

Large Dimensional Data Reduction by Various Feature Selection Techniques: A Short Review

Bharti Swarnkar
Dept. of Electrical Engineering
Kalinga University
Raipur, Chhattisgarh, India

Prateek Pratyasha
Dept. of Electrical Engineering
Veer Surendra Sai University of
Technology
Sambalpur, Odisha, India

Aditya Prasad Padhy
Dept. of Electrical Engineering
Kalinga University
Raipur, Chhattisgarh, India

ABSTRACT

In a data dependent era, dimensions contain a huge number of variables both in rows and columns forming more complex data matrices and these dimensional expansions generate a large dimensional data (LDD). The dimensionality problem of LDD is a massive challenge for analytics purpose and it somehow burdens the machine learning approaches. Due to the faster growing rate in innovative Internet of Things and web-based technologies, static data becomes noisy and non-stochastic that results in data loss and instability. Therefore, the demand for complex data dimension reduction technique (DDR) is growing immensely to improve data prediction, analysis and visualization. Several computational techniques have implemented for DDR which is further segregated into two categories such as feature extraction techniques (FET) and feature selection techniques (FST). But, most of the existing FET methods focus on transforming the higher dimensional data into a lower dimensional space and unable to tackle with the dimensionality reduction problem. Hence, this paper focuses on various dynamic FST that not only reduces the dimensionality load but also catalyze the data analysis process.

General Terms

Machine Learning, Feature Extraction, Feature Selection, Curse of Dimensionality, and Data dimension reduction.

Keywords

Large Dimensional Data (LDD), Dimension Reduction (DDR) Techniques, Feature Extraction Techniques (FET), Feature Selection Techniques (FST).

1. INTRODUCTION

A dataset implies a statistical data matrix with rows as domains or subjects and columns as variables. One measurable data subset taken from the input data is known as features. Recently a term called “large dimensional data (LDD)” has widespread in data science and other related research fields [1, 2]. LDD is the subset of big data which carries huge volumetric database matrices that informs about data analysis and visualization. These data are useful in storing recording patient related information and that number is growing progressively in machine learning, healthcare sector etc. [3]. Its application is widespread in the domain of data interpretation, data management, data analysis and visualization. But, the storage, processing and maintenance of such amass features of LDD need a lots of memory space

which may be the reason for data loss. The problems related with data loss, data privacy and security seem are a few unavoidable issues which lead to the state-of-art of “Curse of Dimensionality”. This phenomenon may increase the space volume of data vulnerably that tends to form a sparse data. In those cases, it is extremely necessary to preserve the necessary information of those massive dataset by DDR techniques. These techniques focus at converting the data from higher dimensional space into a lower dimensional space for better data analysis retaining most of its redundant attributes in original form [4].

To evaluate the discussions statistically, DDRs are further classified into feature extraction technique (FET) and feature selection technique (FST). The most former attribute for dimensional reduction technique is feature extraction. It transforms large number of features into compact dimension and makes it a new additive one to analyze a large static dataset [5]. However, FETs lack in dealing with data complexity, so we are not mentioning those details in this paper. On the other hand, the objective of any FST is to select the best optimal feature set by eradicating the irrelevant features from the original dataset without disturbing the originality of features.

The objectives of feature selection can be executed insight when we focus on model accuracy and for that feature redundancy and feature relevancy are two major factors in improving classification by eradicating non-relevant features from the original dataset.

Various feature selection algorithms takes a major role when:

- (a) Over-fitting problems are needed to get solved with a better performance model,
- (b) A model has to be built which is cost-effective and has a faster computational rate,
- (c) The generated data has to be analyzed in deep.
- (d) The machine algorithm performance need support for big data analytics

2. FRAMEWORK

FST depends on two basic approaches: subset evaluation and individual evaluation. Subset evaluation deals with feature relevancy and redundancy based on a search strategy approach.

