Feature Selection, Online Feature Selection Techniques for Big Data Classification: - A Review

S.Gayathri Devi,
Department of Information Technology,
Coimbatore Institute of Engineering and Technology,
Coimbatore – 641 109, Tamilnadu, India.
gg.govind2007@gmail.com

M.Sabrigiriraj
Department of Electronics and Communication Engineering,
SVS College of Engineering,
Coimbatore – 642 109, Tamilnadu, India.
sabari_giriraj@yahoo.com

***Abstract*: In the recent times, several disciplines have to tackle with huge datasets, which are involved with a huge number of additional features. Feature Selection (FS) techniques target at reducing the noisy, redundant, or unnecessary features, which might degrade the performance of classification. Although there is several numbers of FS techniques, still it remains an active research field among the data mining, machine learning and pattern recognition groups. Several FS techniques are imposed with critical issues with regards to efficiency and usefulness, due to rise in data dimensionality, which happens nowadays. Nonetheless, conventional techniques are deficit of sufficient scalability to deal with datasets consisting of millions of instances and obtain results with success in a less amount of time. Therefore, in this case, an Online Feature Selection (OFS) algorithm can yield a better solution for solving this issue. This work reviews few of the available and well-known FS, OFS techniques by pointing out the pros and cons of those techniques. This technical work studies the details of traditional FS and OFS techniques depending on evolutionary computation that is helpful in getting the subsets of features from huge datasets. As a result, this review also provides a summary, and analysis of machine learning algorithms for huge datasets. In addition, the new machine learning strategies and methodologies are explained with their capacity of dealing with the different challenges with the ultimate goal of assisting the practitioners in selecting the suitable solutions for their use cases. This review work renders a view on the big data domain, finds the research gaps and possibilities, and offers a solid foundation, assistance for more research in the machine learning field that uses big dataset.**

Keywords - Big Data, Machine Learning Algorithms, Data Mining, Data Analysis, Feature Selection (FS) Techniques, Online Feature Selection (OFS) Techniques, and MapReduce (MR) paradigm.

I.INTRODUCTION

Learning from very big databases is a critical challenge for many of the present day's data mining and machine learning algorithms [1]. This issue is generally referred using the term "big data," that indicates the hardships and drawbacks of carrying out the processing and analysis of elaborate chunks of data [2–3]. It has drawn a good amount of interest in a large number of areas like bioinformatics, medicine, marketing, or financial businesses, due to the extensive sets of raw data, which are stored. The recent advancements on cloud computing technologies let the adaptation of standard data mining methods so that they can be successfully applied over enormous amounts of data [4]. Adapting data mining tools for big data challenges may require the redesign of the algorithms along with their incorporation in parallel environments. Simultaneously, the machine learning algorithm has easily occurs dimensionality of the problem. In order to resolve this issue, Feature Selection (FS) techniques can be utilized for reducing the dimensionality, prior to the application of any data mining methods like classification, association rules, clustering and regression.

The objective of FS is to decide on the feature subset that is as tiny as possible. It is one among the necessary pre-processing steps before the application of any data mining tasks. FS chooses the subset consisting of real features, with no loss of resourceful information. It eliminates unnecessary and repetitive features leading to the data dimensionality reduction. Consequently, it enhances the mining accuracy, decreases the computational time and improves result comprehensibility [5]. On applying feature selection step is used to the reduced feature subset produces the same result as with original high-dimensional dataset. FS provides the benefits like decreased storage needs, preventing over fitting, enabling data visualization, accelerating of the mining algorithms' speed of execution and limiting the training times [6]. After this, data mining methods are used over the dataset reduced for increasing the classification results [7].

FS techniques can be categorized into three groups: (i) Wrapper techniques: the selection criterion forms a part of the fitness function and hence is dependent on the learning algorithm [8] (ii) Filtering techniques: the selection is dependent on data-related measures, such as separability or crowding (iii) Embedded techniques: the optimal subset of features is constructed within the classifier generation [8]. To get more information regarding the particular feature selection techniques, the reader can refer to the surveys released with regard to the topic [7–8].

One fascinating study for using the feature selection for huge datasets is identified in [9]. In this work, the authors explained about an algorithm, which is capable of effectively coping with ultra high-dimensional datasets and choose a small subset consisting of exciting features from them. But, the number of chosen features is supposed to consist of a number of orders of magnitude, lesser compared to the total number of features and the algorithm is developed to be run in a single machine. Hence, this technique does not offer scalability for preferentially big datasets.

