International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August, 2019

A Stable SVM-RFE Feature Selection Method for Gene Expression Data

Shaveta Tatwani, Ela Kumar

Abstract: ***Feature Selection techniques are generally employed to remove the inessential attributes before machine learning technique could be applied. It thus plays an extremely important role by eliminating the unnecessary features that do not contribute and sometimes degrade the performance and prediction accuracy of the machine learning technique. With the growth of dimensionality of data, Feature Selection becomes even more important because it helps to reduce the dimensions of data and hence decreases the requisite memory and computational complexity of the machine learning techniques. Support vector machine-recursive feature elimination (SVM-RFE) has proven to be an efficient wrapper feature selection technique which continues to be widely utilized in many applications, especially in classification of gene expression data. From the perspective of this data, not only the precision in classification but also the stability of Feature Selection method plays an important role. Nonetheless, the topic of stability is ignored in study of feature selection algorithms. To improve the stability of RFE method, a fusion of Information Gain and RFE (IG-RFE-SVM) method is proposed in this paper. Experimental studies show that IG-RFE-SVM outperforms SVM-RFE method in terms of stability.***

Keywords: Feature Selection, Gene Expression Data, Machine Learning, Recursive Feature Elimination, Support Vector Machine.

I. INTRODUCTION

Bioinformatics is the interdisciplinary research area that analyzes the biological data i.e. gene expression data using statistical method and software tools. To analyze the gene expression data is difficult for classification because the data is high dimensional and generally, the performance of traditional Machine Learning (ML) techniques degrades when they are applied to High-Dimensional Datasets (HDD) of various applications due to the "curse of dimensionality" [1][2]. This is because the complexity of existing machine learning algorithms is typically proportional to the exponent of the degree of dimensions. As in HDD, the number of dimensions is very large and hence existing algorithms produce computational challenges and thus, become inefficient when applied in the real world. Further, these algorithms are generally based on Similarity Metric which is calculated by distance or nearest neighbor concept and this concept fails in HDD as it is proven in numerous literatures. This is because the expected difference between the Euclidean distances of any point from its closet point to the farthest point decreases, as the dimensionality increases [3]. Due to this the machine learning method becomes ineffective. Also over-fitting may happen, significantly impacting the performance of machine learning technique on test data set [4]. Furthermore, in the HDD, several features are redundant or irrelevant. These large numbers of redundant and irrelevant features require huge memory and storage space, thus having consequences on machine learning algorithms towards its performance and computational cost.

Due to these drawbacks of the existing machine learning algorithms, data dimensionality reduction is required to make them effective. Feature Selection (FS) techniques are the simplest and most frequently used techniques so that the dimensionality of data is decreased [5][6][7]. In the FS technique the selection of the subset, which is optimal, is done from high dimensional set according to the certain rules. It enhances the ML efficiency with regard to result precision and speed through removal of the extraneous data. In FS techniques, the new features are a subset of the original feature set formed without any transformation and hence, do not lose their physical meaning. FS Technique, when employed on gene expression data, is also referred to as Gene Selection. It mainly serves two motives: 1) it reduces the volume of data through removal of genes which are non-relevant and increases the classification accuracy; 2) it helps in identifying the genes which are the cause of certain diseases. Hence, classification accuracy as well as the set of features, which are end products of feature selection, is what matters most in analysis of gene expression data. Stability of FS techniques refers to the sensitivity of that technique to varying condition due to changes in sample data. Therefore, it is necessary that with slight change or perturbation in the dataset, chosen feature set does not change [8][9][10]. Even without perturbation of data sample, sometimes the FS techniques produce the different results. The inconsistent results could cause confusion and hence, fail in identifying the genes which might cause of certain diseases. Hence, while analyzing the gene expression data, stability is an important parameter. The rest of the paper is organized to initially examine FS technique process in Section II, then to explore FS techniques stability issue in Section III, followed by working principle of the proposed FS technique. Finally, experimental setup and results are presented in section V, with conclusion under section VI.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Shaveta Tatwani*, Department of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, New Delhi-110006, India. E-mail: shaveta.tatwani@gmail.com
Ela Kumar, Department of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, New Delhi-110006, India. E-mail: ela_kumar@rediffmail.com

II. FEATURE SELECTION (FS) TECHNIQUES

FS technique is a way of finding optimal set of features from N dimensional data having 2^N possible subsets of features. They are initially applied in pre-processing step to remove the noisy data. But the growth of high dimensional data makes the feature selection an important step as it chooses only those attributes which are imperative and useful for the given situation.