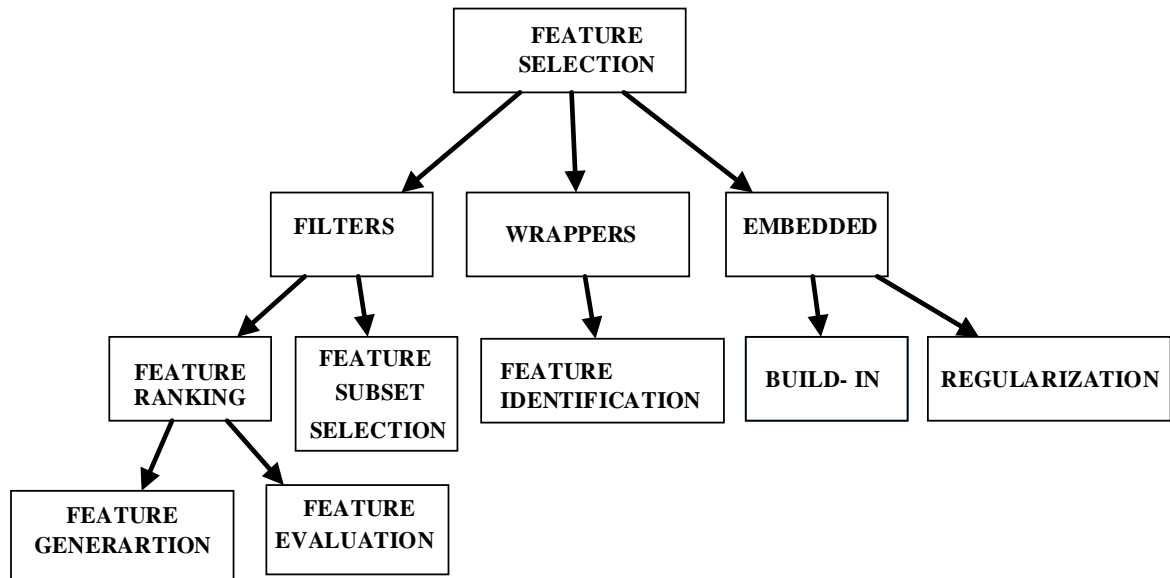


Fig. 1 Taxonomy of FST

Some feature candidates are generated from the original feature subset. One of the feature candidates is evaluated by certain computational measurement and the measurement value is compared with previously evaluated feature. On the other hand, individual evaluation assigns weights to the features based on their relevancy ranking priority in a way that the top ranked feature is assigned with first preference for evaluation. If we take care of the redundancy implicitly in accordance with relevant feature, then most of the optimal subset may be approximated [6-8].

Although these heuristic methods are employed to approximate the computational time by searching for combined redundant and relevant features, but this exhaustive search strategy fails when we deal with a bulk quantity of features. It has been observed that the optimal feature selection is based on the data population with recognized data distribution. In higher-dimensional coordinates, the training data is a part of data population and over-searching of redundant features may cause over-fitting problem for training data. Thus, we have to find some alternate methods to get rid of this problem as well as beneficial in DDR for optimal feature subset selection [9-11].

On the basis of various methods of searching, FSTs are further branched as filter, wrapper and embedded methods as shown in Table 1. The most discriminative features are opted from the character of data by filter methods. Generally, filter methods perform feature selection before classification and

usually involves a two-step process. In the first step, all the features are ranked according to a certain criteria. In the next

step, the features with the highest ranking are selected. Wrapper methods use the intended learning algorithm to select a feature subset as search problem where various problems are prepared, evaluated and compared with other combinations. Wrappers are quite faster than filters as they use model hypothesis by taking the training data in focus. Embedded methods perform feature selection in the process of model construction [12-15].

Table 1. Attributes of FST

Types of FST	Attributes	
	Classifier Dependency	Computational cost
Filter	Independent	Low
Embedded	Dependent	Low
Wrapper	Dependent	High

From fig 1, we observe that, to distinguish various FST, various feature ranking methods are taken into account. The obtained results from feature ranking are highly dependent on individual features. The results of feature subset selection are also applicable to the target class. The results of embedded methods are prognostic models that are configured by feature subsets. However, the prior objective of FST is to opt a optimal feature subset, but filter and wrapper methods may produce an uncertain specific model associated with the feature subset. On the other hand, the objective of embedded methods is to produce a predictive model. Features remain in the model are the byproduct of the modeling process [16]. The results are not generally optimal but for calculation purpose, it is computationally efficient. It involves algorithms such as HGAPSO, ReliefF, Minimum Redundancy Maximum Relevance (MRMR), Recursive Feature Elimination (RFE) and Simultaneous Perturbation Stochastic Approximation (SPSA).

2.1 Hybridised Genetic Algorithm and Particle Swarm Optimization (HGAPSO)

Various schematic FST demand for more training samples to evaluate the dataset accurately. In aid to that, the exhausted CPU processing time and computational time of samples is a prolonged issue for finding the “best” optimal features. To overcome this barrier, a

