ISSN 2007-9737

On the Performance Assessment and Comparison of Features Selection Approaches

Seyyid Ahmed Medjahed[1,*], Fatima Boukhatem[2]

[1] University of Relizane,
Algeria

[2] University Djillali Liabes,
Algeria

seyyidahmed.medjahed@univ-relizane.dz, fatima.boukhatem@univ-sba.dz

Abstract. In many supervised learning problems, feature selection techniques are increasingly essential across various applications. Feature selection significantly influences the classification accuracy rate and the quality of SVM model by reducing the number of features, remove irrelevant and redundant features. In this paper, we evaluate the performance of twenty feature selection algorithms over four databases. The performance is conducted in term of: classification accuracy rate, Kuncheva's Stability, Information Stability, SS Stability and SH Stability. To measure the feature selection algorithms, multiple datasets from the UCI Machine Learning Repository are utilized to assess both classification accuracy and stability variations.

Keywords. Feature selection, classification, stability, support vector machine.

1 Introduction

In recent years, the motivation behind applying feature selection techniques has evolved significantly. What was once merely an illustrative example has now become a crucial prerequisite for effective model building. This shift in emphasis can be attributed to several factors, including improved generalization performance, reduced running time requirements, and the need to address constraints and interpretational challenges inherent in the problem domain.

Feature selection is a vital dimensionality reduction technique in data mining, involving the selection of a subset of original features based on specific criteria.

This process is important and commonly utilized to enhance the efficiency and effectiveness of data analysis tasks [1, 2, 3].

It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications: speeding up a data mining algorithm, and improving mining performance such as predictive accuracy and result comprehensibility.

Therefore, it is essential to employ an effective feature selection method that considers the number of features used for sample classification to enhance processing speed, predictive accuracy, and comprehensibility.

The correlation between features significantly impacts classification outcomes. Removing important features can reduce classification accuracy and negatively affect the quality of SVM models.

Similarly, certain features may have no discernible effect or may be laden with high levels of noise [4]. Their removal increases the classification accuracy rate.

The aim of feature selection is to find the smallest feature subset that increases the classification accuracy rate.

The optimal features subset is not unique; it may be possible to achieve the same accuracy rate using different sets of features, because if two features are correlated one can replace by other.

Note that feature subset selection chooses a set of features from existing features, and does not construct new ones; there is no feature extraction or construction [5, 6].

doi: 10.13053/CyS-28-2-4211

Table 1. Some feature selection criteria and algorithms

Methods	Full Name
MRMR	Max-Relevance Min-Redundancy [20,18]
CMIM	Conditional Mutual Info Maximisation [13,18]
JMI	Joint Mutual Information [14,18]
DISR	Double Input Symmetrical Relevance [15,18]
CIFE	Conditional Infomax Feature Extraction [16,18]
ICAP	Interaction Capping [17,18]
CONDRED	Conditional Redundancy [18]
BETAGAMMA	BetaGamma [18]
MIFS	Mutual Information Feature Selection [19,18]
CMI	Conditional Mutual Information [18]
MIM	Mutual Information Maximisation [12,18]
RELIEF	Relief [18]
FCBF	Fast Correlation Based Filter [21,27]
MRF	Markov Random Fields [26]
SPEC	Spectral [22,27]
T-TEST	Student's T-test [27]
KRUSKAL-WALLIS	Kruskal-Wallis Test [23,27]
FISHER	Fisher Score [24,27]
GINI	Gini Index [25,27]
GA	Genetic Algorithm