A specific means of dealing with feature selection is by making use of evolutionary algorithms [10]. Generally, the set of features gets encoded in the form of a binary vector, where every position decides whether a feature gets selected or not. This permits carrying out the feature selection with the investigation capabilities of evolutionary algorithms. But, they do not have the scalability required for dealing with huge datasets (starting with millions of instances). In spite of its significance of classification accuracy, many of the studies involving feature selection are limited to batch learning. Dissimilar to conventional batch learning techniques, online learning techniques stands for a potential family consisting of effective and scalable machine learning algorithms to be used for massive-scale applications. Many of the available studies of online learning need reaching out all the attributes/features of training instances. A classical setting such as this is not always suitable for practical-world applications if data instances are of high dimensionality or it is costly to get the entire set of attributes/features. In order to deal with this setback, the issue of Online Feature Selection (OFS) is examined, where an online learner is just permitted to maintain a classifier that involves just a small and fixed number of features. The organization of the rest of the section of the work is as below: Section 2 explains about the review carried out on the Literature, Section 3 studies about the Inference obtained from the Review Work. Section 4 explains about the solutions, Section 5 reviews the Results and discussion and at last, Section 6 studies about the conclusion and future work.

II.LITERATURE REVIEW

This section of literature review is structured as below. This section explains about the FS techniques, OFS techniques and big data classification techniques. Along with, the strengths and drawbacks of the above mentioned techniques are highlighted.

FEATURE SELECTION (FS) TECHNIQUES

Sun et al [11] developed a local learning based feature selection to be used for high dimensional data analysis. The key concept is to disintegrate a randomly complicated nonlinear problem into a group of locally linear ones by means of local learning, and thereafter perform the learning of feature relevance universally within the huge margin framework. The newly introduced algorithm is dependent on well-formed machine learning and numerical analysis methods, without assuming about the data distribution underneath. Theoretical analysis carried out on the sample complexity of the algorithm shows that the algorithm offers a logarithmical sample complexity corresponding to the number of features.

Fong et al [12] developed a suitable classification model for finding an accurate set of features from high dimensional data. A novel FS algorithm known as Swarm Search (SS) is designed for the identification of an optimal feature set by making use of meta-heuristics. As it possesses the flexibility to include any classifier to act as its fitness function, the swarm search is termed to be advantageous.

Fong et al [13] suggested a novel feature selection technique for attaining a classification model with good prediction accuracy. For having an optimal balance between generalization and over fitting, the technique of new and efficient feature Clustering Coefficient of Variation (CCV) is developed. The CCV search is presented for optimal subset of attributes taking the coefficient into consideration for improving the accuracy of classification. At last, Hyper-pipe i.e. fast discrimination technique is employed for examining the group which produces better accuracy with respect to classification.

Peralta et al [14] suggested a feature selection algorithm depending on evolutionary computation, which makes use of the MapReduce paradigm to get subsets of features from huge datasets. This algorithm divides the real dataset into blocks of instances with the purpose of learning from them in the map phase; and after this, the reduce phase integrates the partial results obtained into one final vector of feature weights. The FS algorithm is assessed by making use of three popular classifiers (Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB)) realized within the Spark framework for dealing with big data challenges.

Harde et al [15] proposed the Ant Colony Optimization Swarm Search (ACO-SS) based FS algorithm for Data Stream Mining of Big data. With the aim of solving these issues on the high dimensional and streaming structure in data feeds seen in big data, a novel light weight ACO based feature selection is introduced in this research work. The new feature selection processes will carry out the validation of the big data from their high level of dimensionality and streaming structure.

Selvi and Valarmathi [16] developed an enhanced firefly heuristics for effective FS. Firefly Algorithm (FA) is remarkable for local searches. But it may get pushed into local optima and hence cannot carry out global searches in a refined manner. In this, the classifications of the chosen features are conducted employing NB, K - Nearest Neighbor (KNN) and Multilayer Perceptron Neural Network (MLPNN) classifiers. Such a kind of firefly based FS presented here enhances the classification of the twitter data and also minimizes the time complexity. The result obtained from the experiments show that the new FA based FS method improves the effectiveness of classifiers.

Viegas et al [17] demonstrated a genetic programming strategy for feature selection in high dimensional skewed data. A method developed this manner targets at integrating the most distinguishing feature sets chosen by unique FS metrics. This is carried out so as to get a more efficient and impartial set consisting of the most distinguishing features, moving away from the hypothesis that unique FS metrics generate diverse feature space projections. The new solution designed not just raises the resourcefulness of the learning process by decreasing the data space up to 83% size, but also considerably increases its efficiency in conventional classifiers.

ONLINE FEATURE SELECTION (OFS) TECHNIQUES

An online Feature Selection (OFS) algorithm tries to solve the feature selection problem in an online manner by efficiently examining online learning methodologies. Particularly, the aim of OFS is to design online classifiers, which are involved with just a small and constant number of features for the purpose of classification. This section renders the background information regarding the OFS techniques and their disadvantages when addressing the problems of big data classification.

Zhou et al [18] explained about stream-wise feature selection, a class belonging to FS techniques. In this work, features are regarded to be in sequence, and either get added to the model or ignored. Two ordinary stream-wise regression algorithms, information-investing and alpha-investing are employed for producing simple and accurate model. It shows that stream-wise FS is hugely strong even with no early knowledge regarding the structure of the feature space.