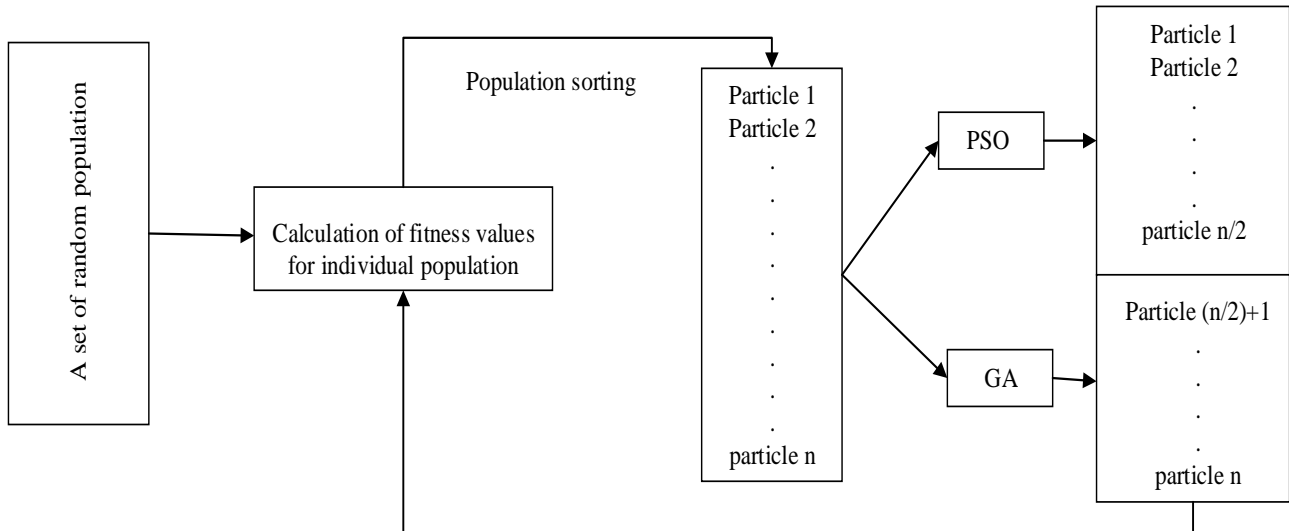


Fig 2. Flowchart of HGAPSO

faster FST has been introduced known as HGAPSO with the concepts of two most popular optimization algorithms such as GA and PSO [17-19]

GA can be incorporated with PSO by integrating the steps such as GA based selection, cross-over and mutation with an updated position and velocity of particles from PSO. Hence, HGAPSO is an integrated techniques used to select the features optimally from a set of population in which the number of features must be equal with total dimension of particles. In this technique, the dimension of position is equal with dimension of velocity of each particle as:

$$v_i^k = x_i^k \quad (1)$$

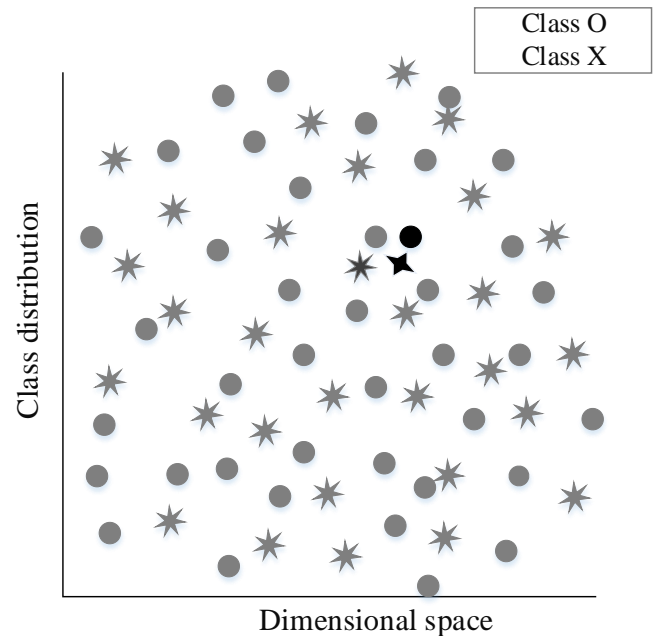
where x_i^k and v_i^k is the dimension of position and velocity of a particle respectively. The positions are taken in terms of binary values of 0's and 1's representing absence and presence of features respectively. Fig.2, Shows the flowchart of HGAPSO for feature selection. In Fig 3, a set of random population is rendered and the fitness values for each parameter of the population are carried out by applying optimization algorithm. Then, by sorting method, the features have ordered in descending order. After that, combinations of GA and PSO have been implemented where the individual population is carried out as chromosomes for GA and as particles for PSO. Then the whole set is split into two different subsets, each retaining its own attributes [20].

2.2 ReliefF

It is a unique family of filter type feature selection algorithm that maintains a satisfactory balance between the objectives and computational accuracy. However, there are some shortcomings of this algorithm such as frequency sampling is uncertain and there is a fluctuation in feature weight instances. Hence, some modification in reliefF algorithm is needed to resolve the accuracy related problem on the basis of mean variance model [21].