There are three types of FS techniques: Wrapper method, Filter method and Hybrid method, as discussed below:

A. Wrapper Methods

In Wrapper methods, the importance of features is calculated depending on the performance of machine learning technique. All possible subsets are generated using some search techniques and then fitness of subset is evaluated using machine learning technique itself.

I. Guyon et al. [11] proposed the wrapper based Recursive Feature Elimination for Support Vector Machines (SVM-RFE) method, which is based on iterative procedure with three main stages. To start with, it trains the SVM classifier, then in second step it calculates the ranks of each feature which are given according to weight assigned by the SVM and in the final phase it removes the features having the lower rank. This method has been experimentally proven better for Feature Selection and classification problem on microarray data in comparison to alternative correlation based Feature Selection methods [11][12].

Another wrapper method proposed by M. Kabir et al. [13] is, Constructive Approach for Feature Selection (CAFS). In this method, the Neural Network is used in subset evaluation process using a constructive approach by identifying the correlation between the dataset features. The search strategy selects less associated features if they improve the accuracy of the Neural Network, resulting in compact databases.

J. Leng et al. [14] used the Genetic Algorithm based wrapper method. In this method, Genetic algorithm is used for subset generation, and then the subset is evaluated depending upon the performance of KNN (K-Nearest Neighbor Classifier). Experimental results proved than Genetic based FS technique improve the performance of KNN.

Wrapper methods are accurate as they use Machine Learning algorithm itself to assess the significance of a characteristic. But they are most expensive because FS algorithm itself is a combinatorial problem and the calculation of importance of a feature using machine learning algorithm makes it more time consuming.

B. Filter Methods

The criterion on which Filter method selects the optimal set of features depends upon the general characteristics of dataset which is independent of the machine learning technique. This method selects the feature based upon the rank given to each feature after estimating its relevance and then filtering out less useful, lower rank features which do not play an important role in prediction of data.

The RELIEF Feature Selection method is proposed by Kira and Randell [15] that calculates the relevance of each feature based on its Euclidean distance. More the difference between the Euclidean distances of a feature from the instance of the identical category and instance of the dissimilar category, higher the applicability of that feature. It works well for HDD but this method does not eliminate the redundant features.

Correlation-based Feature Selection (CFS) method [16][17] is another technique of this category. It selects the subset having features which are relevant with the class but they have less inter feature relevance. In this method, relevance is measured in terms of Pearson Correlation. As it also measures inter feature correlation, hence it eliminates the redundant features.

Fast Correlation-Based Feature Selection (FCBF), which is developed by Yu and Liu [18][19], removes both irrelevant and redundant features. It applies the sequential forward search for subset generation by using correlation and consistency measures as evaluation function to guide the search. Since this method uses sequential forward search, it is more suitable for HDD.

The Minimum Redundancy Maximum Relevance (mRMR) criterion is used by H. Peng et al. for feature selection [20]. In this technique, redundancy calculations are done on the basis of the Mutual Information shared among the features, whereas relevance is computed depending on the Mutual Information shared among individual feature and its associated class. The mRMR method shows better results when applied to pattern recognition problem and gene expression classification problem [20][21]. Further, a recent work done by Franay B [22] proves by experimental results that the Mutual Information based Feature Selection techniques do not always enhance the performance of ML techniques. In spite of its property to maximize the Mutual Information of each individual feature with its group, it does not always guarantee decrease in the generalization error. A survey [23] discusses the use of Feature Selection filters in the field of Bioinformatics.

