Integrated Intelligent Research (IIR)
International Journal of Computing Algorithm
Volume: 06 Issue: 02 December 2017 Page No.64-68
ISSN: 2278-2397

Stability of Indexed Microarray and Text Data

T.Velmurugan[1] S.Deepa Lakshmi[2]
[1]Associate Professor, PG and Research Department of Computer Science, D.G.Vaishnav College, Chennai, India.
[2]Research Scholar, Bharathiar University, Coimbatore, India.
Email: velmurugan_dgvc@yahoo.co.in, deepa.dgvc@gmail.com

Abstract: The common challenge for machine learning and data mining tasks is the curse of High Dimensionality. Feature selection reduces the dimensionality by selecting the relevant and optimal features from the huge dataset. In this research work, a clustering and genetic algorithm based feature selection (CLUST-GA-FS) is proposed that has three stages namely irrelevant feature removal, redundant feature removal, and optimal feature generation. The performance of the feature selection algorithms are analyzed using the parameters like classification accuracy, precision, recall and error rate. Recently, an increasing attention is given to the stability of feature selection algorithms which is an indicator that requires that similar subsets of features are selected every time the algorithm is executed on the same dataset. This work analyzes the stability of the algorithm on four publicly available dataset using stability measurements Average normal hamming distance(ANHD), Dices coefficient, Tanimoto distance, Jaccards index and Kuncheva index.

Keywords:Stability measurements, Jaccard Index, Kuncheva Index, Tanimoto Distance, Dice Coefficient.

I. INTRODUCTION

Data mining and machine learning tools were proposed to automate pattern recognition and knowledge discovery process. The growth of technology has led to exponential growth of data with respect to dimensionality. Storing and processing high dimensional data has become more challenging[1]. Dimensionality reduction is a technique to reduce the size of the dataset and can be categorized into feature extraction and selection[2]. Feature extraction projects the features into new space with low dimension. Examples of feature extraction are principal component analysis, linear discriminant analysis etc. Feature selection aims to select a small subset of relevant features[3]. Feature selection is classified into three methods namely filter method, wrapper method and embedded method[4]. The proposed algorithm CLUST-GA-FS is an embedded feature selection method. Commonly used evaluation metrics include classification accuracy, precision, recall and error rate. The selection stability has drawn an increasing attention recently.

The selection stability is desired characteristic for feature selection algorithms. Selection stability is the sensitivity of a feature selection algorithm to perturbation in the training data[5]. Stability metrics are used to assess the stability of multiple feature selection results. A feature selection algorithm is considered as stable if the features selected are consistent from multiple execution of the algorithm with the same dataset[6]. Petr Somol et al. proposed a modified form of Average Normalized Hamming Index and the family of inter measures was introduced [7]. An extension to Rapid Miner was proposed in which Least Angle Regression algorithm was used. Kuncheva index was used to measure the stability of MRMR and an ensemble version of MRMR [8]. Suphakit et al. proposed a method for measuring the similarity between words using Jaccard Coefficient which was developed using Prolog language[9]. Yvan Saeys et al. used multiple feature methods to produce more robust result on Microarray datasets. Stability metrics Spearman correlation and Jaccard index were used to determine the robustness of the feature selectors[10]. A Graph based feature selection method that assess the nodes using Eigen Vector Centrality was proposed and the stability was assessed using Kuncheva index[11]. Iman Kamkar et al. proposed a new method that uses intrinsic graph structure of the electronic medical records in which the graph structure was exploited using Jaccard similarity[12]. David Dernoncourt et al. introduced ATI_{PA} stability measure which is a modification of ATI that aims to avoid a bias on the number of selected features [13]. Sarah et al. illustrated the benefits of stability measures and showed that Pearson's correlation similarity measure is a generalization of Kuncheva index[14]. Kehan Gao et al. used six filter feature selection techniques , three sampling approaches and the stability measure Consistency index on software measurement data to study the impact of data sampling on Stability of Feature Selection[15]. The work presented in this paper analyses the different stability metrics and applies to the proposed algorithm.

The paper is organized as follows: the various stability measurements are explained in chapter 2, the proposed algorithm CLUST-GA-FS is discussed in chapter 3. Chapter 4 discusses the stability of the proposed algorithm.