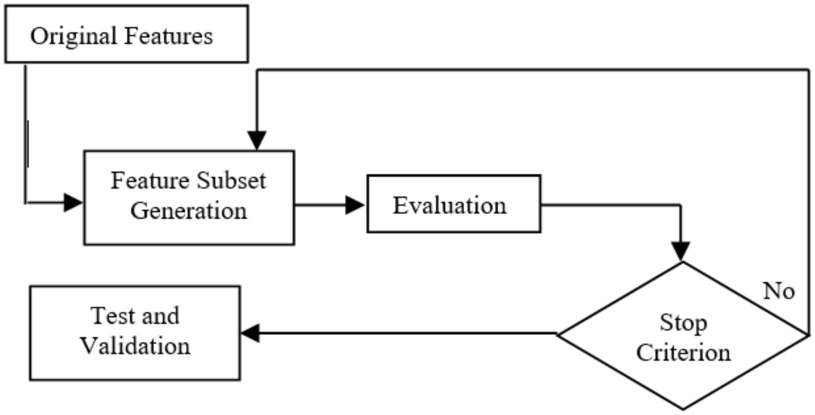

Fig. 1. A unified view of feature selection process

In this study, we analyze and evaluate the performance of several feature selection techniques (20 algorithms) by using the criterion of stability and the classification accuracy rate calculates with SVM-SMO.

The experimentation is conducted over 4 datasets obtained from UCI machine learning repository.

The paper is organized as follows. In section 2, we give an overview of SVM.

Table 2. Datasets from the UCI ML repository

Datasets	Number of classes	Number of instances	Number of features
Breast Cancer	2	699	9
Cardiotocography	2	1831	21
ILPD	2	583	9
Mammographic Mass	2	961	5

Table 3. Number of instances used for training and testing steps

Datasets	Missing instances	Training set	Testing set
Breast Cancer	16	411	272
Cardiotocography	0	1101	730
ILPD	0	351	232
Mammographic Mass	131	500	330

