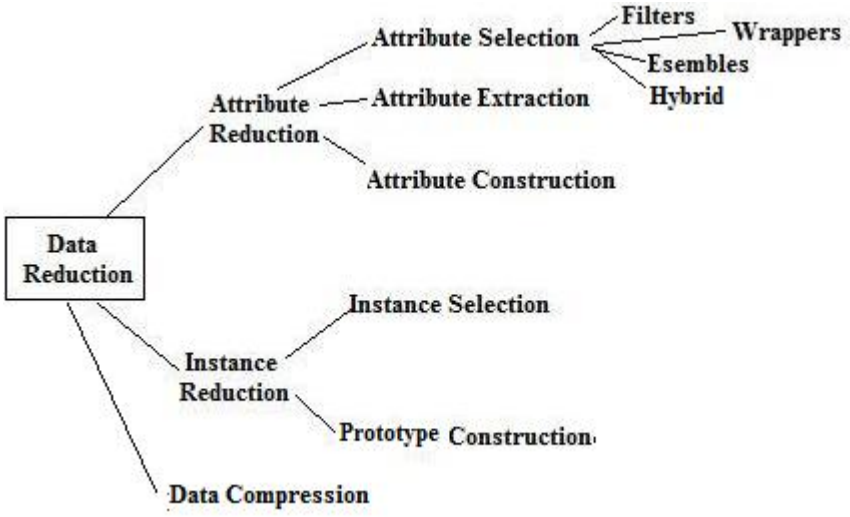


Fig. 1 Data reduction methods classification [25]

1.2 Data partition:

Distributed parallel processing has many data storage methods. Data storage on different nodes that means split whole dataset among different nodes. Our paper includes survey of available data store methods with data reduction methods which are mention pros and cons of each method. Data storage method scale data on different node such that data processing or mining process make easy, fast, less costly and scalable. Distributed parallel environment has one concept which provide like local data processing called replication. but data replication required data consistency problem means each node communication each time when data updated on any site. So, the solution is handle data partitioning such a way that less processing cost, fast outcome and less inter node communication. In this paper, literatures of data partitioning methods with feature selection methods are discussed.

Different framework may follow different partitioning but generally they perform horizontal partitioning like hadoop framework used hash or range base horizontal portioning. If their is requirement for vertical partitioning then dataset will be transposed as mentioned by some authors in their papers [19].

ISBN : 978-93-87793-82-8

II. Feature selection:

High dimensional data has large number of features or attributes in big data, may have irrelevant and missing features which is not helpful to analyze data properly[9][14]. So, feature selection plays an important role as to select relevant and important features. Feature selection methods are filters, wrappers, hybrid and embedded. Filters are based on statistical methods which is fast but less accurate compare to others. Filters are used mostly in big data for fast calculation. Wrappers are feature selection based on learning algorithm or search algorithm[18]. Embedded methods are built/embedded with model. Hybrid means combination of filter and wrappers used as a feature selection [20]. In big data, main question is how to perform proper feature selection on distributed partitioned dataset. That means there is a need to study relation between data partition and its effect on feature selection methods. In this paper, we compare some exist literature for data partition and effect of feature selection method in big data environment.

III. Importance of Data partition:

For large scale data processing, data are divided onto different nodes which work parallel execution. For big data, same way dataset are partition among nods or sites and execution are distributed or send to each node to run locally and global results are merged. This is basic important feature of big data, rest of other imperative features are listed below [21][24]:

1) Scalability: whole dataset is stored distributed among nodes instead of centralized manner. So, when data size increase, there is no hardware limit for storage. It would be easy to add nodes in cluster as and when required. So, it solves scalability issue of centralized approach.
2) Security: As data partition applied on dataset, there is a possibility to split sensitive and non-sensitive data and apply separate security control on local dataset.
3) Flexibility: partition dataset and store on nodes such a way that administrative and transactional flexibility increase.
4) Availability: Failure in any node or data, can be managed by replication on other nodes without knowledge of end users and performed task properly.
5) Performance: parallel distributed execution of data for specific work decrease time complexity in terms of data access or processing. For data mining task like pattern finding, classification, clustering get outcome without degrading performance. Repartition/dynamic partition of data for increase performance of given task is an important feature[1][2].

Disadvantage of data partition:

1) Communication cost: As data partitions are stored on different nodes, local processes work with some little communication of data. The question is – how to minimize or lower the communication cost among nodes. Second thing is minimize the sensitive data in communication.
2) Data integration: different nodes generate some outcome after given task then global integration policy needed to be design for any data mining task.