II. STABILITY MEASUREMENTS

The Stability of the feature selection algorithm is the sensitivity of the selection to the variation of the dataset. The stability can be assessed by pair wise comparison between the results. If the similarity is greater, the stability is higher. There are three different representations of the output of the feature selection algorithms namely indexing, ranking and weighting.

A. Stability by Index

Let the dataset be X and the set of features be F. The selected subset of features is represented as a vector of indices that correspond to the features $f \subset F$. Let m be the number of features in the dataset and k be the number of features selected and $k \leq m$. Stability measurement assesses the amount of overlap between results.

Average Normal Hamming Distance(ANHD): ANHD measures the amount of overlap between the features

International Journal of Computing Algorithm
Volume: 06 Issue: 02 December 2017 Page No.64-68
ISSN: 2278-2397

selected on multiple execution of the algorithm using the same dataset[6]. It represents the selected feature subset as f_{ii}, where f_{ik} has a value 1 if the k^{th} feature was selected in the i^{th} run and a value 0 if the k^{th} feature was not selected in the i^{th} run. The stability between two set of features selected by the algorithm f_i and f_j is given by

$$H(f_i, f_j) = \frac{1}{m}\sum_{k=1}^{m} |f_{ik} - f_{jk}| \quad (1)$$

H value is in the interval [0,1], where 0 indicates more stability and 1 means that it is not stable. As the value of m is large for high dimensional data, the value of H is small and it indicates a more stable algorithm. The property of ANHD will mislead to higher stability when k<m where majority of features are not selected. It cannot deal with different size of selected features.

Dice's Coefficient: Dice's coefficient is a similarity measure used to calculate the overlap between two set of features[16]. It takes values between [0,1] where 0 means no overlap and 1 means the two set of features are identical. Dice's coefficient is given by

$$Dice(f_i, f_j) = \frac{2|f_i \cap f_j|}{|f_i| + |f_j|} \quad (2)$$

Where f_i and f_j are the two feature subsets generated by the feature selection algorithm.

Tanimoto Distance: Tanimoto measures the overlap between two set of features and has value in the same range as Dice[17]. It produces values in the range [0,1]. Tanimot distance is given by

$$Tanimoto(f_i, f_j) = 1 - \frac{|f_i| + |f_j| - 2|f_i \cap f_j|}{|f_i| + |f_j| - |f_i \cap f_j|} \quad (3)$$

Jaccard's Index: Jaccard index known as intersection over union is used for comparing the similarity of feature subsets[18]. It produces values in the range [0,1]. The advantage is that it can deal with sets of different cardinalities and Dice, Tanimoto and Jaccard do not take the dimensionality into account. Jaccard index is given by

$$Jaccard(f_i, f_j) = \frac{|f_i \cap f_j|}{|f_i \cup f_j|} \quad (4)$$

Kuncheva Index KI: The drawback of the stability measurements is that the overlap between the selected feature subset is more due to chance as the cardinality of the feature subset is larger[19]. Kuncheva index contains a correction term to avoid intersection of the selected subset of features by chance. It produces values in the range [-1,1] where -1 means there is no intersection between the subsets and 1 means the subsets are identical. Kuncheva index is given by

$$KI(f_i, f_j) = \frac{|f_i \cap f_j| . m - k^2}{k(m-k)} \quad (5)$$

B. Stability by Ranking

Stability by rank evaluates the correlation between the ranking vectors. They can handle vectors with different cardinality or vectors that have different set of features. Some of the stability measurements are Spearman's rank correlation coefficient and Canberra distance[16]. Filter methods can be assessed using rank based metrics and Wrapper methods can be assessed using set based stability metrics. In rank based metrics, there are two types depending on whether the list obtained is full ranking or partial ranking. For full ranking list, the stability is assessed based on the rank of all features and for partial list, a threshold is specified and the features greater than the threshold are used for stability assessment. Spearman's correlation is the widely used full ranking metric and its value ranges between -1 to +1 with 0 indicating no correlation, +1 indicating positive correlation and -1 indicating negative correlation. Partial ranking metrics used are Jaccard index and Kuncheva index.