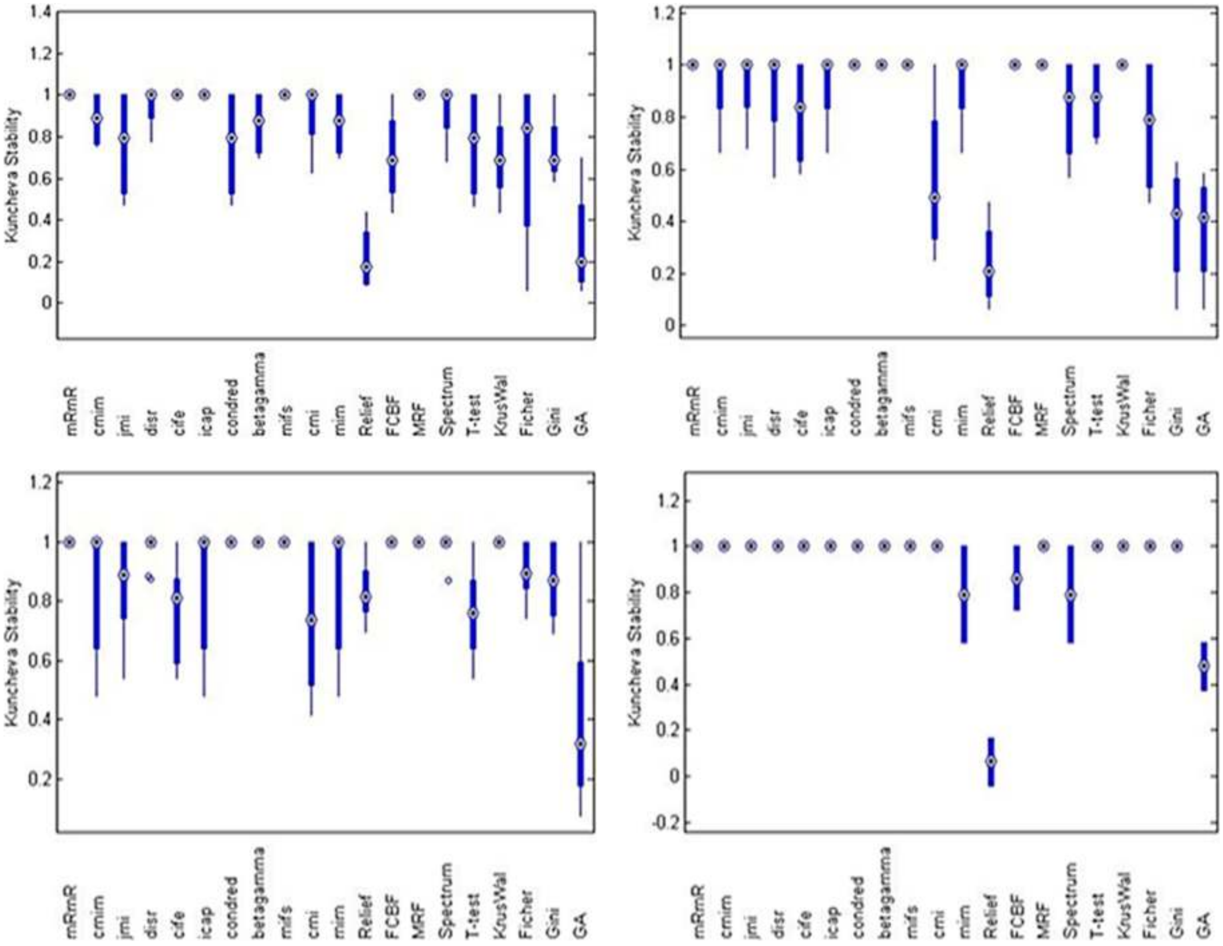

Fig. 2. Kuncheva's stability over the 4 data sets. The box indicates the upper and the lower quartiles. The small circle shows the median values, while the blue line indicates the maximum and the minimum values