The input to find the optimal features is a feature weight vector $F_w(A)$ for class C and for other attribute values. The output is aimed at determining the qualities of attributes for the input raw signal. To achieve this output some steps are followed. In step 1, a weight F_w of sample data matrix is fixed at 0. Then, the process begins for $i = 1, \dots, n$. An

instance r_i is randomly chosen and class (r_i) is computed for each class C . In the next step, the hits h_j points and the misses point $n_j(c)$ of k nearest points are determined for each class C . Again, the weight vector is computed as, $F_w(A) = (a_i - h_j)^2 + (a_i - n_j(c))^2$ (2)



Instances	Attributes
★	Target instance for class O
●	Nearest neighbor instances for class O
★	Nearest neighbor instances for class X

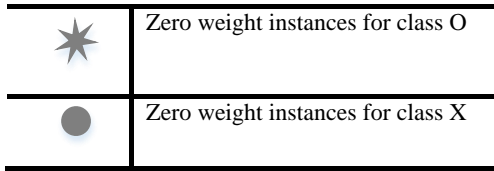


Fig 3. Optimal Feature selection by ReliefF

Figure 3 gives a graphical representation for two different classes such as class O and class X showing how features are selected optimally. ReliefF algorithm is not bounded with feature selection of a single class; it also deals with multi-classes. Hence, the state-of-art of reliefF algorithm is more effective than any other wrapper methods by calculating the feature weight vector and also figures out convex optimization problems. In the meanwhile, the selection of instances occurs randomly, due to which the sampling frequency becomes uncertain and hence this algorithm reduces the accuracy [22].

2.3 Minimum Redundancy Maximum Relevance (MRMR)

Mutual information is the most powerful tool in finding and selecting relevant features from a feature space. Often, some of these subsets contain some characteristics of features which may be redundant. Since our aim is to select the most relevant feature subset with a limited size factor, a new approach such as MRMR selection algorithm has been developed to eradicate this problem. The selections of optimal features involve 2 steps: (a) Measure the dependency between the features to be selected and the target dataset by mutual information. (b) Maximize the dependency. Since, the calculation of dependency is seemed to a tough job; hence we choose to find out the dependency by MRMR criteria as per the formula:

Dependency = relevance - redundancy

$$D = I_M(a_j; c) - \frac{1}{n-1} \sum_{a_i \in S_{n-1}} I_M(a_j; a_i) \quad (3)$$

where a_i is the total number of feature subset and a_j is the features chosen for consideration under the set $A - S_{n-1} \cdot S_{n-1}$ is the already selected feature subset. Let S_i represents the feature set to be selected and δ be the classification target. Redundancy between S_{n-1} and S_i can be represented as;

$$R_e = \frac{1}{n} \sum_{a_i \in S_{n-1}} I_M(a_i, a_i) \quad (4)$$

The proper defined MRMR function can be obtained by combining equation (3) and (4) as:

$$\max_{a_j \in S_{n-1}} \left[I_M(a_j, \delta) - \frac{1}{n} \sum_{a_i \in S_{n-1}} I_M(a_j, a_i) \right] \quad (5)$$

Equation (5) justifies the maximum correlation with minimum redundancy among the a_j features in S_i . In this algorithm, relevance is computed by using F-static values and mutual information for continuous and discrete features respectively whereas, redundancy is computed by applying Pearson and mutual information [23, 24].

2.4 Recursive Feature Elimination(RFE)

REF is a technique which contains the weights of the feature vectors as an important attribute for optimal feature subset prediction.. It must be noted that the features which can predict the best result even though some of the feature cannot predict well in target data. When we deal with noisy data, algorithm can be useful in extracting sufficient information to respective weight values. The iteration process continues till the prediction errors are settled to a minimum value or a specified feature values are eliminated. In [25], A proposed RFE conjoint with Support Vector Machine (SVM) was studied for the analysis of a large dimensional dataset. An RFE based SVM can be represented as a backward selection algorithm to solve the basic optimization problem [26].

2.5 Simultaneous-Perturbation Stochastic Approximation(SPSA)

Often it is observed that gradient effects are incapable to compute the selected features because the objective functions are not optimized properly. So, to avoid this problem, a gradient approximate method is used known as SPSA. This method is proved to be effective as it needs only two objective functions for stochastic gradient approximation [27]. The detail algorithm of SPSA can be studied from figure 8 which depicts the random selection of a constant finite coefficient α to obtain two finite constants such as α_0 and α_1 . After simulating both the constants, the satisfying criterions have also been calculated. Then, the optimality condition has been checked. In case of non-feasible solutions, the gradient has been searched and values of α has been updated for each iterations [28].

Let us consider an objective function $S: Z^d \rightarrow Z$ such that v^* is an optimum of S that fulfill the characteristics of gradient space which

disappears at point. To compute the gradient effectively, SPSA method is introduced which incorporates within which $\hat{g}(v^*)$ is considered to be an approximate for true value of $\hat{g}(v^*)$. objective function performs the gradient descent in a stepwise manner such as:

$$v_{m+1} = v_m + d_m \hat{g}_m(v_m) \quad (6)$$

Where $d_m \in Z$ is the size of the optimization step for each m iteration.