A new Genetic Algorithm was proposed by Jungjit and Freitas [24] for Multi-Label Correlation–Based Feature Selection (GA-ML-CFS). They applied this method on different multi-label data sets and compared the results with Hill-Climbing method; to prove that GA based feature section provides higher predictive accuracies. In their extended work, a Lexicographic Multi-Objective Genetic Algorithm (LexGA-ML-CFS) [25] is proposed and compared with the previous two techniques on multi-label data. The end results reveal that the LexGA-ML-CFS has enhanced the execution of classification when compared to GA-ML-CFS and HC-ML-CFS.

As these Filter approaches assess the fitness of the subset by using statistical measures considering the inherent properties of the data, rather than the Machine Learning algorithm, hence they are fast to compute and more suitable than wrapper methods for HDD.

C. Hybrid Methods

With the available literature it is observed that Filter methods provide better time and space efficiency but are not accurate. In contrast, Wrapper methods, though more precise, are expensive when applied on high dimension data. Thus, Hybrid methods were proposed to combine the above two methods into a single system for Feature Selection, which would help attain the best characteristics of both the methods.

The Feature Selection method based on Greedy Randomized Adaptive Search Procedure (GRASP) was suggested by M. Esseghir [26].

It is hybrid method having two phases; Constructive phase and Improvement phase, where Filter techniques are used in Constructive phase and Wrapper techniques are used in Improvement phase. An enhancement over this method was also proposed which speeds up the GRASP method by significantly decreasing the number of Wrapper evaluations [27].

A Wrapper-Filter Feature Selection Algorithm (WFFSA) is proposed on the concept of a memetic framework [28]. This memetic framework is based on a ranking method as filter technique, Genetic algorithm and Nearest Neighbor Classifier as Wrapper technique. Subset of features are generated using Genetic algorithm, the population of features are assessed and further selected based on Classifier performance. Afterwards, the ranking method is used to refine the subset generated using local search. The ranking methods used by authors are: ReleifF, Gain Ratio and Chi-square parameter. They showed that this technique outperforms the existing techniques in terms of generalization error, size of subset and time complexity.

A Correlation-based Memetic Framework proposed by Kannan and Ramaraj [29], also uses Genetic algorithms search technique for subset generation. Subset is evaluated using Naïve Byes Classification followed by the local search in which Filter method refines the process of Genetic Algorithm either by inserting or by removing certain features considering the ranking calculated according to symmetrical uncertainty.

Another recent work by Juanying Xie et al. proposed a framework named Improved F-score and Sequential Forward Floating Search [30]. In this method, Wrapper component constitute of Sequential Forward Floating Search (SFFS) and SVM; and Filter phase constitute of improved F-score measure. This method is tested on erythemato-squamous disease dataset.

Ahuja and Ratnoo [31] proposed a hybrid approach for feature selection which employs a Multi-Objective Genetic Algorithm at filter phase based on several criteria and a simple GA at the wrapper phase which optimizes based on SVM classifier.

III. ISSUE OF STABILITY IN FEATURE SELECTION

Stability describes the measure of change in the degree of the feature subset when data sample is changed. In other words, this is the ability of the feature selection method to produce unaltered subset if data sample is changed. The instability is therefore, the main issue in Feature Selection process as it may degrade the performance of machine learning technique due to failure in identification of the most relevant features.

Stability is measured in terms of similarity between the various subsets produced by altering data sample. Commonly used stability measure in the literature is the Tanimoto Index (TI) given by A. Kalousis et al. [32] and it works as follows:

$$\begin{aligned} TI(S_i, S_j) &= 1 - S_k(S_i, S_j) \\ &= 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \\ &= 1 - \frac{|S_i| + |S_j| - 2|S_i \cap S_j|}{|S_i| + |S_j| - |S_i \cap S_j|} \end{aligned}$$

Where Sk is the similarity measure between two subsets Si and Sj.

In order to calculate the stability of FS technique, it is applied n times on n different perturbation of datasets, producing n feature subsets. Hence, the overall stability of FS technique is measured in terms of average Tanimoto index given by:

$$ATI = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S_k(S_i, S_j)$$

Randall Wald et al. [33] compared the various filter and wrapper methods based on Tanimoto index and concluded that the Filter methods are superior as compared to Wrapper based methods in terms of stability. The end results also revealed the higher stability of Correlation based Filter method in comparison to Consistency based Filter method.

