XVII International Conference on Recent"Trends in Engineering, Applied Science and Management"
(IETE) Institution of Electronics and Telecommunication Engineers,Chandigarh TEAM-18
24th June 2018 www.conferenceworld.in ISBN: 978-93-87793-31-6

Big Data Feature Selection Methods: A survey

Falguni N. Patel[1], Dr. Hitesh Shah[2], Dr. Shishir Shah[3]

PhD Research scholar, Gujarat Technological University, Gujarat, India[1]

Professor & Head of E.C. Department, GCET, Gujarat, India[2]

Professor, CS Department, University of Houston, USA[3]

ABSTRACT

In a high speed era, digital information is increase in exponential manner which are useful in corporate, institute, science, engineering and technology etc. area for making specific decision and prediction. Big data analytics play an important role as data mining techniques are not capable to handle these big data. big data having large, complex and velocity characteristics which are research area now a days. For large volume data, it having large high dimensions need new or modified existing feature selection techniques. In this paper, we have discussed difference feature selection methods like filters, wrappers, embedded and hybrid. We have also discussed use of feature selection method in big data are till now introduced for specific applications. Here, in this paper, some feature selection filter based methods are tested with distributed parallel environment of big data and it performed better compare to original dataset in terms of time and accuracy.

Keywords: *Big Data Analytics, Feature Selection, Feature Extraction, Distributed Parallel Processing, Data Reduction*

1. INTRODUCTION

As in a fast growing digital world, data arrival is growing rapidly and need fast processing for data mining or analysis to make accurate decisions which are useful in real scenarios like weather forecasting, sentiment analysis, prediction and management. Machine learning and data mining play important role but for big data, it is not easy to handle huge speedy and complex data. Big data has 3 V's as volume, velocity, variety mainly and more V's are available like veracity, value etc. large and huge amount of data is not always important for analysis[1][2]. This data may contain noisy, irrelevant and redundant features or instances which may decrease accuracy and take more time for classification and clustering. In data pre-processing step, there are two approach used. First is data preparation and second is data reduction. Data preparation is compulsory steps for any classification, clustering etc. data preparation include data cleaning, transformation, normalization etc. Data reduction is optional step as it include feature reduction, instance reduction, data compression etc.

From high dimensional dataset, selecting small important features method called feature reduction. Feature reduction can be done by feature selection and feature extraction. Feature selection means selecting some important features/attributes from all features which reduce its feature size[3][4]. While feature extraction means transform features from one dimension to other dimensions such a way that it reduce feature set size[5].

Some of the big datasets as per application wise lists are discussed in second section. Feature selection methods in details are discussed in third section. Feature selection related some experiments results define in fourth section.

1.1 Big Dataset list

The mass generation of high dimensional and huge data in engineering, science etc area, need to be analyzed. This big data in different domain are collected and this domains and dataset are listed here :

Plants, Nutrient, soil, weather dataset in agriculture domain, Microgene, protein datasets in biology domain, sea climate, weather batch and real time dataset in climate/weather domain, patent, scopus, graph dataset in complex network domain, NASA earth observation, earthquake, water resource datasets in earth science domain, stoke exchange, trade market datasets in economics domain, student result and financial status, full/part time course outcome datasets in education domain, power utilization and consumption datasets in energy domain, graph and map of safelight dataset in GIS domain, feedback of policy, scheme of government in government domain, disease prediction like tumour, cancer, diabetes etc in health care, face, text reorganization as in image processing domain, neuroscience domain, public domains like Google news, yahoo news, Amazon, social networking domain like facebook, twitter [8] etc. analysis, software testing domain like software error checking and solution analysis, bug detection, in sports domain like winner analysis, in domain where time series and spatiotemporal data, real time transportation analysis and updation, prediction of events etc. in transportation domain. Above mentioned all domain with specific activities, need more enhancement and some new concepts to solve real application problem which we face day to day life[17].