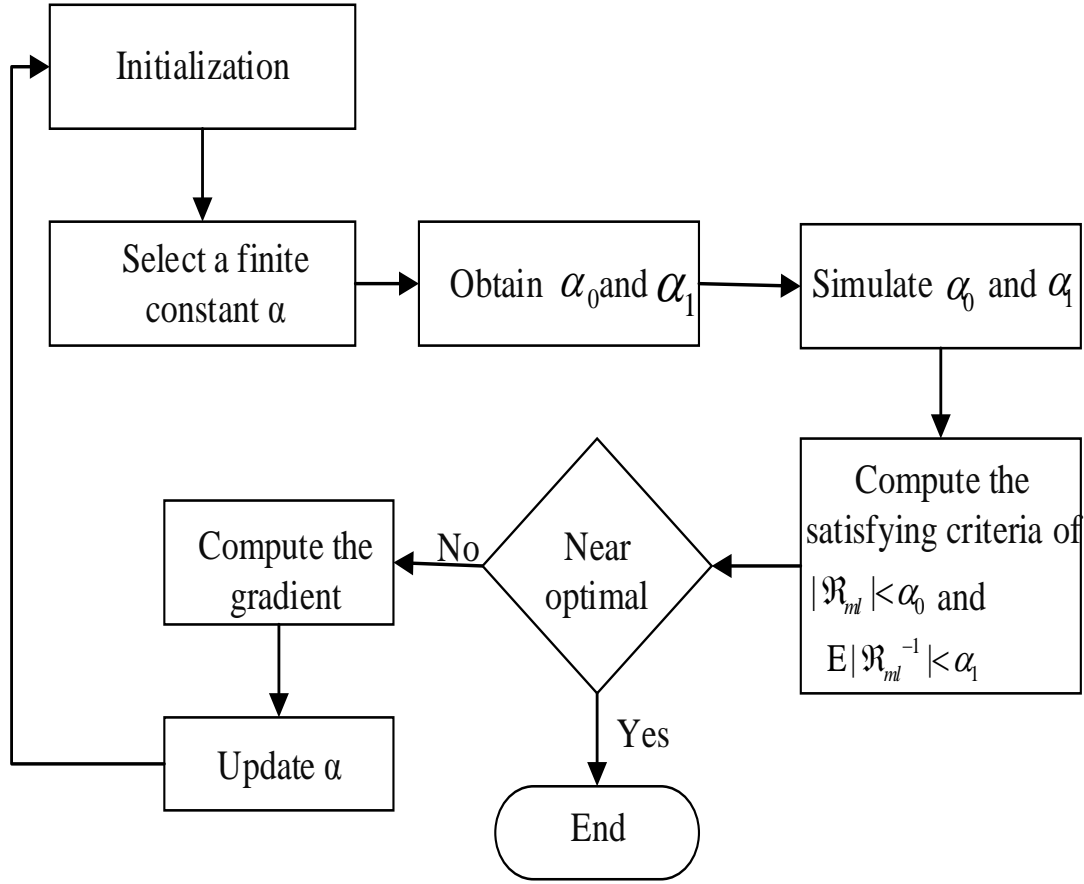


Fig 4. Flowchart of SPSA

Let us consider a simultaneous perturbation vector $\mathfrak{R}_m \in Z^d$. Hence, the gradient approximation is calculated for the given \mathfrak{R}_m as:

$$\hat{g}_m(v_m) = \begin{bmatrix} \frac{1}{\mathfrak{R}_{m1}} \\ \frac{1}{\mathfrak{R}_{m2}} \\ \vdots \\ \frac{1}{\mathfrak{R}_{md}} \end{bmatrix} \cdot \frac{S(v_m + C_m \mathfrak{R}_m) - S(v_m - C_m \mathfrak{R}_m)}{2C_m} \quad (7)$$

where $\mathfrak{R}_m = \{\mathfrak{R}_{m1}, \mathfrak{R}_{m2}, \dots, \mathfrak{R}_{md}\}$ is a vector of d and the elements are mutually independent mean-zero random variables. The satisfying criteria for two finite constants α_0 and α_1 are $|\mathfrak{R}_{ml}| < \alpha_0$ and $E|\mathfrak{R}_{ml}^{-1}| < \alpha_1$ respectively. When we deal with extreme non-linear and noisy data, multiple numbers of SPSA gradients can be calculated and their mean can be accomplished for each iteration. The detail algorithm of SPSA can be studied from figure 8 which depicts the random selection of a constant finite coefficient α to obtain two finite constants such as α_0 and α_1 . After simulating both the constants, the satisfying criterions have also been calculated. Then, the optimality condition has been checked. In case of non-feasible

solutions, the gradient has been searched and values of α has been updated for each iterations.