C. Stability by Weighting

Stability by weight deals with the weight of the feature set. Pearson's correlation is the measurement that takes two set of features and returns stability which is the correlation between them. It cannot deal with different subset size. It produces values in the range [-1,1] where 1 means the feature subset are correlated, -1 means they are anti-correlated and 0 means they are not correlated[16].

Table 1: Stability measurement Categories

Measures	Stability by	Different Size	Range
ANHD	Index	Yes	[1,0]
Dice	Index	Yes	[0,1]
Tanimoto	Index	Yes	[0,1]
Jaccard	Index	Yes	[0,1]
Kuncheva	Index	Yes	[-1,1]
Spearman	Rank	No	[-1,1]
Canberra	Rank	No	[0, ∞]
Pearson	Weight	No	[-1,1]

III. CLUST-GA-FS ALGORITHM

The feature selection algorithm CLUST-GA-FS has three components: irrelevant feature removal, redundant feature removal, and optimal feature generation. The first component removes the irrelevant features by using mutual information, a filter method. The second component removes the redundant features by choosing the representatives from each cluster[20]. The genetic algorithm is used as the third component to find the optimal set of features. The irrelevant feature removal obtains features relevant to the class by eliminating the features which are irrelevant to the target class using mutual information. Redundant feature removal removes features that are redundant in 3 steps: Constructing a minimum spanning tree from the relevant features, grouping the features in the forest into clusters and selecting the representative feature from each cluster. The set of a representative feature from each cluster, the class variable and the number of features desired is provided as input to a genetic algorithm. Genetic algorithm (GA) is an adaptive heuristic search that uses optimization techniques to find true or approximate solutions based on the evolutionary ideas of natural selection and genetics. GA begins with a

International Journal of Computing Algorithm
Volume: 06 Issue: 02 December 2017 Page No.64-68
ISSN: 2278-2397

set of chromosomes called the population. New populations are evolved by mutating solutions according to their fitness value. The fitness function used in this proposed method is based on the principle of min-redundancy and max-relevance.

Proposed Algorithm: CLUST-GA-FS

Inputs: Data set S{f_1,f_2,….,f_m C}, C-class, Θ - threshold value
Output: representative feature subset
//=====Part 1 : Irrelevant feature removal=======
Step 1: for each feature f_i in the data set
compute MI(f_i,C)
if MI(f_i,C)<Θ
remove the feature f_i
end
end
relevant feature set F'={f'$_1$,f'$_2$,….,f'$_l$}(l≤m).

//===== Part 2 : Removal of Redundant feature=====

Step 2: for each feature f_i in F'
construct graph G=(V,E) with feature f_i as vertices
and MI(f_i,f_j) as edges
end
generate the minimum spanning tree of G
forest=minimum spanning tree of G
for each edge in forest
if MI(f_i,f_j)<MI(f_i,C) and MI(f_i,f_j)<MI(f_j,C)
remove edge from the forest
end
end
for each tree in the forest
find the maximum MI(f_i,C)
select f_i as the representative feature of the cluster
end
representative feature set is F''={f'$_1$,f'$_2$,….,f'$_k$}(k≤l).

//======= Part 3 : Generating Optimum set of features using Genetic Algorithm======

Step 3: Input: Representative feature set F''={f'$_1$,f'$_2$,….,f'$_k$},
Class C, desired no of features
Output: Optimal set of features

Max_gen=no.of generations desired
find the entropy of F''- H_f and C- H_C and mutual information between the features - MI_{ff}
and mutual information between the features and class C- MI_{fC}
generate the population consisting of the feature set
while generation is less than max_gen
find the fitness function=$max_S\left[\frac{1}{|F''|}\sum_{f_i\in F''} MI(f_i,c) - \frac{1}{|S|^2}\sum_{f_i f_j\in F''} MI(f_i,f_j)\right]$
rearrange the population according to their fitness values
create a new generation
if the chromosomes generated are identical
sel=population rearranged
end
end
sel=optimal set of features

Steps1 and 2 generate the relevant set of features and remove the redundant feature. The representative set of features is given as input to a genetic algorithm that determines the optimal set of features.