ISBN : 978-93-87793-82-8

3) Real time data with replication: real time data arrive as per time basis, how to store and replication among nodes is not defined properly and so processing cost may increase.

IV. Types of data partition:

Basic introduction of data partition with its advantage and disadvantage are discussed above sections. In this section, different types of data partitions are discussed bellow [3][2]:

4.1 Attribute level data partitioning algorithms: partition of dataset depends on features group which reduce communication cost and maintain accuracy of particular application [5]. Set of features in each partition may be overlapped or non-overlapped. The scenario shown in fig -2.

4.2 Workload aware/transaction based data partitioning: As per processing cost on each node that means maximize local processing and minimize communication cost, data partition strategy applied. This type of partition method suited for business transaction based applications. That means load balancing and transaction balancing based partition occurs.

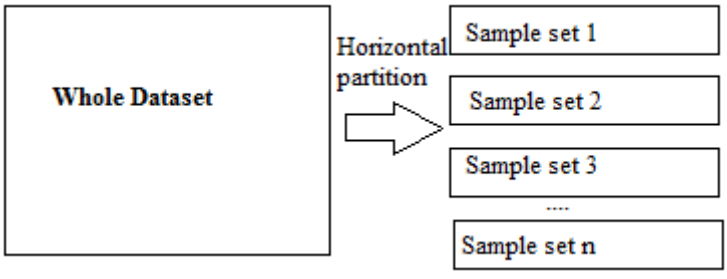
4.3 Graphical / location based data partitioning: database tuples are mentioned node in a graph and authors are trying to minimize complexity of graph.

4.4 Dynamic data partitioning / re-partitioning: in static partition, data fragment is same on each node but in dynamic partitioning, data fragments or partition after some time or in each phase depend on transaction requirement.

4.5 Clustering based data partitioning: for frequent attributes finding in some dataset is get by clustering method of features. That means for large dataset, affinity measure between features are not handled by other methods, clustering method works.

4.6 Sample based data partitioning: In this method, data samples are divided into data sample subsets on different nodes called horizontal partitioning, which is useful for large dataset mining as shown in fig. - 2.

4.7 Proportional partitioning: data is divided with percentage proportional values on different node for example 5% data or sorted data with top 10 % on one data node etc.



ISBN : 978-93-87793-82-8

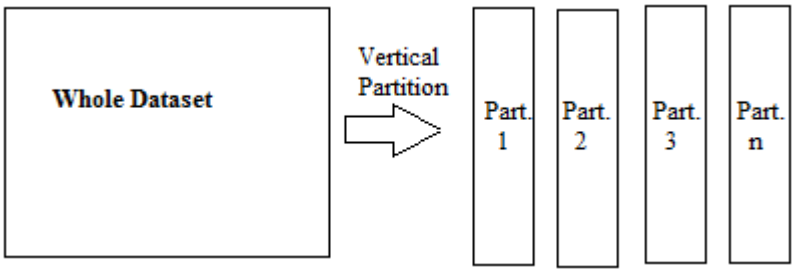


Fig. 2 Horizontal and Vertical Partitioning [22]

V. Data partition and Feature selection method:

In big data, data partition is a significant step for data analysis which reduces complexity and cost. Data partition on distributed parallel nodes [17][21], mainly focus on attribute wise partition/vertical partition, sample partition/horizontal partition, hybrid partition (vertical and horizontal), workload/load balance partition, location based dynamic partition. Others types are data access/process based partition, user's data usage based partition, user's required output accuracy based partition etc. as shown fig. – 3 define feature selection methods on vertical partitioning dataset as local and combine as global feature selection[13][14]. In given table, list of literature regarding feature selection and partition methods jointly discussed with its effect as advantage and future suggestions as well.