Various algorithms have been introduced for online learning. Forget on [19] is termed the first online budget learning algorithm with the aim of guaranteeing over the number of mistakes. During every iteration, when the classifier does a mistake, it performs a three-step updating. Randomized Budget Perceptron (RBP) [20] eliminates a randomly chosen support vector if the number of support vectors goes beyond the predetermined budget. It accomplishes an identical mistake bound and empirical performances like the forgetron algorithm. One more strategy projectron [21], which is an online, perceptron like technique, is limited in space and time complexity. It follows a projection strategy to fix the number of support vectors. Particularly, in every iteration, if the training instance gets misclassified, it first builds a novel kernel classifier by using the updating rule of perceptron to the present classifier; then it projects the new classifier onto the space that is spanned by all of the support vectors with an exception being the new instance got. The classifier will stay unmodified, when the difference between the new classifier and its projection is not bigger compared to a threshold given. Empirical studies indicate that Project on generally performs better than forgetron in terms of classification but with a lengthier running duration. One important setback of projectron is that even though the number of support vectors of projectron is limited, however it has no clarity about the precise number of support vectors attained by projectron theoretically. Moreover, its high computational expense renders it unfavorable for big-scale applications.

Kivinen et al [22] used the kernel-based algorithms to be used for an online setting. As in this setting, data comes in a sequential manner, the examination offers solution for the three key issues seen in the online setting. 1) over fitting problem with standard online settings for linear techniques, 2) the functional representation of conventional kernel – based estimators and, 3) the training time taken of batch and incremental update algorithms.

Hoi et al [23] deals with the issue of OFS where the online learner is just permitted to maintain a classifier that involves a small and constant number of features. Sparsity regularization and truncation methodologies have been used in this research work. The empirical performance is analyzed here on both simple and massive-scale datasets. The results shows that the algorithms used are quite efficient to be used for feature selection tasks pertaining to online applications. It also indicates that they offer more efficiency and scalability compared to few of the batch learning feature selection methodologies.

Wu et al [24] suggested a new Online Streaming Feature Selection (OSFS) technique for selecting potentially relevant and non- repetitive features. A new framework dependent on FS has also been aimed at dealing with streaming features. This study indicates that applications involving streaming or an infinite size of features, a small number of features can be chosen for training a potential model, rather than making use of the important features.

Yu et al [25] provided an extension of a Scalable and Accurate Online Approach (SAOLA) for FS for large dataset. SAOLA uses new online pair-wise comparison methodologies and maintains a parsimonious model over time online. The two issues treated here include 1) high dimensionality that continues growing 2) feature selection to be greatly scalable, in an online fashion. SAOLA is remarkably stronger compared to the other state-of-the-art batch FS techniques in terms of precision, while being much faster in running time, and offers potential performance in comparison with the other state-of the-art FS methods.

Wang et al [26] deals with two diverse tasks of OFS: (1) learning with the entire input in which a learner is allowed to have access to all the features to determine the subset of active features, and (2) learning apart of the input where just a less number of features are permitted to have access to every instance by the learner. For both of these tasks, the algorithms have been advised to consider the issue of OFS for binary classification. Experimentation with Sparse online learning and assessment on big scale data has also been carried out. The results reveal that the algorithms are efficient for feature selection tasks pertaining to online applications, and are considerably more resourceful and scalable compared to batch feature selection method.

Anusuya et al [27] mentioned about an Apriori Particle Swarm Optimization (APSO), which is a distributed frequent sub graph mining method over MapReduce (MR). In this, the issues related to data with high dimensionality and the streaming behavior of incoming data has been dealt with. An incremental computation technique, with the capability of monitoring large- scale data in a dynamic fashion has also been explained making use of PSO in the form of a swarm search technique.

BIG DATA CLASSIFICATION METHODS

Even though, conventional machine learning approaches were designed in another era, and are dependent upon several assumptions, like the dataset fitting wholly into memory. But unluckily, these assumptions are no longer true in these new times. These damaged assumptions, along with the big data features, are generating problems if making use of the conventional methods. This section discusses different strategies for an effective big data classification.

Jun et al [28] developed a big data image classification method, which used Extreme Learning Machine (ELM) for the generation of positive and negative fuzzy rule. Here, in this work, the Extreme Learning Machine is exploited for pulling out the positive and negative fuzzy rule from the single-hidden layers. It reduces the computational speed and enhances accuracy via speed learning approach of Extreme Learning Machine method.

Huang et al [29] investigated different learning techniques with ELM. In this work, Proximal Support Vector Machine (PSVM) and Least Square Support Vector Machine (LS-SVM) with ELM have been studied and it is observed that ELM does the classification of any disjoint regions of big data along with small optimization limitations compared to the other two methods. It is proven from the analysis that ELM operates with more efficiency and outperforms PSVM and LS-SVM in terms of big data classification.