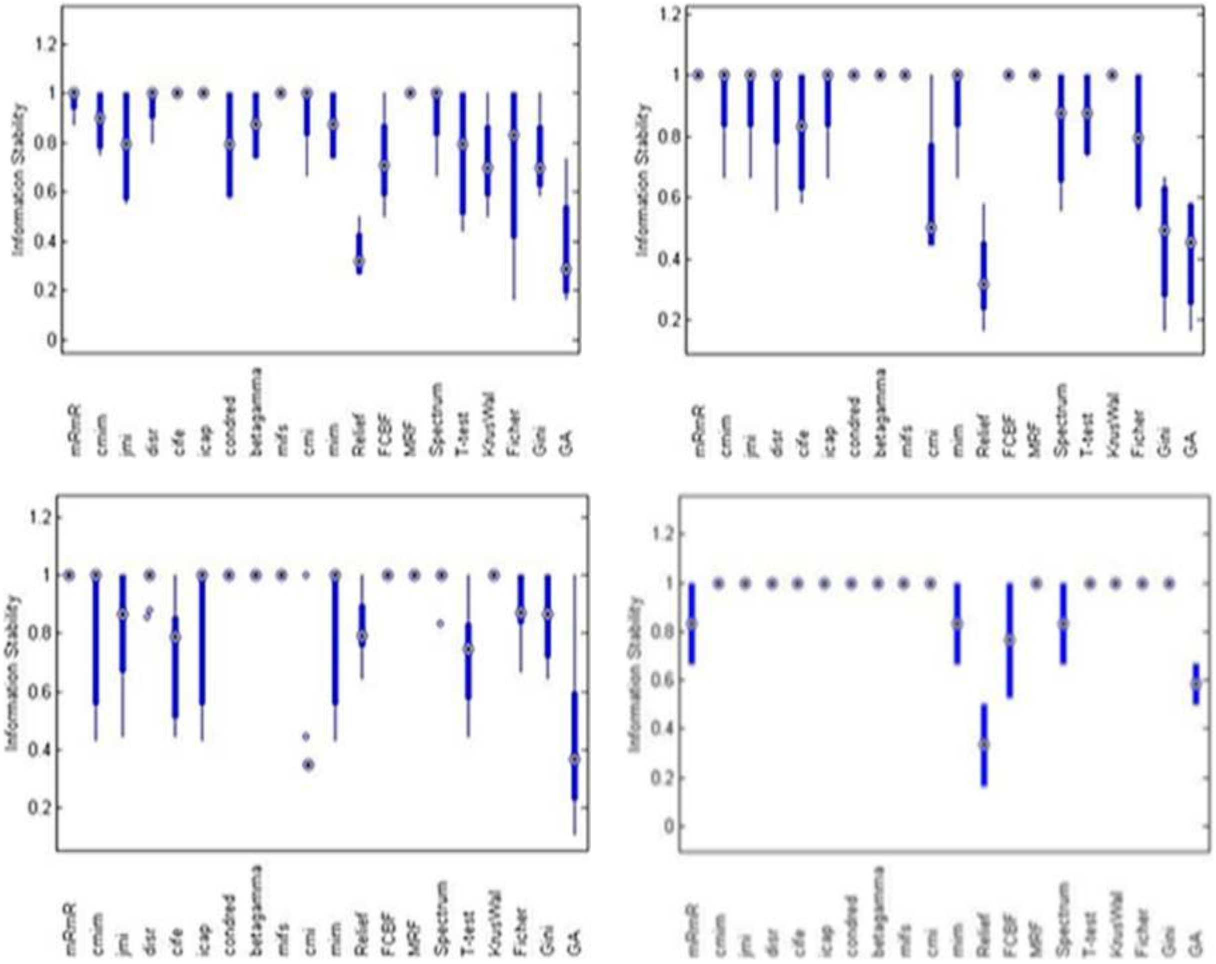

Fig. 3. Information stability over the 4 data sets. The box indicates the upper and the lower quartiles. The small circle shows the median values, while the blue line indicates the maximum and the minimum values

In section 3, we describe the stability criteria used in the literature. In section 4, we discuss the different feature selection techniques.

Section 5 describes the results obtained by the approaches. Finally, concluding remarks are made in section 6.

2 Overview of Support Vector Machine

SVM can be briefly described as follows [7, 8, 9]. Consider $(x_1, y_1), \cdots, (x_n, y_n)$ with $y\{-1, +1\}$ denote a set of training data. The goal of Support Vector Machines (SVM) is to create a separating hyperplane in the attribute space that maximizes the margin between instances of different classes. This task involves reformulating the classification problem into a quadratic optimization problem aimed at finding the optimal hyperplane:

$$\begin{gathered}\min_{\alpha} -\sum_{i=1}^{N}\alpha_i + \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle, \\ \text{s. c.} \sum_{i=1}^{N} \alpha_i y_i = 0, \\ \forall i \in \{1, \dots, N\}, \qquad \alpha_i \geq 0. \end{gathered} \tag{1}$$