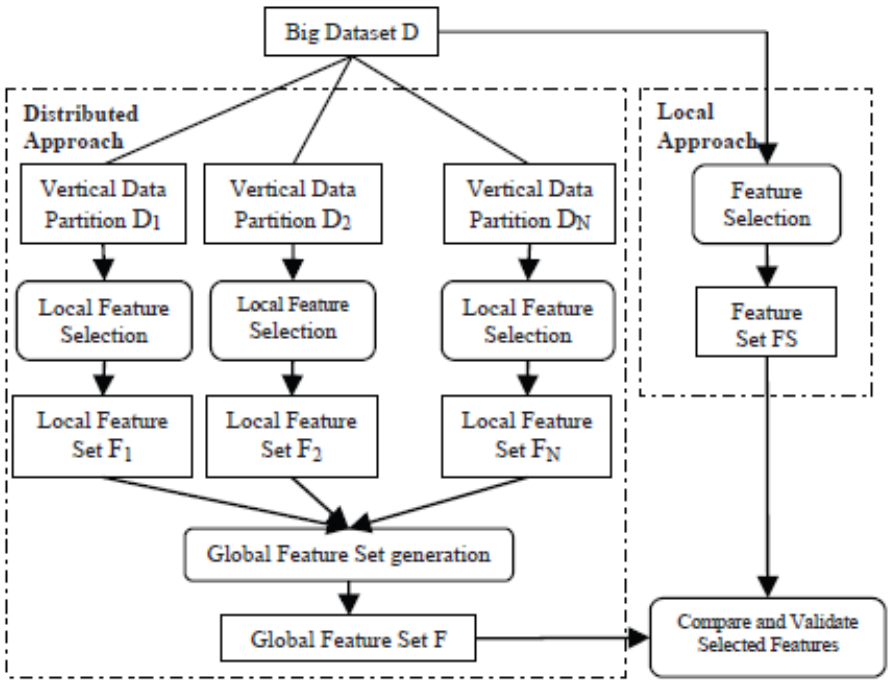


Fig. 3 Feature selection on vertically partition distributed data method and Centralized feature selection method [4]

ISBN : 978-93-87793-82-8

Title	Publication	Partition type	Dataset and Outcome	Advantage	Future Work
Towards parallel feature selection from vertically partitioned data[5	ESANN, 2014	Vertical Partition	1)Isolet-617 features, 7474samples,26 classes 2)Madelon-500 features,2400 samples, 2 classes 3)Mnist-717 features,60000 samples, 2 classes -Accuracy and computational time parameters are compared.	-Accuracy is maintained as local node execution with parallel dist. Nodes. -computational time is decreased in parallel environment compare to local node. -Accuracy is high compare to parallel horizontal partition.	-for large dataset, computational time increase linearly. -Filters are only compared with classifiers. other feature selection methods will be applied.
Distributed feature selection: an application to microarray data classification[7]	Elsevier, 2015	Vertical Partition	1)colon-2000 features,62 samples 2)DLBCL-4026 features,47 samples 3)CNS-7129 featurs,60 samples 4)Leukemia-7129 featurs,72 samples 5)Prostate-12,600 features,136 samples 6)Lung-12,533 features,181 samples 7)Ovarian-15,154 features, 253 samples 8)Breast-24481 features,97 samples -Accuracy and computational time parameters are compared.	- This scheme handles classification even if features are more compare to number of samples. -Accuracy of parallel environment is maintained and less computational time compare to centralized classification.	-Rank based method [15] proved good but larger dataset which are unable to fits in partitions, Question is- how to find rank and perform rank based partition? -wrappers are used but accuracy is same as filters in parallel nodes.
A time efficient approach for distributed feature selection partitioning by features [6]	Springer, 2015	Vertical Partition	1)Isolet-617 features,7474samples 2)Madelon-500 features,2400 samples 3)Mnist-717 features,60000 samples 4)Breast-24481 features,97 samples 5)Lung-12,533 features,181 samples	- Repartition of dataset are used in each round with count of selected each features. Remove features by threshold. -Author defines new threshold equation with data complexity instead of classification error.	-Difficult to find redundant features when data partitions are distributed.
Hadoop based feature	IJST, 2016	Horizontal Partition	Kdd1,2,3,4, dataset with 60000 samples approximate in each	-Rough set as feature selection method is used. -Random forest is good	-In real-time distributed data, uncertainty and