L.I. Kuncheva [34] proposed another way to compute stability using Kuncheva Index (KI), which overcomes the disadvantage of 'By Chance' condition in Tanimoto Index. As Tanimoto Index does not consider the actual count of features in the dataset, therefore, once the subset size nears the total count of features, the Tanimoto distance is invariably nearly 1. Kuncheva Index (KI) for two Subsets S_i and S_j of original feature set X is shown by:

$$KI(S_i, S_j) = \frac{r.n - k^2}{k(n-k)}$$

Where S_i and S_j are of same size as k; n is the size of original feature set X; $r = |S_i \cap S_j|$ is the cardinality of intersection of two subsets.

Generalizing the above formula for set of feature subset $S = \{S_1, S_2, \ldots, S_k\}$, the Average Consistency Indices (ACI) is given by:

$$AKI = \frac{2}{K(K-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} I_c(S_i(k), S_j(k))$$

Another approach based on Hamming Distance is developed by Dunne Kevin et al. [35]. Given a feature subset S_i and S_j, Hamming distance between them is given by:

$$H(S_i, S_j) = \sum_{k=1}^{n} |S_{ik} - S_{jk}|$$

Where, n is total number of features and $S_{ik} = 1$; if Kth feature belongs to feature subset S_i else $S_{ik} = 0$.

Generalizing the formula for w feature subsets, average Hamming distance is given by:

$$H = \sum_{i=1}^{w} \sum_{j=i+1}^{w} H(S_i, S_j)$$

IV. A STABLE RECURSIVE FEATURE ELIMINATION METHOD

I. Guyon proposed the wrapper based Recursive Feature Elimination for Support Vector Machines (SVM-RFE), which is based on iterative procedure with three main stages. To start with, it trains the SVM classifier, then in second step it calculates the ranks of each feature which are given according to the weight assigned by the SVM and in the final phase, it removes the features having the lower rank. This procedure is replicated till a single feature is remaining in the subset. At the end, the output of this FS technique is that feature subset which is having highest accuracy. This method has been experimentally proven (in numerous literatures) better for feature selection in terms of classification accuracy in comparison to alternative Feature Selection methods. But it is less stable feature selection method than other filter methods. Hence, to increase the stability of SVM-RFE method we propose the fusion of Information Gain feature selection method and SVM-RFE method. The working principal of our proposed method IG-RFE-SVM is illustrated in figure1:

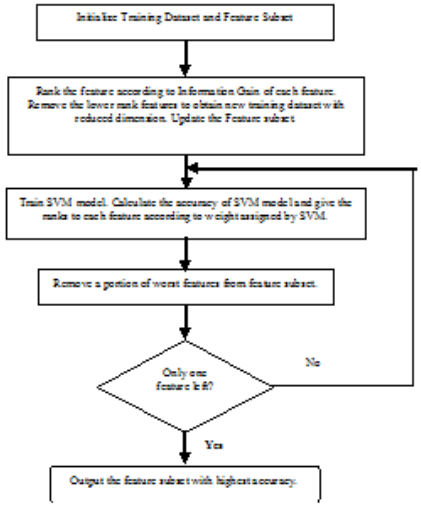

Fig. 1: Working Principle of IG-RFE-SVM

V. EXPERIMENTS AND RESULTS

A. Data Sets

In this study, the experiments are preformed on three gene-expression data sets, which is available in R package 'datamicroarray' [36].