Chen et al [30] developed a novel technique for land cover classification known as ELM-MapReduce (ELM-MR). In this newly introduced work, ELMs are trained by mean so FMR framework in which the massive amount of data are provided as input to MR and it divides and then distributes the data to different nodes and locates an intermediate key value pairs; and these processes are conducted in parallel. It limits the computation and expense of classifiers. The results finally at the end of the prediction are combined together with voting strategy.

Tekin and Van der Schaar [31] presented a distributed online big data classification technique employing the context information. In this technique, the data is collected by means of distributed data sources and then processed with the help of a heterogeneous set consisting of distributed learners. In order to solve the issue of joint classification, due to the distributed and heterogeneous learners, where the data is received from different sources, it is considered to be a distributed contextual bandit problem and every data is treated by a particular context.

Grolinger et al [32] suggested methodology of big data classification with MR model. This study is focused on the issues that are involved in the realization of MR for the big data and presents the MR based classification model for overcoming the challenges. Such a kind of an approach enhances the parallelism seen in computing and decides on the classes for samples.

Rebentrost et al [33] advised on the usage of Quantum Support Vector Machine (QSVM) for having an effective classification of the big data. This technique makes use of SVM in the quantum computer for analyzing the complexity logarithmic in the size of the vectors and also the number of training examples. The resourceful evolutionary under sampling methodologies presented by

Triguero et al [34] has proven to be of great potential in big data classification. The issue of limited number of instances has to be resolved and therefore the MapReduce approach is proposed, which distributes the working of evolutionary under sampling algorithms into a cluster with computing elements.

Lopez et al [35] introduced linguistic fuzzy rule based classification system that is built on the MR framework for an effective classification of big data. This technique is represented as Chi-FRBCS-Big data yields a superior predictive accuracy in big data. Chi-FRBCS-Big data has been designed with two versions Chi-FRBCS-Big data-Max and Chi-FRBCS-Big data-Ave to render an effective data clustering for various degrees employing fuzzy rule based data classification.

Anbalagan and Chandrasekaran [36] proposed Parallel Weighted Decision Tree (PWDT) classifier for the classification of the complicated spatial landslide big data in the MR model. The technique is inclusive of the various degrees of significance to the diverse landslide factors. The three data structures like attribute table, count table, hash table are employed for building the parallel decision tree classifier. This enhances the classification by choosing a best splitting attribute that possesses the best gain ratio. Divya and Singh [37] mentioned the realization of artificial intelligence to enhance the big data classification. Based on the Pollination based Optimization, the big data classification is carried out here. This technique with a distinct mix of classification algorithm, parallelism and artificial intelligence, enhances the big data classification with efficiency.

Triguero et al [38] suggested a new distributed partitioning technique for prototype reduction methods in the closest neighbor classification. The technique efficiently minimizes the number of instances and therefore maximizes the speed of the classification that, in turn, immensely limits the storage needs and the noise sensitive measures. For the purpose of reducing prototypes in bigger datasets, the MR based framework is presented with techniques to merge several partial solutions into one single solution and there by prevents the slip in the rates of classification accuracy. This technique can be used for several applications that need conceptual work merging.

Horta et al [39] suggested the idea of active learning strategy by combining data stream based ELM and Hebbian learning for effectively classifying big data with the help of the linearization of the input data present in the ELM tree. The issue with this technique is that it is focused on multiple big data challenges such that the classification performance is not satisfying.

Raikwal et al [40] provided the analysis of machine learning algorithms such as Support Vector Machine (SVM) and Decision Trees (DTs) for assisting big data classification [40]. The analysis carried out on machine learning algorithms yields a clarity regarding the needs for having a huge chunk of data classified with efficiency. Tejasviram et al[41] developed a novel model known as Auto Associative Extreme Learning Machine (AAELM) with Multiple Linear Regression (MLR) (AAELM + MLR). It is the integration of AAELM and MLR. Here, in this newly introduced work, the input data gets trained with AAELM model. Afterwards, the hidden nodes from the AAELM generate few results. Depending on the input data it is regarded to be nonlinear principal components. Then the output of the hidden nodes of AAELM is provided as input to MLR to generate big data regression model.

Xin et al [42] introduced a novel approach for the issue seen in Extreme Learning Machine known as Elastic Extreme Learning Machine (E^2LM). The chief short coming of ELM technique is that more of matrix computation is necessary in the hidden layer of ELM. It deteriorates the performance of ELM with regard to computation expense and memory. This way, the newly introducedE^2LMmakes use of MapReduce framework for matrix multiplication computation. It generated an intermediate result of matrix multiplication and the old results are substituted by intermediate results. Depending on the results that are replaced, the final result are calculated with centralized computing.