ISBN : 978-93-87793-82-8

selection and decision making models on big data			set	compared to decision tree classifiers on big data.	missing values are big issues. -Different attributes have values whose data distribution is uniform. -Required new feature selection method to handle mixed attributes.
Prediction with partitioning: big data analytics using regression techniques [1]	IEEE, 2017	Horizontal Partition	Bike sharing dataset	-Regression model on each partition used locally and compare value at globally. -Multivariate linear regression is good compare to single, multiple linear regression.	-To defeat big data issue, new serial partition method of subsets as an alternative of samples partition.
Scalable feature selection via distributed diversity maximization	AAAI - 2017	Vertical Partition	-biological, text, image dataset like -Diversity maximization as feature selection used. (distance among features and MI among feature-class)	-vertical feature selection with handling redundancy. -minor accuracy increase in case of specific dataset.	- Develop more promising objective function to deal with redundancy in vertical partition data.
A new horizo-vertical distributed feature selection approach	Cybernetics and IT,2018	Horizontal and Vertical Partition	-Isolet, 11-tumor, Madelon dataset - CFS and mRmR feature selection methods[16] used for irrelevant and redundant features respectively.	-accuracy is little increase compare to centralized, horizontal, vertical partitioned data. C4.5, kNN, Naïve Bayes, SVM classifiers are used. -Features majority voting is used globally after combining horizontal and vertical selected features.	- Authors used small samples and features data for experiment, but in future large high dimensional dataset will be used. -Time complexity is not compared for all approach like centralized, horizontal, vertical partition.
Feature selection in high-dimensional dataset using mapreduce	arXiv, 2017	Horizontal partition	-Four artificial dataset with large sample or high dimensions. -mRmR used as feature selection[16].	-horizontal partition is applied on transposed high dimensional data so feature wise horizontal partition can perform and better result compare to sample wise partition of data.	-develop novel feature selection method which works with distributed feature-sample data matrix on nodes.
Centralized vs distributed	Elsevier, 2016	Horizontal and Vertical Partition	-Connect4,Isolet, Madelon, Ozone, Spambase, Mnist dataset used	-authors compared centralized approaches with distributed approach have horizontal, vertical partition.	-feature retain threshold calculation time is increase overall execution

ISBN : 978-93-87793-82-8

feature selection methods based on data complexity measures [22]			- time complexity(speed) and accuracy is compared	-for high dimension data, vertical part. Is good. -for feature selection, percentage of feature retain is depend on complexity which is calculated by fishers score.	process. -new approach for horizontal-vertical partition combines together in future.
Data partitioning view for mining big data[2]	arXiv, 2016	Horizontal partition	-sample synthetic temperature dataset	-authors discuss attribute based vertical partition and percentage proportional partition method (PPP). -PPP method used for some users want specific filter data which may be part of whole data like 5% / 10% etc.	-define new global model for local patterns integration.
Distributed feature selection using vertical partitioning for high dimensional data[4]	IEEE,2016	Vertical partition	-ipums, mfeat, mfeat-fouriers, covtype, optdigit -information gain used as feature selection(FS) is used.	-centralized and distributed vertical partition used IG then selected top 5features are same in both case means accuracy maintained and execution time is decrease in dist.env.	- Instead of IG other FS measure can be used in future. - Horizontal and vertical both partition can be used for reduce computational time.

Table 1. Literature of data partition with feature selection methods

VI. Conclusion:

Our paper includes brief introduction about data partitioning methods with its advantage and disadvantage for distributed parallel processing specific for data mining task. We discussed the data partition and data reduction-feature selection methods in literature with proper table format which help to find out till research work and give new idea to researcher by referring future work from literature. In future, instance selection methods and data partitioning methods comparative survey will be projected for distributed parallel environment.

REFERENCES:

[1] K. Saritha and S. Abraham, "Prediction with partitioning: Big data analytics using regression techniques," 2017 International Conference on Networks & Advances in Computational Technologies (NetACT), Thiruvanthapuram, 2017, pp. 208-214.doi: 10.1109/NETACT.2017.8076768.

[2] Shichao Zhang, “Data partitioning view of mining Big data”, Arxiv 2016.

ISBN : 978-93-87793-82-8

[3] S. Phansalkar and S. Ahirrao, "Survey of data partitioning algorithms for big data stores," 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, 2016, pp. 163-168.doi: 10.1109/PDGC.2016.7913137.

[4] B. R. Prasad, U. K. Bendale and S. Agarwal, "Distributed feature selection using vertical partitioning for high dimensional data," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 807-813.doi: 10.1109/ICACCI.2016.7732145, IEEE,2016.

[5] V. bolon canedo, noelia sanchez, joana cervino, "Toward parallel feature selection from vertically partitioned data", ESANN, 2014.

[6] V. bolon canedo, L. marono, A. Alonso, "A time efficient approach for distributed feature selection partitioning by features", Proceedings of the 16th Conference of the Spanish Association for Artificial Intelligence on Advances in Artificial Intelligence - Volume 9422, Pages 245-254, November 09 - 12, 2015, Springer-Verlag New York, Inc. New York, NY, USA ©2015.