Table- I: Datasets

S. No.	Dataset	Sample Size	No. of Classes	No. of Features
1	Colon Cancer	62	2	2000
2	Leukaemia	72	2	7129
3	Prostate Cancer	102	2	12,533

- The Colon Cancer data includes the gene expression data of about 2000 genes and 62 samples taken from colon-cancer patients. Out of these 62 samples, 40 are tumorous marked as 'negative' and 22 are non- tumorous marked as 'positive'
- The Leukaemia data set consist of 72 samples of different patient, out of which 47 patients are having acute lymphoblastic leukaemia (ALL) and 25 patients are having acute myeloid leukaemia (AML). The data matrix contains 7129 gene expressions.
- Prostate Cancer Data Set contains probes for approximately 12,533 genes and 102 samples. Out of these 102 samples: 52 are tumour samples and 50 are non-tumours.

B. Experimental Setup

All the involved algorithms are implemented in R 3.5.1. Kuncheva Index (KI) is used to measure the stability of different Feature Selection techniques. To assess the stability of the proposed method, independent training and validation sets have been generated from the aforementioned datasets. The training dataset has undergone a slight variation of 10% to measure the stability. The random sampling of training dataset is repeated 10 times with 90% of overlap, and KI is averaged over all samples. The KI value has been scaled from its original [-1, 1] range to [0, 1].

C. Results

To study the stability of our proposed method, IG-RFE-SVM, it is implemented and its stability is compared with varying subset size. It is also compared with other three FS technique: Correlation based FS (CFS) and Relief method and feature Selection based on Random Forest. The results are shown below:

Table- II: Comparison of stability of various Feature Selection methods with varying subset size based upon the KI on Colon Cancer dataset

S.No.	Subset Size	Kuncheva Index				
		IG-RFE-SVM	CFS	RELIEF	Random Forest	SVM-RFE
1	10	0.80	0.75	0.65	0.54	0.70
2	20	0.80	0.77	0.65	0.62	0.70
3	30	0.81	0.78	0.64	0.66	0.73
4	40	0.79	0.75	0.65	0.63	0.78
5	50	0.80	0.74	0.66	0.64	0.78
6	100	0.83	0.75	0.66	0.69	0.81
7	200	0.84	0.76	0.67	0.66	0.82
8	500	0.84	0.78	0.61	0.59	0.82
9	800	0.87	0.78	0.60	0.65	0.81
10	1,000	0.98	0.79	0.58	0.87	0.80

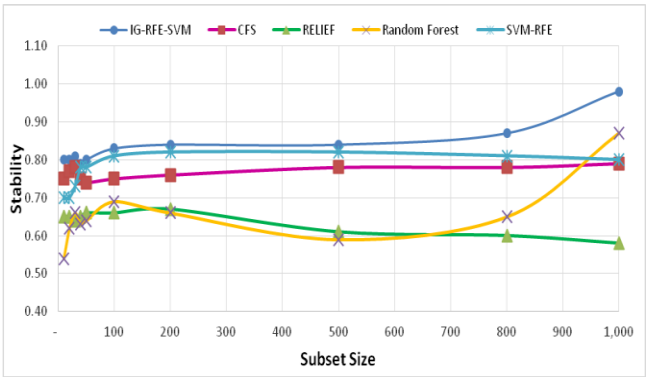

Fig. 2: Comparison of stability of various feature selection methods with varying subset size based upon the KI on Colon Cancer dataset.

Table- III: Comparison of stability of various feature selection methods with varying subset size based upon the KI on Leukaemia dataset

S.No.	Subset Size	Kuncheva Index				
		IG-RFE-SVM	CFS	RELIEF	Random Forest	SVM-RFE
1	10	0.79	0.75	0.73	0.75	0.75
2	20	0.76	0.76	0.72	0.77	0.76
3	30	0.76	0.76	0.71	0.78	0.73
4	40	0.81	0.78	0.69	0.81	0.79
5	50	0.81	0.78	0.68	0.78	0.80
6	100	0.82	0.79	0.68	0.77	0.86
7	200	0.83	0.79	0.66	0.72	0.88
8	500	0.86	0.81	0.69	0.64	0.89
9	1,000	0.86	0.81	0.68	0.70	0.87