2. BIG DATA FEATURE SELECTION METHODS

Big data having characteristics of large amount of volume may have large number of samples or large number of high dimensions, complex data like relational database like structure, email like semi-structure, text, audio, video like unstructured data, speedy data arrival and analysis with batch, less speed or high speedy collected data with real time analysis. This all characteristics are some way connected to all domains which are explained in above section. In this paper, we focus on high dimensional data. To analysis high dimensional data, process is affected with noisy, redundant or irrelevant features[20]. In data mining, many feature reduction techniques are exist but for big data, they face many problems. This is growing area and required more research. In big data, some of the feature selection methods are introduced till now, among them most of methods are classified in four methods, discussed below which are helpful for further research work.

Fig. 1: Feature Selection Method - Filters Method

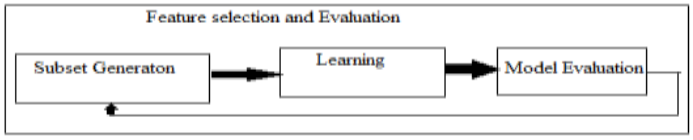

Fig. 2: Feature Selection Method - Wrappers Method

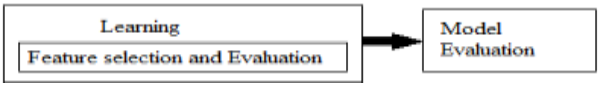

Fig. 3: Feature Selection Method - Embedded Method

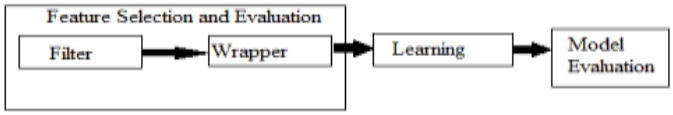

Fig. 4: Feature Selection Method - Hybrid Method

Methods	Variables	Advantage	Disadvantage	Examples
Filters	Univarite	-Fast -Scalable -Independent of the classifier	-Ignore feature dependencies -less accuracy because of classifier not used for FS	-Chi Square -Information Gain -Gain ratio
	Multivariate	-Models feature dependencies -Independent of the classifier -Better computational complexity than wrapper	-Slower than univariate -Less scalable than univariate -less accuracy because of classifier not used for FS	-Correlation based FS -Fast correlation based FS
Wrappers	Deterministic	-simple -interact with classifier (feedback) -models feature dependencies -less comp. comple. Compare to randomized	-over fitting risk -more prone than randomized to get stuck in local minima -classifier dependent selection	-Sequential Forward selection -Sequential Backward selection

	Randomized	-less prone to local minima -interact with classifier (feedback) -models feature dependencies	-computational intensive -over fitting risk than deterministic approach -classifier dependent selection	-Simulated annealing -randomized hill climbing -genetic algorithm -estimation of distribution algo.
Embedded		-interact with classifier Better computational complexity than wrapper -models feature dependencies	-classifier dependent selection -less prone to overfitting	-decision tree -weighted naïve bayes -FS using weight vector of SVM

Table 1: Comparison of Feature Selection Methods [24]

2.1 Filters Method:

For feature reduction, statistical methods and random sampling are used to select or extract features which are helpful to learn classifier accurately. Feature reduction methods are introduced in classification and clustering also. In this paper, we are concentrating only on classification methods with feature reduction in big data area. Filters are classified as univariate and multivariate. Univariate means feature and class dependency is considered for feature selection while multivariate means between features dependency is considered to select features-see figure -1 and table -1. Filters has feature selection methods like chi-square, mRmR (max. relevance min. redundancy), information gain etc. and filters has feature extractors like feature transformers – PCA (principle component analysis)[18], SVD (singular value decomposition) etc[6][7][19].

In big data, data having large, complex and speed characteristics so some of filter methods are introduced till now which are capable to run on distributed parallel data processing units are discussed here.

Claudio reggiani and etc used artificial binary dataset as high dimensional dataset and apply mRmR (maximum relevancy minimum redundancy) as feature selection method. Authors tried to define that for high dimensional dataset, alternative (features as rows list in data matrix) method is speedy compare to conventional method (features as columns) with mRmR. Here, authors are not compared accuracy of conventional and alternative methods with mRmR[7].

Mihail popescu and etc used random projections for big data classification and compare with PCA method. Random projections are set of projection and applied classifiers as ensemble method with fuzzy majority voting for kNN. PCA is a feature extractor does not work for large dataset and so RP is used. RP is proved good compare to PCA method. Authors work on synthetic and activity datasets[9].