INFERENCE FROM THE REVIEW WORK

Feature Selection has found extensive application in several domains like image processing, data mining and text mining, mainly for the issues that involve high-dimensional data. Feature selection over high-dimensional data has been commonly seen as an issue involved with the searching done for a minimal subset of features, which, in turn, results in the prediction model that is most accurate. Nonetheless, the scalability required for addressing big datasets is lacking (starting from millions of instances). In spite of its significance, many studies of FS are confined to batch learning. Dissimilar to conventional batch learning techniques, online learning indicates a potential family of effective and scalable machine learning algorithms to be used for large-scale applications. Many of the available studies of online learning need the access to all of the attributes/features of training instances. This online learning is proven to be a suitable technique for practical applications with the data instances that are being involved are of greater dimension.

SOLUTION

Online Feature Selection (OFS) is specifically significant and required if a practical application has to tackle with sequential training data with high dimensionality, like

online spam classification tasks, in which conventional batch learning FS strategies cannot be used directly. Even though there has been an interesting advancement by the already available OFS techniques, they still are imposed with the burden of scalability when the dimensionality is about the scale of millions or more. In accordance, this issue inspires to design a new scalable and online learning technique to cope up with dataset having a hugely high dimensionality in the form of future work.

II.RESULTS AND DISCUSSION

This section explains about the experiments that are performed and also their results. Various realizations of the MR programming model have shown up in the earlier years. The most prevalent one includes Apache Hadoop, which is an open-source framework written with Java that, in turn, lets the processing and management of big datasets in a distributed computing environment. In this work, performance of various feature selection techniques like OSFS, SAOLA, and APSO has been measured with diverse Hadoop based classifiers including Parallel Weighted Decision Tree (PWDT), Quantum Support Vector Machine (QSVM) and Auto Associative Extreme Learning Machine (AAELM) with Multiple Linear Regression (MLR) known as (AAELM + MLR).

Hadoop framework and Net Beans Integrated Development Environment (IDE) version 8 were utilized for the execution of the MapReduce framework. Moreover, this work employs the dataset used at the data mining competition of the Evolutionary Computation for Big Data and Big Learning" held on July 14, 2014, in Vancouver (Canada) that is considered to be ECBDL14 from http://cruncher.ncl .ac.uk/bdcomp/. This dataset contained 631 features (along with both numerical and categorical attributes), and it comprises of nearly 65 million instances. The primary characteristics of these datasets are tabulated in Table 1. For each dataset, the number of instances used for both training and test sets in addition to the number of attributes are indicated, also with the number of splits, which divides every dataset.

Table 1: Summary of the Big Datasets used for Classification

Dataset	Training instances	Test instances	Features	Instances per split
ECBDL14-ROS	65003913	2897917	631	~1984

With Table 2, different performance measures can be obtained. Few of the performance measures obtained from the Table 2 are generally employed in imbalanced learning. They include Precision, Recall, F-measure and Accuracy correspondingly.

Precision

Precision (*P*) is defined to be the fraction of retrieved instances, which have relevance and is expressed as

$$\text{Precision } (P) = TP/(TP+FP) \quad (1)$$

Recall

Recall (*R*) is defined as the ratio of relevant instances, which are retrieved.

$$\text{Recall } (R) = TP/(TP+FN) \quad (2)$$

F-measure

A measure combiningaccuracy and recall gives the harmonic mean of accuracy and recall, known as the conventional F-measure or balanced F-score:

$$\text{F-measure} = 2.(P.R)/(P+R) \quad (3)$$

Accuracy

The classification accuracy of an individual samples dependent on the number of samples rightly classified (true positives plus true negatives) and is assessed by the formula:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (4)$$

Table 2: Performance Measures Comparison of OFS Techniques Vs Classifiers

OFS Techniques	Classifiers	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
OSFS	PWDT	83.33	88.24	85.71	80.00
	AAELM + MLR	85.16	89.42	87.24	82.00
	QSVM	86.90	89.62	88.24	83.25
SAOLA	PWDT	89.66	89.90	89.78	85.24
	AAELM + MLR	90.41	91.29	90.85	86.74
	QSVM	91.44	91.38	91.41	87.49
APSO	PWDT	91.81	92.06	91.93	88.24
	AAELM + MLR	93.27	93.52	93.33	90.23
	QSVM	94.68	94.94	94.81	92.22

OFS techniques like OSFS, SAOLA, and APSO has been measured with various Hadoop based classifiers like

Parallel Weighted Decision Tree (PWDT), Quantum Support Vector Machine (QSVM) and Auto Associative Extreme Learning Machine (AAELM) with Multiple Linear Regression (MLR) known as (AAELM + MLR).

The figures 1- 4 below illustrates the results of performance comparison of precision, recall, f-measure and accuracy with three diverse classifiers like the PWDT, AAELM+MLR and QSVM classifiers that are measured with the help of three various OFS techniques including the OSFS, SAOLA and APSO. It can be concluded from the results that the QSVM with APSO yields superior results for all the metrics.