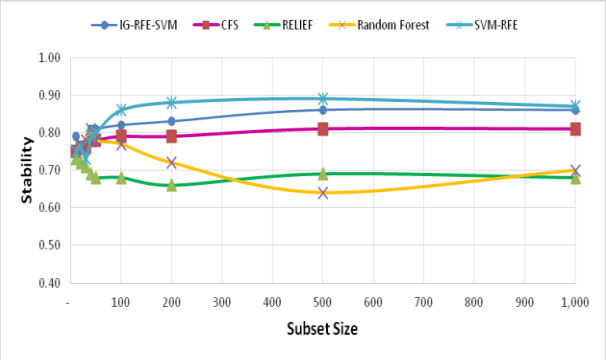

Fig. 3: Comparison of stability of various feature selection methods with varying subset size based upon the KI on Leukaemia dataset

Table- IV: Comparison of stability of various feature selection methods with varying subset size based upon the KI on Prostate Cancer dataset

S.No	Subset Size	Kuncheva Index				
		IG-RFE-SVM	CFS	RELIEF	Random Forest	SVM-RFE
1	10	0.80	0.79	0.67	0.75	0.80
2	20	0.85	0.77	0.67	0.77	0.89
3	30	0.89	0.75	0.65	0.79	0.88
4	40	0.89	0.76	0.65	0.80	0.88
5	50	0.89	0.77	0.64	0.78	0.84
6	100	0.86	0.80	0.65	0.73	0.85
7	200	0.90	0.81	0.66	0.70	0.89
8	500	0.91	0.81	0.67	0.65	0.88
9	1,000	0.91	0.82	0.70	0.60	0.88

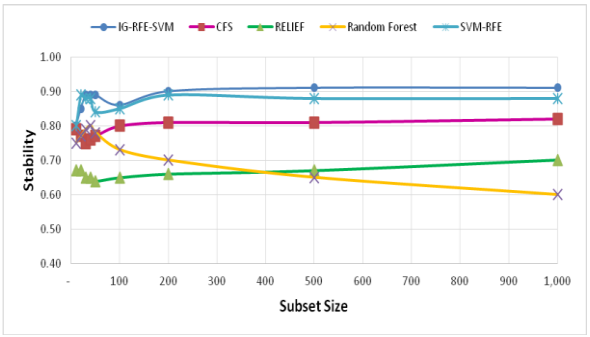

Fig. 4: Comparison of stability of various feature selection methods with varying subset size based upon the KI on Prostate Cancer dataset

From the above table and graphs, the summarized results can be stated as follows:

1. The resulting data depicts that the stability achieved by proposed method IG-RFE-SVM is better than all other four FS techniques. It has significantly improved the stability of SVM-RFE.
2. In contrast, Relief filter method performs worst on all three gene expression datasets. This method suffers from instability due to randomly selection of instances from same and different class for each feature weight calculation.
3. The graphs show very clearly and significantly that stability increases with increase in subset size.
4. Random Forest FS technique behaves differently with varying subset size. The stability of Random Forest first increases with increase in subset size and then it starts decreasing.

VI. CONCLUSION

Stability plays a crucial role in the performance of any Feature Selection method on high-dimensional gene expression data. Hence, it is essential to identify the stable feature selection method which could efficiently classify the gene expression data. In order to achieve this, we have proposed a fusion of Information Gain and Recursive Feature Elimination method for Support Vector Machine. It is evident from the results that, proposed method IG-RFE-SVM has significantly improved the stability of SVM-RFE. The stability of IG-RFE-SVM is also compared with three other Feature Selection methods: The Relief Algorithm, Correlation based Feature Selection (CFS) and Random Forest FS technique. The experiments are successfully performed on three high dimensional gene expression datasets and compared the stability of five Feature Selection methods on the basis of Kuncheva Index stability measure. Also, the stability is measured with varying subset size selected for feature selection. With the outcomes, it is apparent that the IG-RFE-SVM method is superior to other methods on the stability criterion. From the results, it can also be concluded that stability of these methods generally increases with subset size.