IV. EXPERIMENTAL RESULTS

The proposed algorithm CLUST-GA-FS is executed on microarray and text data. Sample sets are generated by random sampling or l-fold cross validation from a given dataset. To analyze the stability of the algorithm, the algorithm is applied to each sample set and selects optimal set of features. Similarity measures are applied to evaluate the similarity between the features lists obtained. The dataset used and the stability assessment are discussed below.

A. Dataset Description

The implementation of CLUST-GA-FS algorithm was done on four publicly available dataset. Three microarray datasets and one text dataset are used in this work. The number of features of the dataset varies from 1800 to 10000, the instance also varies from 38 to 5000 and the number of classes is 2.

Table 2: Data Set Description

S. No.	Data set	No. of features	No. of instances	No. of Classes	Domain
1	Colon	2000	62	2	Microarray
2	Leukemia	7129	38	2	Microarray
3	Arcene	10001	200	2	Microarray
4	SMS spam	1833	5574	2	Text

B. Stability Assessment

Sample dataset are obtained from the original dataset X by using random sampling or l-fold cross validation and the sample datasets are labeled X1 and X2. The algorithm is executed on all the sample dataset and the optimal set of features is generated. Let the feature list generated by the proposed algorithm on the sample X1 be F1 and the feature list generated by the proposed algorithm on the sample X2 be F2. Let S() be the measurement to assess the stability of the algorithm. S(F1,F2) is the stability measure of the algorithm. Since the output representation of the proposed algorithm is an indexed set of optimal features, stability by index measurements are used to assess the stability of the algorithm. The Stability by index measurement used is ANHD, Dice, Tanimoto, Jaccard and Kuncheva Index.

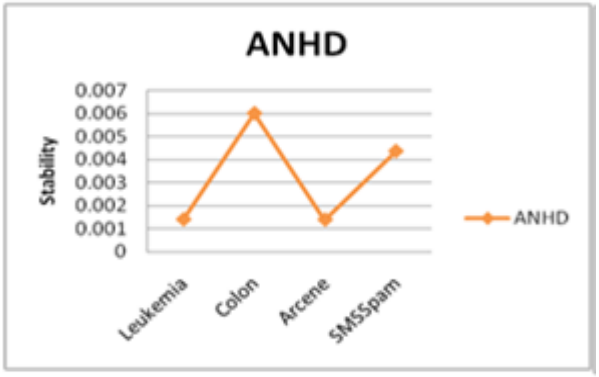
The assessment of stability of the proposed feature selection algorithm CLUST-GA-FS is performed using the stability metrics ANHD, Dice, Tanimoto, Jaccard and Kuncheva. Different samples(X1 and X2) of the dataset(X) are obtained

International Journal of Computing Algorithm
Volume: 06 Issue: 02 December 2017 Page No.64-68
ISSN: 2278-2397

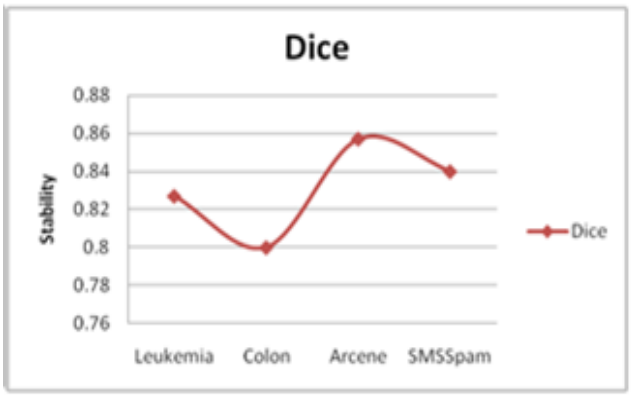
using l-fold cross validation. The proposed algorithm is executed on each sample of the dataset. The output of the CLUST-GA-FS algorithm is an indexed list of optimal features F1 and F2 respectively. The stability metric is applied on the optimal features list obtained using two samples of the dataset S(F1,F2).