Liang zhao, zhikui etc. authors define efficient economical model using distributed feature selection method on distributed parallel environment. Initially, data pre-processing methods are used like instance

selection with noise removal, missing value imputation and min-max normalization. Then for feature selection, first authors used density based clustering among instances as horizontal reduction and second find each attribute's contribution by correlation measure as vertical reduction. That attribute has less correlations are removed[10].

Sergio ramirez and other authors are used information theory based filter method mRmR on spark framework on distributed parallel processing environment. They used epsilon, dna, ecbdl14, url and kddb datasets as big data on spark. Authors find that selected features on centralized and distributed environment are same and accuracy is not degraded. Total feature selection time on big data is decrease in distributed parallel environment[12].

In conclusion of above survey regarding filter methods, big data with volume characteristic used information theory based mRmR, chi-square, PCA, SVD, mutual information, clustering are used till now. More filters methods can be enhanced in this area is a research for future work. Filters are fast compare to other feature selection methods and this will helpful to big data area but less accuracy compare to wrappers, embedded etc, so accuracy in filters need to be more focus in future.

2.2 Wrappers Method :

They are depends on learning algorithm and therefore it gives good accuracy compare to filters but slow in speed as in figure-2. It has fitness function for select relevant important features. On single machine, wrappers method is performed well but for ultra high dimensional dataset on distributed approach, it is difficult to scale. In big data, some of the authors try to implement this approach as we discuss in further paragraphs.

Daniel peralta, sara del rio ate., used evolutionary algorithm as feature selection for big dataset classification on spark framework. They used CHC(cross generational elitist selection, Heterogeneous recombination and cataclysmic mutation) evolutionary algorithm which include half uniform crossover- HUX, Elitist selection, Incest prevention and restarting process in map process on each separate node on distributed separate dataset and by reduce process number of features selected. These selected features apply on each distributed dataset and applied classification algorithms. Authors used Epsilon, ECBDL14 with ROS (Random Over Sampling because of imbalanced dataset) datasets as big data. kNN algorithm work as fitness function in CHC algorithm for feature selection. Evolutionary algorithm as feature selection performed better in execution time as well as accuracy compare to sequential execution of large dataset[11].

Jiaheng Wang, Bing Xue ate. used binary differential evolution but improved version used. Simple evolutionary used generally classifier as fitness function which is time consuming process. But authors used fitness function measure with filter which is processed fast. This function is depends on interclass Distance and intraclass spread measure. That means inter class has instances should has more distance and intra class has how much class is spread out. Generally smaller spread is preferred in intra class. Authors used fitness function with distance and spread measure in feature selection, instance selection, feature-instance selection combinelly. They prove that IIC-FS, IIC-IS,IIC-FIS is good in execution time and little accuracy compare to kNN-FS, kNN-IS, kNN-FIS. For larger big dataset, this method need for improvement in fitness function is a research scope[13].

2.3 Embedded Method :

It has feature selection process as part of model construction and efficient compare to wrappers because it has no iterative steps to feature selection see figure -3. Embedded methods like LASSO (least absolute shrinkage and selection operator), REF-SVM (recursive feature elimination – support vector machine) etc. used and can be applicable in big data for large, complex and online data.

2.4 Hybrid approach:

For feature selection, sometimes filter and wrappers are used together seen in figure -4. First, applied filter to remove features which reduce size and then wrapper is used which work fast because of small data and get good accuracy with more small features set. In big data, some research is finding hybrid approach. Chia Tien dan lo etc. are work with malware dataset which is unstructured text dataset with 9448 cases and 682936 features big in size. Authors use dimensionality/ feature/attribute reduction method like chi-square in first stage that reduce features up to 68000 features. Then after for further reduce feature authors are used random forest repetitively as reduce features up to 9 and accuracy performance is increased compare to previous accuracy[14].

In conclusion, feature selection using any of above methods proved better in accuracy and fast in execution with reduced dataset in simple dataset. it is also helpful in big data with volume means large data. for complex big data, feature selection is a unsolved issue till now. For velocity big data, online classification with feature selection is improved by researchers but also there is a scope to enhance these characteristics of big data. Experiment: The following section covered datasets which are used for preliminary testing, methods and feature selection and classification results on distributed parallel environment.