REFERENCES

1. David L. Donoho, “High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality,” AMS Math Challenges Lecture, 2000.
2. W. Wang, J. Yang, “Mining High-Dimensional Data”, The Data Mining and Knowledge Discovery Handbook, Springer Heidelberg, pp. 93–799, 2005.
3. B. Goldstein, R. Shaft, “When is “nearest neighbor” meaningful?” Proceedings of International Conference on Database Theory, 1999.
4. Sun-Yuan Kung and Man-Wai Mark, “Feature Selection for Genomic and Protemic Data Mining”, Machine Learning in Bioinformatics, Wiley, 2009.
5. Z.M. Hira, D.F. Gillies, “A review of feature selection and feature extraction methods applied on microarray data, Advances in Bioinformatics”, ID 198363, 1-13, 2015.
6. Artur J. Ferreira, Mario T. Figueiredo, “Efficient feature selection filters for high-dimensional data”, Pattern Recognition Letters, Vol. 33(13), pp. 1794-1804,2012.

7. V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, J.M. Benitez, F.Herrera, "A review of microarray datasets and applied feature selection methods", Information Science 282, pp. 111–135, 2014.
8. S. Nogueira, K. Sechidis, and G. Brown, "On the Stability of Feature Selection Algorithms," Journal of Machine Learning Research, vol. 18(174), pp 1−54, 2018.
9. S. Nogueira and G. Brown, "Measuring the stability of feature selection with applications to ensemble methods," in Proc. Int. Workshop Multiple Classifier Syst., 2015, pp. 135–146.
10. P. Mohana and K. Perumal, "A Survey on Feature Selection Stability Measures," International Journal of Computer and Information Technology, vol. 5, no 1, pp. 98–103, Jan 2016.
11. I. Guyon, J.Weston, S. Barnhill, and V. Vapnik. "Gene selection for cancer classification using support vector machines", Machine Learning, 46.1-3, pp. 389–422, 2002.
12. P. Mundra and J. Rajapakse. "SVM-RFE with MRMR filter for gene selection", IEEE Transactions on NanoBioscience, 9,1 pp. 31–37, 2010.
13. M. Kabir, M. Islam, K. Murase, "A new wrapper feature selection approach using neural network", Neurocomputing, Vol. 73, No. 16-18, pp. 3273 –3283, 2010.
14. J. Leng, C. Valli, and L. Armstrong. "A wrapper-based feature selection for analysis of large data sets". Proceedings of 2010 3rd International Conference on Computer and Electrical Engineering (ICCEE 2010), Chengdu, China. IEEE, pp. 167–170, 2010.
15. K. Kira, L. Rendell. "The feature selection problem: traditional methods and a new algorithm." Proceedings of the Association for the Advancement of Artificial Intelligence Conference, Cambridge, MA, USA: AAAI Press and MIT Press, pp. 129–134, 1992.
16. M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning", Proceedings of the International Conference on Machine Learning (ICML), Morgan Kaufmann, pp. 359–366, 2000.
17. M. Hall, "Correlation-based feature selection for machine learning", PhD thesis. Hamilton, New Zealand: Waikato University, Department of Computer Science, 1998.
18. L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy". Journal of Machine Learning Research (JMLR), 5, pp. 1205–1224, Dec. 2004.
19. L. Yu, H. Liu. "Feature selection for high-dimensional data: a fast correlation based filter solution", Proceedings of the International Conference on Machine Learning (ICML), pp. 856–863, 2003.
20. H. Peng, F. Long, C. Ding. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 27, 8, pp. 1226–1238, 2005.
21. Chris Ding, H. Peng, "Minimum redundancy feature selection from microarray gene expression data", Proceedings of the Computational Systems Bioinformatics (CSB'03), pp. 523–529, 2003.
22. B. Franay, G. Doquire, M. Verleysen, "Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification", Neurocomputing, 112.0, pp. 64 –78, 2013.
23. C. Lazar, J. Taminau, S. Meqanck, "A survey on filter techniques for feature selection in gene expression microarray analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9, pp. 1106–1119, 2012.
24. S. Jungjit and A.A. Freitas, "A New Genetic Algorithm for Multi-Label Correlation-Based Feature Seletion", Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN2015), 22-24 April 2015.
25. S. Jungjit and A.A. Freitas, "A Lexicographic Multi-Objective Genetic Algorithm for Multi-Label Correlation-Based Feature Selection", Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation, ACM, pp no 989-996, 2015
26. M. Esseghir, H. Liu, H. Motoda, R. Setiono, Z. Zhao. "Effective wrapper-filter hybridization through GRASP schemata". Proceedings of the MLRWorkshop and Conference, ACM Press, Vol. 10, pp. 45–54, 2010.
27. P. Bermejo, J. Gámez, J. Puerta, "A GRASP algorithm for fast hybrid (filterwrapper) feature subset selection in high-dimensional datasets", Pattern Recognition Letters 32, 5, pp. 701–711, 2011.
28. Z. Zhu, Y. Ong, M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework", IEEE Transactions on Systems, Man, and Cybernetics, Part B 37.1, pp. 70–76, 2007.
29. S. Kannan and N. Ramaraj. "A novel hybrid feature selection via symmetrical uncertainty ranking based local memetic search algorithm", Knowledge-Based Systems, 23, 6, pp. 580 –585, 2010.
30. J. Xie, W. Xie, C. Wang, and X. Gao. "A novel hybrid feature selection method based on IFSFFS and SVM for the diagnosis of erythemato-squamous diseases", Journal of Machine Learning Research (JMLR) - Proceedings Track 11, pp. 142–151, 2010.
31. Jyoti Ahuja, Saroj Dahiya Ratnoo, "Feature Selection using Multi-Objective Genetic Algorithm: A Hybrid Approach", INFOCOMP, Vol. 14, No. 1, pp. 26-37, June 2015.
32. A. Kalousis, J. Prados, and M. Hilario, "Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces," Knowledge and Information Systems, vol. 12, no. 1, pp. 95-116, 2007
33. Randall Wald, Taghi Khoshgoftaar, Amri Napolitano, "Comparison of Stability for different Families of Filter-Based and Wrapper-Based Feature Selection", 12th International Conference on Machine Learning and Applications, 2013.
34. L.I. Kuncheva, "A stability index for feature selection.", Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and applications, Innsbruck, Austria, pp. 390-395, 2007.
35. K. Dunne, P. Cunningham, F. Azuaje, "Solutions to instability problems with sequential wrapper-based approaches to feature selection", Technical Report., Journal of Machine Learning Research, 2002.
36. https://github.com/ramhiser/datamicroarray/wiki/