Table 3 : Stability measurements

Dataset	ANHD	Dice	Tanimoto	Jaccard	Kuncheva Index
Leukemia	0.0014	0.827	0.705	0.705	0.826
Colon	0.006	0.8	0.666	0.666	0.796
Arcene	0.00139	0.8571	0.75	0.75	0.856
SMSSpam	0.00436	0.84	0.724	0.724	0.8377



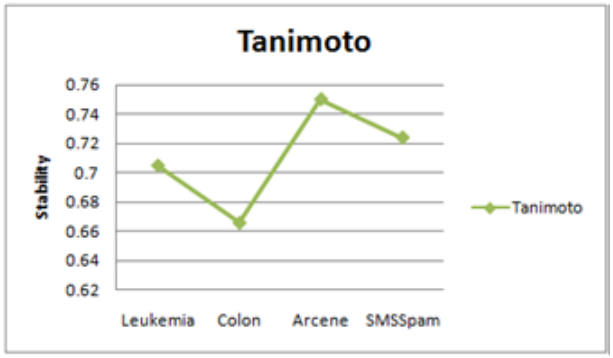
a



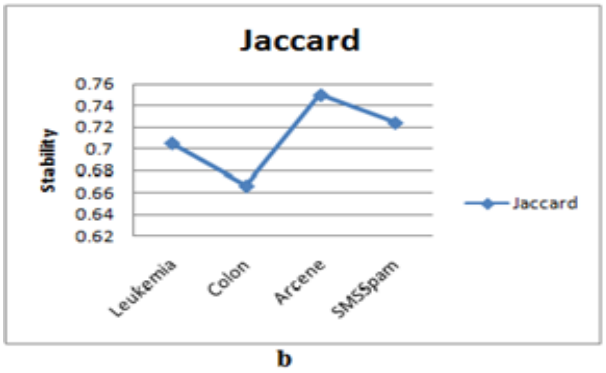
b

Figure 1: ANHD and Dice Stability measurements

The Stability metric ANHD(F1,F2) produces values in the range [0,1] where the value 0 indicates high stability and 1 indicates no stability. From the figure 1 a, it is found that ANHD has produced stable results for all the datasets and Arcene dataset in particular. The stability measurement of all the datasets are close to the value 0 indicating high stability. Dice(F1,F2) produces values in the range [0,1] where 0 indicates no stability and 1 indicates high stability. The stability measurement Dice produces high stability values for all the datasets Leukemia,Colon, Arcene and SMSSpam. A value of 0.8 indicates high stability and all the datasets have a Dice stability close to 0.8 as shown in the figure 1 b.



a



b

Figure 2: Tanimoto and Jaccard Stability measurements

Jaccard Stability measure deals with binary attributes and Tanimoto is an extended version of Jaccard. The presence of a feature in the optimal feature list has a value 1 and the absence of a feature has a value 0. From figure 2, it is evident that Tanimoto and Jaccard produce the same similarity measure values for all the datasets. Tanimoto and Jaccard produce values in the range [0,1] where 1 indicates high stability. Tanimoto and Jaccard have produces high stability for Leukemia, Arcene and SMSSpam datasets. The stability measure produced for Colon dataset is comaparatively less.

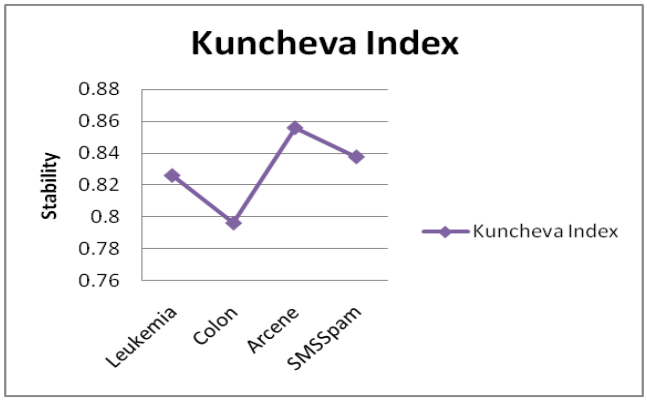


Figure 3: Kuncheva Index Stability measurements

Kuncheva index Stability measurement produces values in the range [-1,1]. The higher the value, more stable is the algorithm. The stability metric produced for all the datasets varies between 0.79 and 0.85 which is closer to the value 1 indicating that the proposed algorithm has produced stable optimal feature set.