3. COMPARISON ANALYSIS

A comparison analysis has the prospect of generalizing valuable understandings between various DDRs based on their qualitative merits and some of the demerit key factors. For a brief observance at a glance, advantages and disadvantages of all the DDRs are mentioned in table 2 and it can be noticed that transformation of original features into a new feature subset may create the data loss and data complexity problems. To avoid this, FSAs are implemented which makes the training model less complicated for further processing. Hence, the state-of-art of FSA is helpful not only for resolving the overfitting issues but also insights new approaches to overcome the ‘‘Curse of dimensionality’’. Hence, the state-of-art of FSA is helpful not only for are compared with several FSAs such as HGAPSO, ReliefF, MRMR, REF and SPSA as illustrated in table 2.

Table 2. Difference between FEA and FSA

Methods	Advantages	Disadvantages
FEA	<ul style="list-style-type: none"> • Makes the domain of signal processing and image processing easier. • Quantifies the behavior of the features. • Discriminative power is more compared to any other data dimensional techniques. • Enhance the effectiveness of ML in supervised data dimensional reduction. 	<ul style="list-style-type: none"> • Feature transformation is expensive • There is a chance of data loss during new feature formation
FSA	<ul style="list-style-type: none"> • Computation time to train the data model is faster. • It simplifies the model and makes it interpretable. • The model accuracy is enhanced by choosing optimal features • It can overcome the overfitting problem 	<ul style="list-style-type: none"> • Optimal feature may not be feasible. • Selection of individual parameter is hard.

Table 3 Advantages and Disadvantages of different FSAs

Techniques used	Advantages	Disadvantages
HGAPSO	<ul style="list-style-type: none"> • Efficient selection of optimal features from a large dataset • Implemented at its best where PCA cannot be used • Finds probability of solution in less time 	<ul style="list-style-type: none"> • Complexity in implementation • Choosing the appropriate parameters is hard • Although the solution is metaheuristic, there is no guarantee of its accuracy.
ReliefF	<ul style="list-style-type: none"> • Adaptive to all types of data • Better computational efficiency • Scalable in many iterative approaches 	<ul style="list-style-type: none"> • Computational time is very expensive • Fail to remove some of the redundant features
MRMR	<ul style="list-style-type: none"> • Provide more accurate solution • Faster algorithm • Maximize the relevance of gene 	<ul style="list-style-type: none"> • Highly sensitive for parametric measurement
	<ul style="list-style-type: none"> • 	<ul style="list-style-type: none"> • Sensitive to feature scaling • Regularization of parameters for every iteration requires many hyperparameters. • Some objective functions are non-smooth to be optimized.

4. VALIDATION FOR HIGH DIMENSIONAL DATA ANALYSIS

In the research area of signal processing, LDD analysis is predominating more than two decades. To evaluate the complexity of the dimension of high dimensional data, we have considered some medical dataset measured from monitoring device and telemetry.

4.1 Description of Dataset

The validation of proposed feature reduction algorithms can be verified by taking some real time dataset into account. Hence, three open source datasets have considered including two microarray datasets (CNS Tumors [29], Leukemia [30] and colon tumor [31]). Table 2 summarizes the dataset with respective feature subset used for dimensional reduction algorithms.

4.2 Experimental Setup

To perform various feature extraction algorithms, the datasets have segregated into training and testing data subset. The training and testing data have not merged to maintain the

individual characteristics of each feature. The test model is formed from 70% of the training mode. For some feature whose attributes is not prominent, we take the nominal values which normalized within the range of [0, 1].

Most of the index features have removed. Then, the complex dimensional high dimensional data has been reduced into a lower dimensional matrix by the transformation original feature set into a extracted feature subset by using five proposed FSA such as HGAPSO, ReliefF, MRMR, RFE, and SPSA have employed.

4.3 Experimental Outcome

Three different high dimensional dataset including microarray data have taken in order to reduce their dimensional size. For that, their necessary feature attributes must be identified and reduced to a limited size. From Table 3, it is shown that MRMR method is proven to be the best among all the FST in selecting optimal features for the dimension reduction of high dimensional datasets. MRMR method is the fastest algorithm to compute the training data and finding the optimal features at the same time.