AUTHORS PROFILE

ShavetaTatwani received her M.Tech in Computer Science and Engineering in 2004 from Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India. From 2004 till 2018, she worked as Assistant Professor at Amity School of Engineering and Technology, New Delhi. She has 14+ years of rich experience in teaching various computer science subjects to B.Tech graduates and guiding 20+ major projects in the field of Big Data, Grid Computing, and Cloud Computing etc. Also, she worked as research scholar at Technical University of Munich, Germany for one year. There, working in the area of High Performance Computing, she got hands-on experience in Parallel Programming Tools. Currently, Shaveta is pursuing PhD in the field of Machine Learning Techniques from Indira Gandhi Delhi Technical University of Woman, Delhi. Her current research interests are Machine Learning Techniques, Big Data Analysis and High Performance Computing.

Ela Kumar is a HOD (Department of Computer Science and Engineering), Dean (Student affairs) and Chief Proctor at Indira Gandhi Delhi Technical University for Woman, New Delhi, India. She has completed her PhD from University of Delhi in 2003 and done M.Tech (Computer Science and Technology) from IIT Roorkee in 1990. In her 29 years of experience, she has guided about 7 PhD students, 43 M.Tech students and more than 50 B.Tech projects in the field of Web Analytics, Information Retrieval and Artificial Intelligence. She is member of various international and national academic and professional bodies. She is member of Doctoral Research Committee of Delhi Technological University, Birla Institute of Technological and Management, and Amity University. She has authored 5 books in field of Artificial Intelligence, Natural Language Processing, Knowledge Engineering and related areas. She has published more than 100 research papers in national/international journals. She is also member of Editorial Board of few technical journals.