3. DATASET AND METHODS:

For experiment, we are selecting one dataset from medical domain DNA dataset and another is text dataset Reuter news 20 dataset used. DNA dataset is binary classification problem and is a very sensitive dataset, has samples around 2000 and features are 180. In Reuter news 20 dataset, it is news documents groups as text data pre-processed as data matrix which has 42 samples and 61381 features with 20 different classes[22]. In Reuter dataset, data samples are randomly selected in equal frequency from each category of class.

Software and hardware: for testing distributed information theory based feature selection, two dataset DNA and news 20 are selected which are related to complex problems. With one name node and six data nodes of 4 GB RAM dual core machines each, Apache hadoop 2.7, apache spark 2.1.1 and scala 2.10.4 are installed. 10 partition of dataset are defined in spark for parallel distributed processing. Information theory based spark package is installed to test DNA and news20 text dataset[12][15].

3.1 Result of Feature selection and classification:

For testing feature selection on big data with distributed parallel tool, information theory based feature selection methods are used. This package is also available on github[16][23]. Basically this method follows feature-feature and feature-class dependency with entropy, which are used to remove irrelevant features. mRmR used difference between relevancy and redundancy of features. Relevancy means mutual information between feature

and class and redundancy means mutual information between features. Following is general feature selection equation and following table-2 is gives various feature selection methods under given constraints. Other derived feature selection methods are mRmR (maximum relevancy and minimum redundancy), MIFS(Mutual Information Feature Selection), MIM(Mutual Information Maximization), CIFS(Conditional Information FS), JMI(Joint Mutual Information), ICAP(Interaction Capping), IF(Informative fragments) and others. For classification, naïve bayes algorithm is used with different number of features and compared accuracy for each feature selection method.

As shown in figure -5 and figure -6, different features sets selected from DNA dataset and tested with naive bayes classifier. Accuracy shown better in mRmR compare to other feature selection methods. Figure- 5 shown feature relevancy and redundancy ratio/value for each feature with different feature selection methods of DNA and News20 datasets. Figure –7 also shown that mRmR is better for news20 text dataset. figure –8 shown comparisons of total execution time of DNA datasets with 150 selected features on single node and multi nodes distributed parallel environment. Figure-8 shown that total execution time is less on multimode, compared to single node and selected features list is same for single node and multi nodes.

$$J = I(X_i; Y) - \beta \sum_{X_j \in S} I(X_j; X_i) + \gamma \sum_{X_j \in S} I(X_j; X_i|Y)$$

Beta	Gamma	FS methods
[0,1]	0	MIFS
0	0	MIM
1/|s|	0	mRmR
1	1	CIFS

Table 2 : General Information theory based Filter method variasion

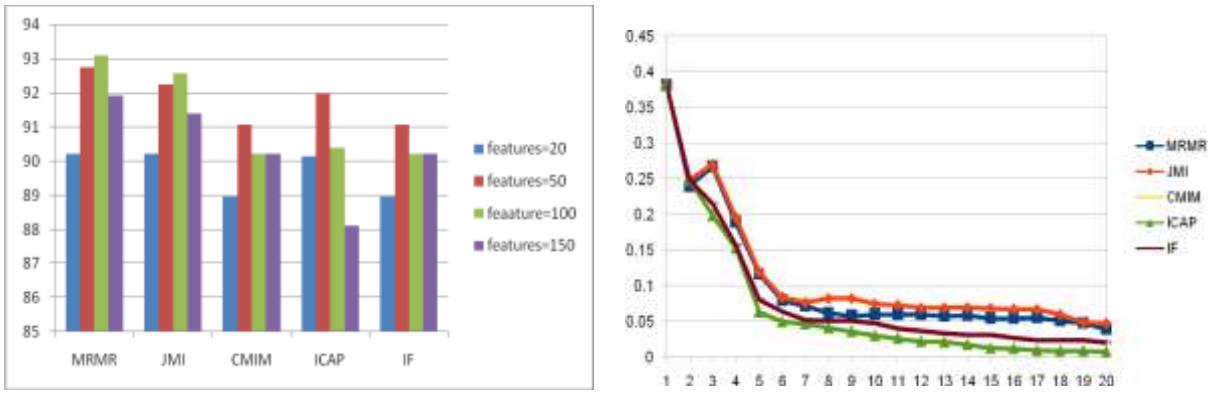

Fig 5: Accuracy chart for different feature with different feature selection methods and feature relevancy-reduction ration chart of DNA dataset