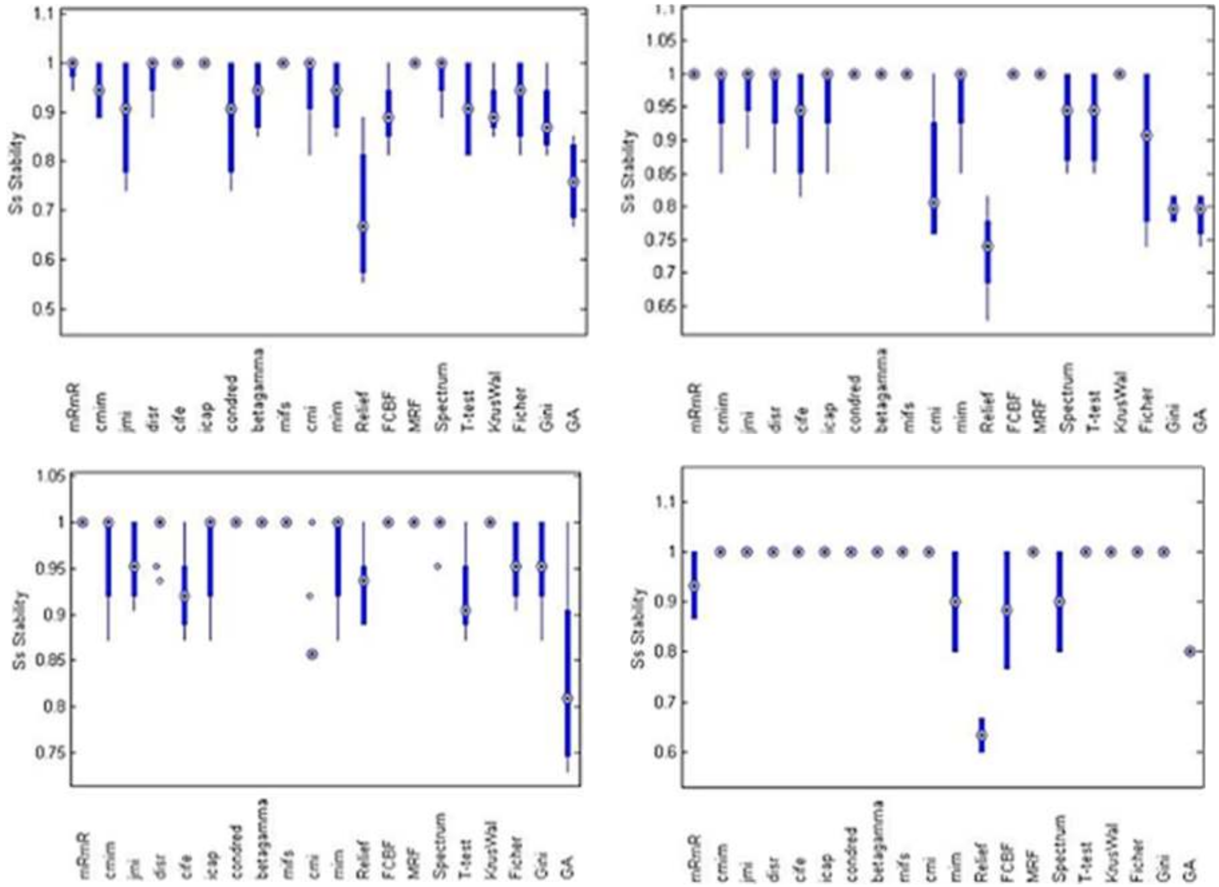

Fig. 4. SS stability over the 4 data sets. The box indicates the upper and the lower quartiles. The small circle shows the median values, while the blue line indicates the maximum and the minimum values

This is the dual form of the quadratic problem, C represents the regularization parameter. To solve the optimization problem in Support Vector Machines (SVM), quadratic optimization algorithms are utilized.

Some commonly used algorithms include: Sequential Minimal Optimization [10, 11], Trust Region, etc. By solving the optimization problem, we determine the Lagrange multipliers, the optimal hyperplane is given by:

$$\begin{aligned} w^* &= \sum_{i=1}^{N} \alpha_i y_i x_i, \\ b^* &= -\frac{1}{2}\langle w^*, x_r + x_s \rangle, \\ H(x) &= \mathrm{sign}(\langle w^*, x\rangle + b^*), \end{aligned} \tag{2}$$

where $\alpha_r, \alpha_s > 0$, $y_r = -1$, $y_s = 1$.

3 Feature Selection Algorithm

Feature selection is a domain garnering growing attention within the realm of machine learning. Numerous feature selection techniques have been outlined in literature dating back to the 1970s.

Feature selection algorithms are categorized into three main types based on their strategies: filter, wrapper, and embedded models.

Filter feature selection methods do not consider classifier properties; instead, they conduct statistical tests on variables. In contrast, wrapper feature selection evaluates various feature sets by constructing classifiers.