Table 3. Number of features selected with different FSTs

Datasets	HGAPSO	ReliefF	MRMR	RFE	SPSA
CNS Tumors	32	30	27	29	28
Leukemia	45	44	42	43	45
Colon tumor	11	8	7	12	10

Table 4. Accuracy (in %) of selected features with different FSTs

Datasets	Accuracy in (%)				
	HGAPSO	ReliefF	MRMR	RFE	SPSA
CNS Tumor	85.47	88.53	94.19	89.56	84.97
Leukemia	86.30	85.32	96.53	83.28	87.01
Colon tumor	88.54	89.24	92.54	84.39	86.67
Average	86.77	87.69	94.42	85.74	86.21

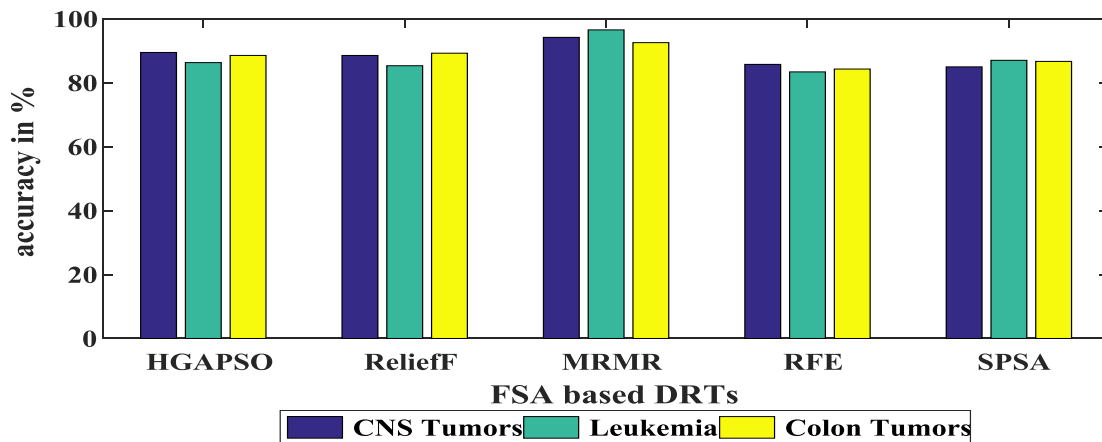


Fig 4. A comparison performance analysis of different FST

From table 4, it can be clearly observed that MRMR bases FST plays a vital role in the reduction of high dimensional data relevantly than the feature selection process individually. No doubt those FSTs are also necessary in preprocessing of data but when dimension reduction of data becomes the major issue of data analysis, it is better to choose a novel FST for two reasons. First reason is that MRMR can reduce the higher dimension of high dimensional data into its lower dimensional form effectively and improve the performance accuracy of the classifier. Secondly, it reduces the over fitting problem by train the dataset at the time of feature selection procedure itself; hence the computational time of training data is also trimmed [32].

To evaluate the results more accurately, we have calculated the accuracy (in %) of all the dimensional reduction techniques. For that, we have taken the most popular SVM classifier. Here, the l_1 norm based SVM for linear regularization has been applied to perform the accuracy of various dimensional reduction techniques. Figure 10 displays the accuracies of all the proposed FEA and FSA based DDRs respectively. From figure 10, it can be clearly observed that FSA play a vital role in the reduction of high dimensional data relevantly than any other FEA. it is better to choose a novel FSA for two reasons. First reason is that FSA can reduce the higher dimension of high dimensional data into its lower dimensional form effectively and improve the performance accuracy of the classifier. Secondly, it reduces the overfitting problem by train the dataset at the time of feature selection procedure itself; hence the computational time of training data is also trimmed.

5. CONCLUSION

This paper has discussed various methods for high dimensional data dimension reduction briefly. From the fundamentals of these selection criteria, feature relevancy and feature redundancy are focused. We have also described the taxonomy of feature selection methods. These methods are segregated into filter, wrapper and embedded type techniques based on their classification dependency and computational cost. Several feature selection algorithms such as HGAPSO, ReliefF, Minimum Redundancy Maximum Relevance (MRMR), Recursive Feature Elimination (RFE) and Simultaneous Perturbation Stochastic Approximation (SPSA) have discussed in brief. A case study has been studied to validate each of the algorithms, for which a medical dataset has tabulated. From the experimented values, MRMR is proved to be one of the finest algorithms to reduce HDD into a valid lower dimensional dataset with minimum iteration compared to all other algorithms. Further extending the experiments, each of the selected features have fed to SVM to find out the accuracy. The resultant is quite clear that MRMR algorithm in addition with SVM provides better accuracy and fits the model well while trimming the computational time.