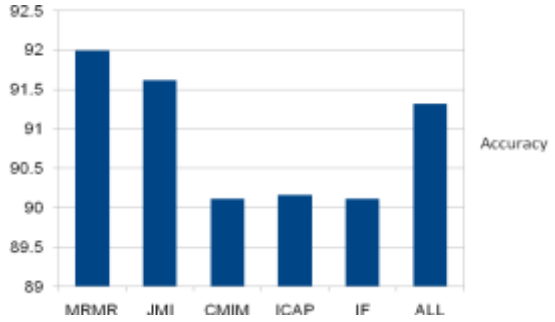

Fig 6: Average Accuracy of different features set with different feature selection methods of DNA dataset

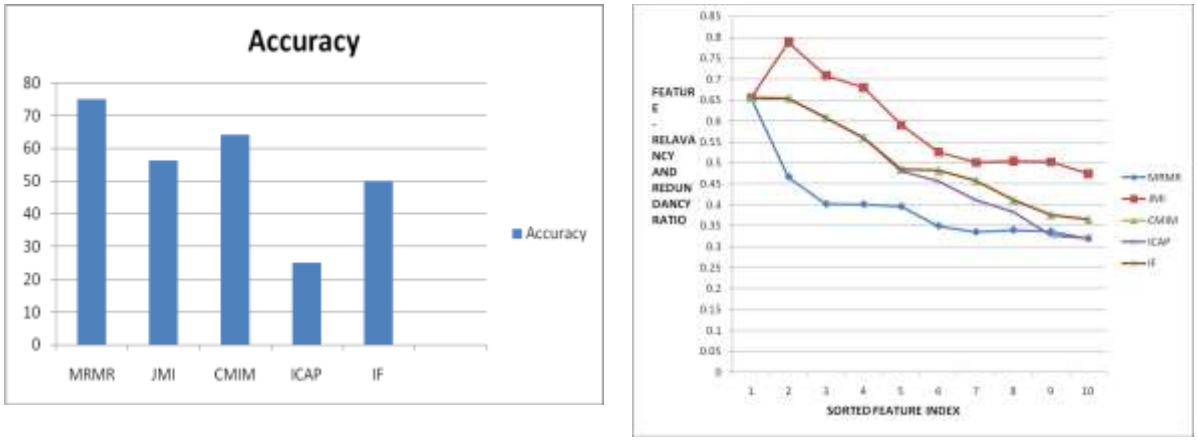

Fig 7: Accuracy chart for different feature with different feature selection methods and feature relevancy-reduction ration chart of New20 text dataset

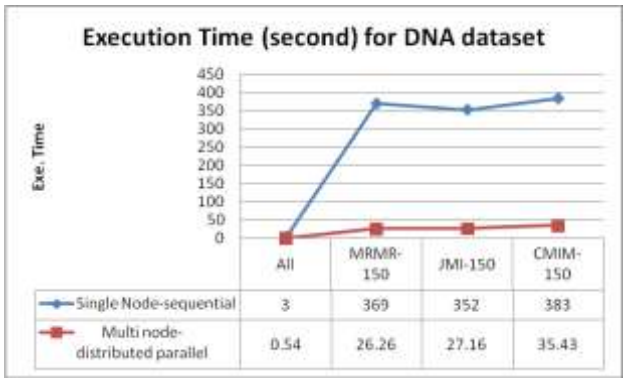

Fig 8 : Total Execution time comparison for sequential and distributed parallel data procession for DNA dataset