[7] V. bolon canedo, noelia sanchez, joana cervino, "Distributed feature selection : An application to microarray data classification", Applied soft computing, Elsevier, Volume 30, Pages 136-150, May 2015.

[8] V. bolon canedo, N. sanchez, A. Alonso, "Recent Advances and emerging challenges of feature selection in the context of Big Data", Journal Knowledge-Based Systems, Volume 86 Issue C, Pages 33-45, September 2015.

[9] Escalante, H.J., Marin-Castro, M., Morales-Reyes, A. et al. "MOPG: a multi-objective evolutionary algorithm for prototype generation " Pattern Anal Applic (2017),Springer, Vol- 20: pp. 33-47,doi:10.1007/s10044-015-0454-6

[10] I Triguero, D Peralta, J Bacardit, S García, F Herrera, "MRPR: a MapReduce solution for prototype reduction in big data classification", neurocomputing, 2015, Elsevier,vol-150,pp. 331-345.

[11] Arnaiz-González, Á., González-Rogel, A., Díez-Pastor, JF. et al., "MR-DIS: democratic instance selection for big data by MapReduce", Prog Artif Intell (2017),Springer. doi:10.1007/s13748-017-0117-5 .

[12] UR Rehman, Muhammad Habib, et al. "Big data reduction methods: a survey." Data Science and Engineering 1.4 (2016): 265-284.

[13] Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, "Feature selection for high-dimensional data", A. Prog Artif Intell (2016) 5: 65. doi:10.1007/s13748-015-0080-y .

[14] Li, Y., Li, T. & Liu, H., "Recent advances in feature selection and its applications" Knowl Inf Syst (2017). doi:10.1007/s10115-017-1059-8

[15] P. Gupta, A. Sharma and J. Grover, "Rating based mechanism to contrast abnormal posts on movies reviews using MapReduce paradigm," 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, 2016, pp. 262-266.doi: 10.1109/ICRITO.2016.7784962.

[16] Ramírez- Gallego, Sergio, et al. "Fast- mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High- Dimensional Big Data." International Journal of Intelligent Systems 32.2 (2017): 134-152.

ISBN : 978-93-87793-82-8

[17] L. Zhao, Z. Chen, Y. Hu, G. Min and Z. Jiang, "Distributed Feature Selection for Efficient Economic Big Data Analysis," in IEEE Transactions on Big Data, vol. PP, no. 99, pp. 1-1.doi: 10.1109/TBDATA.2016.2601934.

[18] Daniel Peralta, Sara del Río,Sergio Ramírez-Gallego,Isaac Triguero, Jose M. Benitez, Francisco Herrera, "Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach", Mathematical Problems in Engineering,Volume 2015 (2015), Hindawi, Article ID 246139.

[19] S. Ramírez-Gallego et al., "An Information Theory-Based Feature Selection Framework for Big Data Under Apache Spark," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. PP, no. 99, pp. 1-13. doi: 10.1109/TSMC.2017.2670926

[20] Wang, Jiaheng, Xue, Bing,Gao, Xiaoying, Zhang, Mengjie, "A Differential Evolution Approach to Feature Selection and Instance Selection", Trends in Artificial Intelligence: 14th Pacific Rim International Conference on Artificial Intelligence, Phuket, Thailand, August 22-26, 2016, Springer International Publishing, pages=588--602, isbn=978-3-319-42911-3,doi=10.1007/978-3-319-42911-3_49

[21] Dr. Shraddha phansalkar, dr. swati ahirrao, "Survey of data partitioning algorithms for big data stores", International conference on parallel, distributed and grid computing, IEEE, 978-1-5090-3669-1/16,2016.

[22] L. Morán-Fernández , V. Bolón-Canedo , A. Alonso-Betanzos , "Centralized vs. distributed feature selection methods based on data complexity measures", Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.09.022..

[23] L'Heureux, Alexandra, et al. "Machine Learning with Big Data: Challenges and Approaches." IEEE Access (2017).

[24]https://www.coursehero.com/file/p1rijh8/Advantages-and-disadvantages-of-horizontal-and-vertical-partitioning-Advantages/

[25] F. N. Patel, "Large high dimensional data handling using data reduction," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 1531-1536. doi: 10.1109/ICEEOT.2016.7754940.