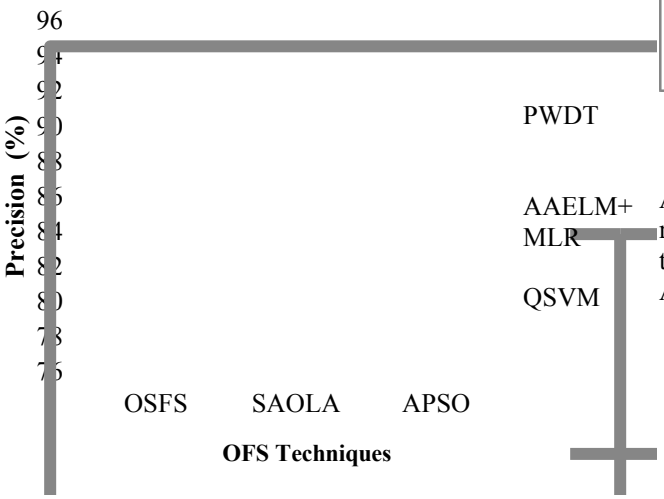

Figure 1: Precision results comparisons of OFS techniques Vs classifiers

As seen in figure 1, QSVM with APSO renders the precision results of 94.68%, while the other classifiers yields precision results of 93.27% and 91.81% for AAELM+MLR and PWDT techniques correspondingly.

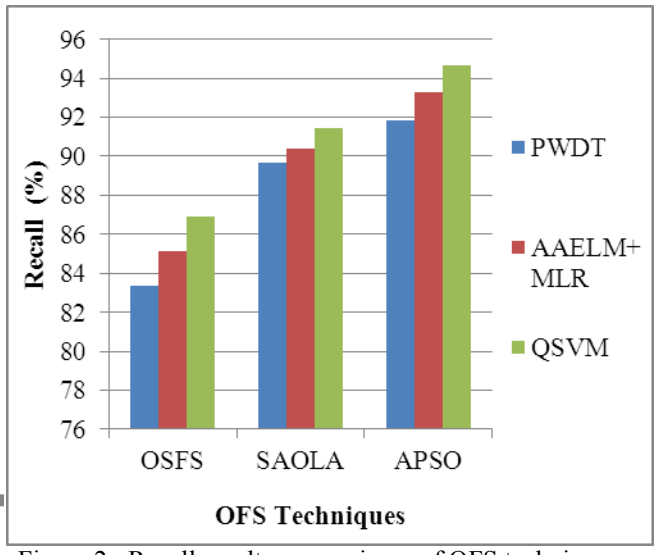

Figure 2: Recall results comparisons of OFS techniques Vs Classifiers

As observed in figure 2, QSVM with APSO yields the recall results of 94.94%, while other classifiers renders the recall results of 93.52% and 92.06% for AAELM+MLR and PWDT techniques correspondingly.

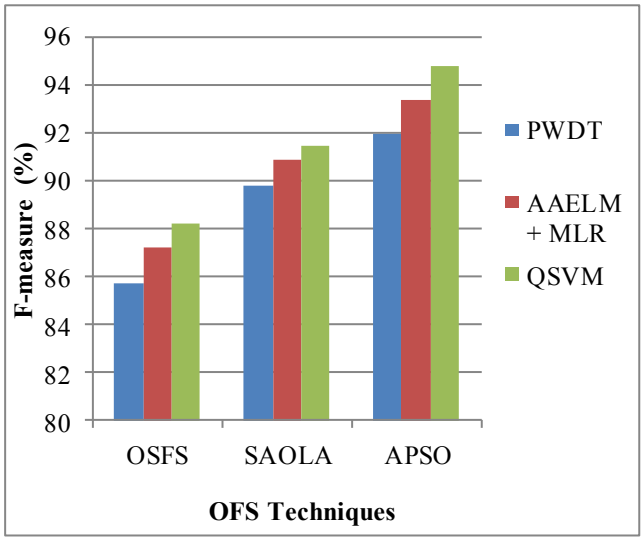

Figure 3: F-measure results comparisons of OFS techniques Vs Classifiers

As observed in figure 3, QSVM with APSO yields F-measure results of 94.81%, while other classifiers renders F-measure results of 93.33% and 88.24% for AAELM+MLR and PWDT methods correspondingly.

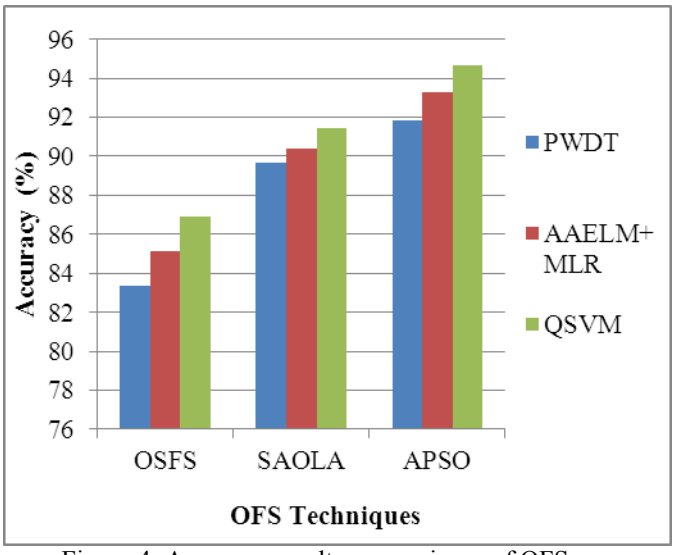

Figure 4: Accuracy results comparisons of OFS techniques Vs Classifiers

As seen in figure 4, QSVM with APSO renders the accuracy results of 92.22%, while other classifiers yields accuracy results of 90.23% and 88.24% for PWDT and AAELM+MLR techniques correspondingly.