6. REFERENCES

- [1] C. Alexander and L. Wang, "Big data in healthcare: A New frontier in personalized medicine," *Am J Hypertens Res*, vol. 1, pp. 15-18, 2017.
- [2] J. A. Basco and N. Senthilkumar, "Real-time analysis of healthcare using big data analytics," *Comput Inf Technol*, vol. 263, p. 042056, 2017.
- [3] J. Archenaa and E. M. Anita, "A survey of big data analytics in healthcare and government," *Procedia Computer Science*, vol. 50, pp. 408-413, 2015.
- [4] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, vol. 2, p. 3, 2014.
- [5] D. Chen and H. Zhao, "Data security and privacy protection issues in cloud computing," in *2012 International Conference on Computer Science and Electronics Engineering*, 2012, pp. 647-651.
- [6] K. Nag and N. R. Pal, "A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification," *IEEE transactions on cybernetics*, vol. 46, pp. 499-510, 2015.
- [7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157-1182, 2003.
- [8] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research*, vol. 5, pp. 1205-1224, 2004.
- [9] A. Chadha, B. R. Iyer, H. Messatfa, and J. Yi, "Dimension reduction for data mining application," ed: Google Patents, 2000.
- [10] J. Tang, S. Alelyani, and H. Liu, "A survey of dimensionality reduction techniques," *Data Classification: Algorithms and Applications*; CRC Press: Boca Raton, FL, USA, 2015.
- [11] D. Giradi and A. Holzinger, "Dimensionality reduction for exploratory data analysis in daily medical research," in *Advanced Data Analytics in Health*, ed: Springer, 2018, pp. 3-20.
- [12] J. Ye and Q. Li, "LDA/QR: an efficient and effective dimension reduction algorithm and its theoretical foundation," *Pattern recognition*, vol. 37, pp. 851-854, 2004.
- [13] C. J. Burges, *Dimension reduction: A guided tour*: Now Publishers Inc, 2010.
- [14] D. L. Sun and C. Fevotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2014, pp. 6201-6205.
- [15] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, pp. 2507-2517, 2007.
- [16] R. Kumar, "Blending roulette wheel selection & rank selection in genetic algorithms," *International Journal of Machine Learning and Computing*, vol. 2, pp. 365-370, 2012.
- [17] S. Mirjalili, "Genetic algorithm," in *Evolutionary algorithms and neural networks*, ed: Springer, 2019, pp. 43-55.
- [18] D. Gong, J. Sun, and Z. Miao, "A set-based genetic algorithm for interval many-objective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 22, pp. 47-60, 2016.
- [19] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft Computing*, vol. 22, pp. 387-408, 2018.
- [20] N. Ghorbani, A. Kasaeian, A. Toopshekan, L. Bahrami, and A. Maghami, "Optimizing a hybrid wind-PV-battery

- system using GA-PSO and MOPSO for reducing cost and increasing reliability," *Energy*, vol. 154, pp. 581-591, 2018.
- [21] T. R. Reddy, B. V. Vardhan, M. GopiChand, and K. Karunakar, "Gender prediction in author profiling using ReliefF feature selection algorithm," in *Intelligent Engineering Informatics*, ed: Springer, 2018, pp. 169-176.
- [22] S. Chikhi and S. Benhammada, "ReliefMSS: a variation on a feature ranking ReliefF algorithm," *International Journal of Business Intelligence and Data Mining*, vol. 4, pp. 375-390, 2009.
- [23] Y. Zhang, C. Ding, and T. Li, "Gene selection algorithm by combining reliefF and mRMR," *BMC genomics*, vol. 9, p. S27, 2008.
- [24] R. Cai, Z. Hao, X. Yang, and W. Wen, "An efficient gene selection algorithm based on mutual information," *Neurocomputing*, vol. 72, pp. 991-999, 2009.
- [25] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353-363, 2015.
- [26] W. You, Z. Yang, and G. Ji, "PLS-based recursive feature elimination for high-dimensional small sample," *Knowledge-Based Systems*, vol. 55, pp. 15-28, 2014.
- [27] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, et al., "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 89-96.
- [28] J. L. Maryak and D. C. Chin, "Global random optimization by simultaneous perturbation stochastic approximation," in *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*, 2001, pp. 756-762.
- [29] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American statistical association*, vol. 97, pp. 77-87, 2002.
- [30] A. M. Brunner, F. Campigotto, H. Sadrzadeh, B. J. Drapkin, Y. B. Chen, D. S. Neuberg, et al., "Trends in all-cause mortality among patients with chronic myeloid leukemia: a Surveillance, Epidemiology, and End Results database analysis," *Cancer*, vol. 119, pp. 2620-2629, 2013.
- [31] E. J. Wang, T. H. Bø, I. Jonassen, O. Myklebost, and E. Hovig, "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data," *BMC bioinformatics*, vol. 4, pp. 1-12, 2003.
- [32] J. R. Berrendero, A. Cuevas, and J. L. Torrecilla, "The mRMR variable selection method: a comparative study for functional data," *Journal of Statistical Computation and Simulation*, vol. 86, pp. 891-907, 2016.