4. CONCLUSION :

As growing digital era, data is also growing in every moment. This big data is large in volume, complex and speedy arrival need analysis process in effective way. Big data have many issues and among them high dimensional data is one issue. Standard feature selection methods are not capable to handle them on distributed parallel environment. So, basic feature selection methods need to be modified or introduced new feature

selection methods for big data. In this paper, comparison of filters, wrappers and embedded methods are shown. Here, some basic feature selection methods like filters, wrappers, embedded and hybrid methods are discussed which are till now applied on big data. Filters are fast and good for big data analysis. Among all discussed methods, in this paper, some basic information theory based filter methods are tested on big data and prove better in terms of total execution time. In future, new feature selection methods for big data need to introduce which increase accuracy for large sampled datasets, high dimensional datasets, complex datasets and online stream datasets.

REFERENCES

[1] Chun-Wei Tsai,Chin-Feng Lai,Han-Chieh Chao, Athanasios V. Vasilakos, "Big data analytics: a survey" ,Journal of Big data, Springer, December 2015.

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

[3] 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 .

[4] 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.

[5] 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.

[6] S. Fong, R. P. Biuk-Aghai and Y. W. Si, "Lightweight Feature Selection Methods Based on Standardized Measure of Dispersion for Mining Big Data," 2016 IEEE International Conference on Computer and Information Technology (CIT), Nadi, 2016, pp. 553-559. doi: 10.1109/CIT.2016.120.

[7] 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.

[8] L. B. Shyamasundar and P. J. Rani, "Twitter sentiment analysis with different feature extractors and dimensionality reduction using supervised learning algorithms," 2016 IEEE Annual India Conference (INDICON), Bangalore, 2016, pp. 1-6.doi: 10.1109/INDICON.2016.7839075.

[9] M. Popescu and J. M. Keller, "Random projections fuzzy k-nearest neighbor(RPFKNN) for big data classification," 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Vancouver, BC, 2016, pp. 1813-1817. doi: 10.1109/FUZZ-IEEE.2016.7737910.

[10] 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.

[11] Daniel peralta, Sara del S., Ramirez Gallego, Isaac Triguero, Jose M.B., Francisco H., "Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach", Mathematical Problems in Engineering,Volume 2015 (2015), Hindawi, Article ID 246139.

[12] 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.

[13] 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.

[14] C. Cepeda, D. L. C. Tien and P. Ordóñez, "Feature Selection and Improving Classification Performance for Malware Detection," 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), Atlanta, GA, 2016, pp. 560-566. doi: 10.1109/BDCloud-SocialCom-SustainCom.2016.87.

[15] Claudio Reggiani, Yann-A¨el Le Borgne, and Gianluca Bontempi , "Feature selection in high-dimensional dataset using Mapreduce",https://arxiv.org, sep 2017.

[16] http://sci2s.ugr.es/BigData.

[17] Q. Liu, B. Ribeiro, A. H. Sung and D. Suryakumar, "Mining the Big Data: The Critical Feature Dimension Problem," 2014 IIAI 3rd International Conference on Advanced Applied Informatics, Kitakyushu, 2014, pp. 499-504.doi: 10.1109/IIAI-AAI.2014.105.

[18] T. Zhang and B. Yang, "Big Data Dimension Reduction Using PCA," 2016 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2016, pp. 152-157. doi: 10.1109/SmartCloud.2016.33.

[19] P. Dhumal and S. S. Deshmukh, "Retrieval and extraction of unique patterns from compressed text data using the SVD technique on Hadoop Apache MAHOUT framework," 2016 International Conference on Computing Communication Control and automation (ICCUBEA), Pune, 2016, pp. 1-5.doi: 10.1109/ICCUBEA.2016.7859996.

[20] Aldehim, Ghadah, Wang, Wenjia, "Determining appropriate approaches for using data in feature selection", International Journal of Machine Learning and Cybernetics, Springer 2017,Jun,volume 8,number 3,Pages: 915-928, ISSN:1868-808X,doi:10.1007/s13042-015-0469-8.

[22] https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/

[23] https://github.com/caesar0301/awesome-public-datasets.

[24] A. Jović, K. Brkić and N. Bogunović, "A review of feature selection methods with applications," 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2015, pp. 1200-1205.doi: 10.1109/MIPRO.2015.7160458.