III.CONCLUSION AND FUTURE WORK

This research work provides detailed review on Feature Selection (FS) and Online Feature Selection (OFS) techniques developed based on the MapReduce (MR) paradigm, focused on the preprocessing of big datasets such that they could be made accessible for other machine learning approaches, like SVM, ELM and DT that currently does not offer sufficient scalability to work with these datasets. The algorithm has been realized making use of Apache Hadoop, and it has been used over big datasets. Various FS and OFS techniques were explained and then compared. Their strengths and drawbacks were also explained. Additionally, this review also investigated different techniques, which include earlier knowledge from different algorithms that is again a means of maximizing the accuracy and minimizing the computational complexity of classifiers. In this work, the performance of the different OFS techniques like OSFS, SAOLA, and APSO has been measured with various Hadoop based classifiers like PWDT, QSVM and AAELM+ MLR. The theoretical assessment of the model focuses on the entire scalability of APSO with regard to the number of features in the dataset, when compared with a sequential technique. This behavior has been ensured further, once the empirical procedures are completed. Although there is an interesting advancement by the already available OFS techniques, still it tends to be hard to satisfy scalability if the dimensionality is of the scale of millions or more. In accordance, these aforementioned challenges inspire to present a new scalable and online processing technique to address dataset with high dimension as the future work.

REFERENCES

[1]. E. Alpaydin, Introduction to Machine Learning, MIT Press, Cambridge, Mass, USA, 2nd edition, 2010.

[2]. M. Minelli, M. Chambers, and A. Dhiraj, Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses (Wiley CIO), Wiley, 1st edition, 2013.

[3]. V. Marx, "The big challenges of big data," Nature, vol. 498, no. 7453, pp. 255–260, 2013.

[4]. J. Bacardit and X. Llorà, "Large-scale data mining using genetics-based machine learning," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, no. 1, pp. 37–61, 2013.

[5]. Liu, H. and Motoda, H., 2012. Feature selection for knowledge discovery and data mining (Vol. 454). Springer Science & Business Media.

[6]. Zeng, Z., Zhang, H., Zhang, R. and Zhang, Y., 2014. A hybrid feature selection method based on rough conditional mutual information and naive Bayesian Classifier. ISRN Applied Mathematics, 2014, pp.1-11.

[7]. H. Liu and H. Motoda, Computational Methods of Feature Selection, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC Press, 2007.

[8]. Saeys, Y., Inza, I. and Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), pp.2507-2517.

[9]. M. Tan, I. W. Tsang, and L. Wang, "Towards ultrahigh dimensional feature selection for big data," Journal of Machine Learning Research, vol. 15, pp. 1371–1429, 2014.

[10]. J. S. Sánchez, "High training set size reduction by space partitioning and prototype abstraction," Pattern Recognition, vol. 37, no. 7, pp. 1561–1564, 2004.

[11]. Sun, Y., Todorovic, S. and Goodison, S., 2010. Local-learning-based feature selection for high-dimensional data analysis. IEEE transactions on pattern analysis and machine intelligence, 32(9), pp.1610-1626.

[12]. S. Fong, X. S. Yang, and S. Deb, Swarm search for feature selection in classification, in Proc. 2nd Int. Conf. Big Data Sci. Eng.,Dec. 2013, pp. 902-909.

[13]. S. Fong, J. Liang, R. Wong, and M. Ghanavati, "A novel feature selection by clustering coefficients of variations," in Proc. 9th Int. Conf. Digital Inf. Manag., Sep. 29, 2014, pp. 205–213.

[14]. Peralta, D., del Río, S., Ramírez-Gallego, S., Triguero, I., Benitez, J.M. and Herrera, F., 2015. Evolutionary feature selection for big data classification: A mapreduce approach. Mathematical Problems in Engineering, 2015.

[15]. Harde, S. and Sahare, V.,, 2015. "ACO Swarm Search Feature Selection for Data stream Mining in Big Data", International Journal of Innovative Research in Computer and Communication Engineering, 3(12), pp.12087-89.

[16]. Selvi, S. and Valarmathi, M.L., 2017. An improved firefly heuristics for efficient feature selection and its application in big data. Biomedical Research, pp.s236-41.

[17]. Viegas, F., Rocha, L., Gonçalves, M., Mourão, F., Sá, G., Salles, T., Andrade, G. and Sandin, I., 2018. A Genetic Programming approach for feature selection in highly dimensional skewed data. Neuro computing, 273, pp.554-569.