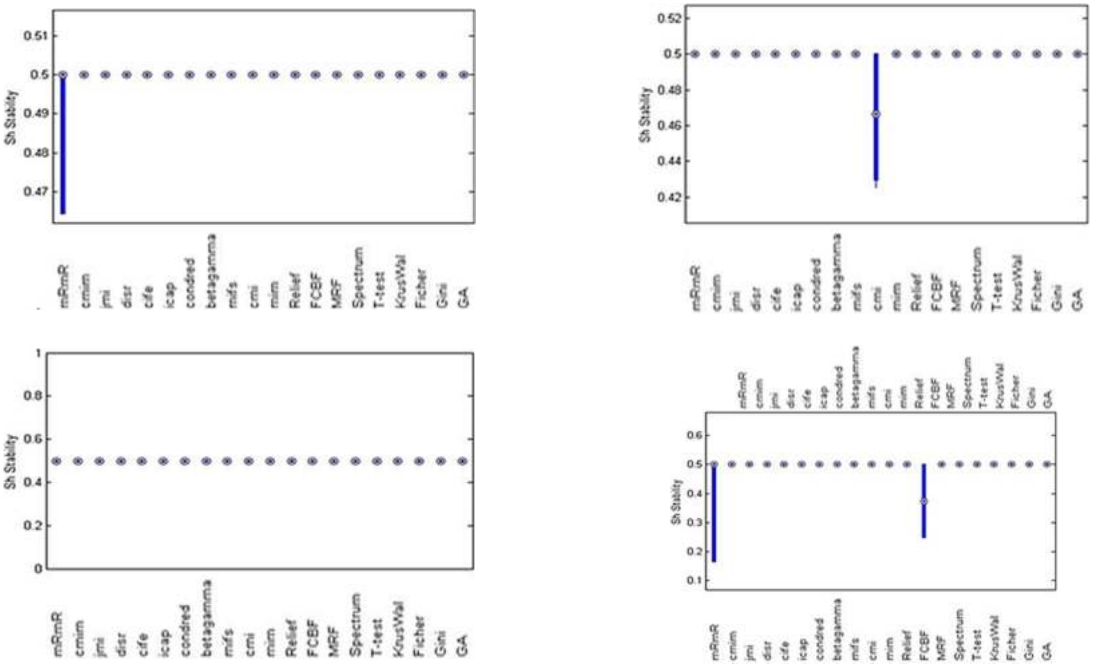

Fig. 5. S_H stability over the 4 data sets. The box indicates the upper and the lower quartiles. The small circle shows the median values, while the blue line indicates the maximum and the minimum values