International Journal of Computing Algorithm
Volume: 06 Issue: 02 December 2017 Page No.64-68
ISSN: 2278-2397

V. CONCLUSION

The performance of an algorithm is analyzed using different metrics classification accuracy, precision, error rate and recall. In this research work, the quality of the feature set obtained by the proposed algorithm is analyzed using Stability measurements. The algorithm removes the irrelevant features using mutual information, removes redundant features by constructing a minimum spanning tree, splitting the minimum spanning tree into clusters and selecting a representative feature from each cluster. The representative features are given to a genetic algorithm to obtain an optimal set of features. Sample datasets are obtained from the datasets and the Stability metrics ANHD, Dice, Tanimoto, Jaccard and Kuncheva Index are applied to the feature sets obtained using different sample datasets. The Stability metrics have produced stable results for the microarray and text datasets Leukemia, Colon, Arcene and SMSSpam.

References

[1] V. Kumar and S. Minz, "Feature Selection," SmartCR, 2014, Vol.4, Issue 3 pp.211-229.

[2] I. Guyon and A. Elisseeff, "An introduction to feature extraction," Featur. Extraction., pp.1-25, 2006.

[3] G. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," In Machine learning: proceedings of the eleventh international conference, pp. 121-129. 1994.

[4] Q. Song, J. Ni, and G. Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data," IEEE transactions on knowledge and data engineering Vol.25, Issue 1, pp.1-14, 2013.

[5] C. Science and N. York, "Stability of Feature Selection Algorithms", Doctoral dissertation, Department of Computer Science Binghamton University, State University of New York, New York, 2010.

[6] K. Dunne, P. Cunningham, and F. Azuaje, "Solutions to Instability Problems with Sequential Wrapper-based Approaches to Feature Selection," Journal of Machine. Learning, pp. 1–22, 2002.

[7] Somol, P. and Novovicova, J, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality", IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(11), pp.1921-1939, 2010.

[8] Schowe, Benjamin. "Feature selection for high-dimensional data with RapidMiner." In Proceedings of the 2nd RapidMiner Community Meeting And Conference (RCOMM 2011), Aachen. 2011.

[9] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard Coefficient for Keywords Similarity," In Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, no. 6, pp.380-384, 2013.

[10] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques", Machine learning and knowledge discovery in databases pp.313-32, 2008.

[11] G. Roffo and S. Melzi, "Feature Selection via Eigenvector Centrality", Proceedings of New Frontiers in Mining Complex Patterns (NFMCP 2016)Oct 2016.

[12] I. Kamkar, S. Gupta, Cheng Li, D. Phung, and S. Venkatesh, "Stable clinical prediction using graph support vector machines," in 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 3332–3337, 2016.

[13] D. Dernoncourt, B. Hanczar, and J. D. Zucker, "Analysis of feature selection stability on high dimension and small sample data," Computational Statistics & Data Analysis, vol. 71, pp. 681–693, 2014.

[14] S. Nogueira and G. Brown, "Measuring the stability of feature selection,", In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 442-457. Springer International Publishing, 2016.

[15] K. Gao, T. M. Khoshgoftaar, and A. Napolitano, "Impact of Data Sampling on Stability of Feature Selection for Software Measurement Data", IEEE 23rd International Conference on Tools with Artificial Intelligence, pp. 1004–1011, 2011.

[16] S. Alelyani, "On Feature Selection Stability: A Data Perspective", Arizona State University, no. May, 2013.

[17] A. Kalousis, J. Prados, and M. Hilario, "Stability of Feature Selection Algorithms," Fifth IEEE International Conference, Data Mining, pp. 218–225, 2005.

[18] T. M. Khoshgoftaar, A. Fazelpour, H. Wang, and R. Wald, "A survey of stability analysis of feature subset selection techniques," 2013 IEEE 14th International Conference Information Reuse and Integration, pp. 424–431, 2013.

[19] L. I. Kuncheva, "A stability index for feature selection," International Multi-conference Artificial Intelligence Applications, pp. 390–395, 2007.

[20] S. Deepalakshmi and T. Velmurugan, "A Clustering and Genetic Algorithm based Feature Selection (CLUST-GA-FS) for High Dimensional Data," International Journal of control theory and Applications, vol. 10, no. 23. pp. 63–76, 2017.