[18]. Zhou, J., Foster, D.P., Stine, R.A., and Ungar, L.H. (2006). Streamwise feature selection. Journal of Machine Learning Research, pp.1861-1885.

[19]. Dekel, O., Shalev-Shwartz, S., and Singer, Y. (2008). The forgetron: A kernel-based perceptron on a budget. SIAM Journal on Computing, Vol. 37, No. 5, pp. 1342-1372.

[20]. Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2007).Tracking the best hyperplane with a simple budget perceptron. Machine Learning,Vol. 69, No. 2, pp. 143-167.

[21]. Orabona, F., Keshet, J., and Caputo, B. (2008).The projectron: a bounded kernel-based perceptron. In Proceedings of the 25th international conference on Machine learning, pp. 720-727.

[22]. Kivinen, J., Smola, A.J., and Williamson, R.C.(2010). Online learning with kernels. IEEE Transactions on Signal Processing, Vol. 100, No. 10, pp.1-12.

[23]. Hoi, S.C., Wang, J., Zhao, P. and Jin, R., 2012, Online feature selection for mining big data. In Proceedings of the 1st international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications, pp. 93-100.

[24]. Wu, X., Yu, K., Ding, W., Wang, H., and Zhu, X. (2013). Online feature selection with streaming features. IEEE transactions on pattern analysis and machine intelligence, Vol. 35, No. 5, pp. 1178-1192.

[25]. Yu, K., Wu, X., Ding, W., and Pei, J. (2014).Towards scalable and accurate online feature selection for big data. In International Conference on Data Mining (ICDM), pp. 660-669.

[26]. Wang, J., Zhao, P., Hoi, S. C., and Jin, R. (2014). Online feature selection and its applications. IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 3, pp. 698-710.

[27]. Anusuya, D., Senthilkumar, R.,andSenthilPrakash, T.(2017). Evolutionary Feature Selection for big data processing using Map reduce and APSO. International Journal of Computational Research and Development (IJCRD), Vol. 1, No. 2, pp. 30-35.

[28]. Jun W, Shitong W, & Chung F, Positive and negative fuzzy rule system, extreme learning machine and image classification, International Journal of MACH LEARN and CYB, 2 (2011) 261-271.

[29]. Huang G B, Zhou H, Ding X, & Zhang R, Extreme learning machine for regression and multiclass classification, IEEE T SYST MAN CY B, 42 (2012) 513-529

[30]. Chen J, Zheng G, & Chen H, ELM-MapReduce: MapReduce accelerated extreme learning machine for big spatial data analysis, IEEE IntConf on Control and AUTOM, (2013) 400-405.

[31]. Tekin, C. and van der Schaar, M., "Distributed online big data classification using context information." In IEEE 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1435-1442, 2013.

[32]. Grolinger, K., Hayes, M., Higashino, W.A., L'Heureux, A., Allison, D.S. and Capretz, M.A., 2014, Challenges for mapreduce in big data. IEEE World Congress on Services (SERVICES), pp. 182-189

[33]. Rebentrost, P., Mohseni, M. and Lloyd, S., 2014. Quantum support vector machine for big data classification. Physical review letters, 113(13), p.130503.

[34]. Triguero I, M. Galar, S. Vluymans, C. Cornelis, H. Bustince, F. Herrera, Y. Saeys. "Evolutionary Under sampling for Imbalanced Big Data Classification." IEEE Congress on Evolutionary Computation (CEC), pp. 715-722, 2015.

[35]. Lopez, V., del Rio, S., Benitez, J.M. and Herrera, F., "On the use of MapReduce to build linguistic fuzzy rule based classification systems for big data." In IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1905-1912, 2014.

[36]. Anbalagan P, and Chandrasekaran R.M. "A Parallel Weighted Decision Tree Classifier for Complex Spatial Landslide Analysis: Big Data Computation Approach." International Journal of Computer Applications, vol 124, no. 2, pp.5-9, 2015.

[37]. Divya, A.J. and Singh, G., "Classification of Big Data through Artificial Intelligence." International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 8, pp.17 – 25, 2015

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

[39]. Horta, E.G., Castro, C.L.D. and Braga, A.P., "Stream-Based Extreme Learning Machine Approach for Big Data Problems." Mathematical Problems in Engineering, vol.2015, pp.1-17, 2015.

[40]. Raikwal, J. S., &Saxena, K. (2014). Weight based Classification Algorithm for Medical Data. International Journal of Computer Applications, 107(21).

[41]. Tejasviram V, Solanki H, Ravi V, &Kamaruddin S, Auto associative extreme learning machine based non-linear principal component regression for big data applications, Digital Information Management, Tenth IntConf on (2015) 223-228.

[42]. Xin J, Wang Z, Qu L, & Wang G, Elastic extreme learning machine for big data classification, Neuro computing, 149 (2015) 464-471.