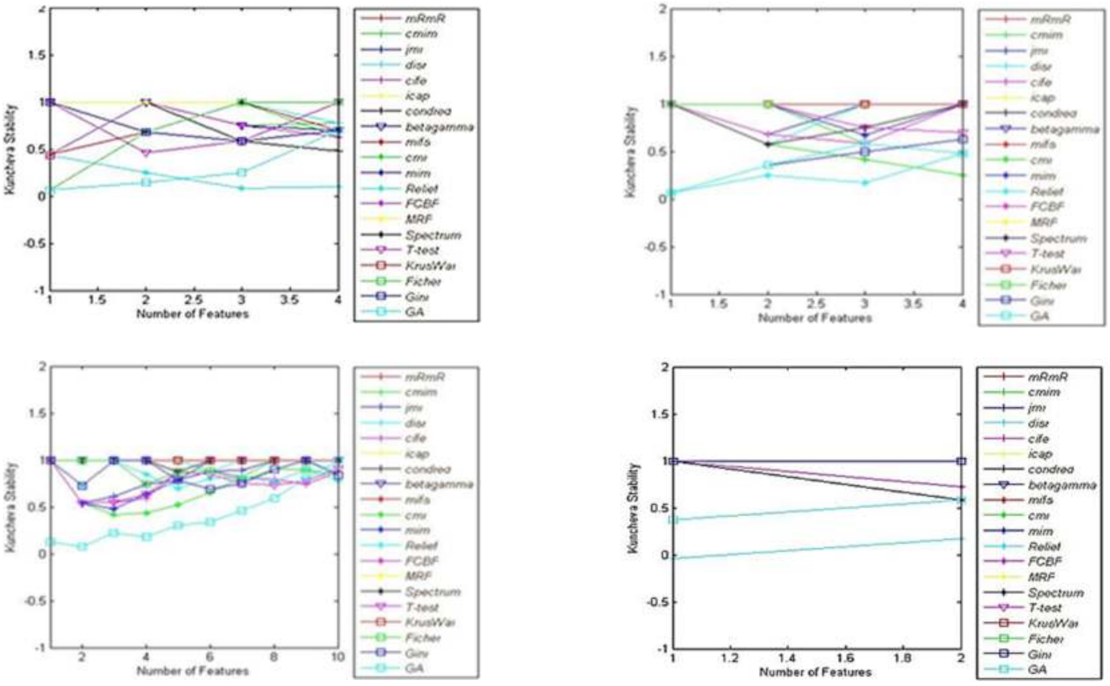

Fig. 6. Kuncheva's stability over the 4 data sets for each number of selected features

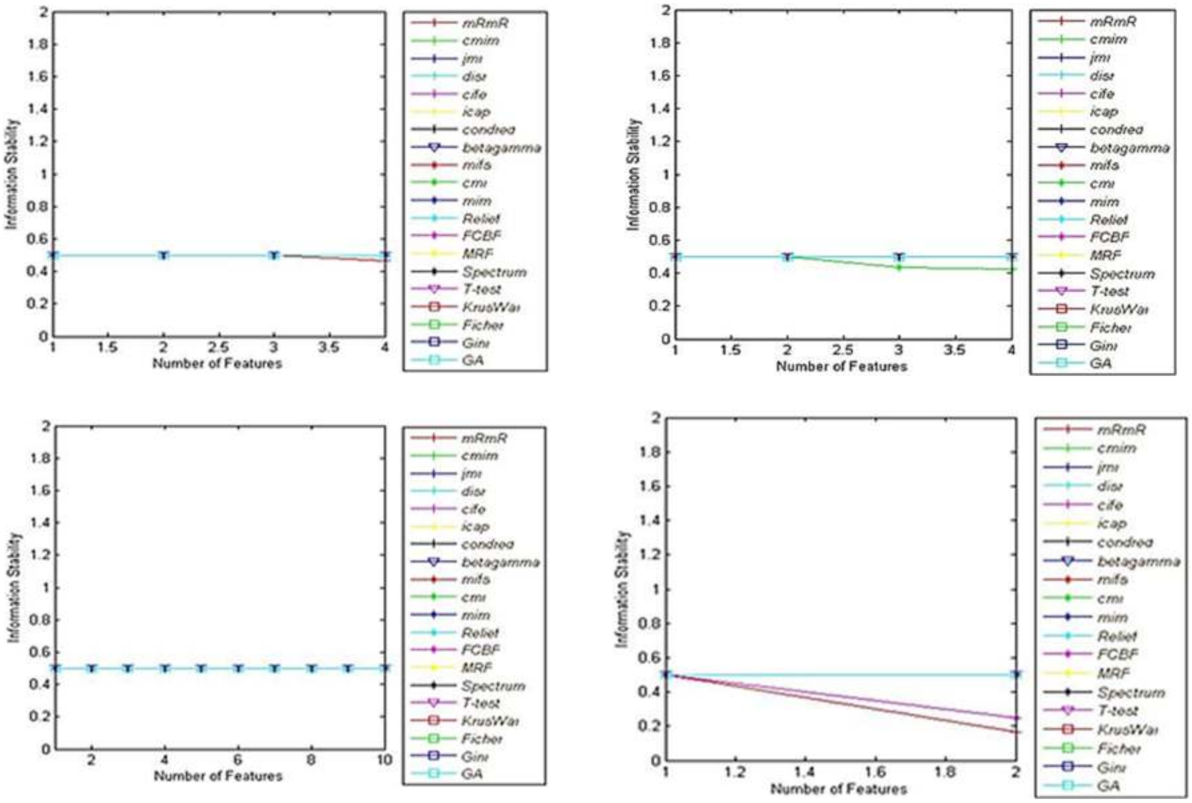

Fig. 7. Information stability over the 4 data sets for each number of selected features

Embedded model algorithms integrate variable selection into the training process, deriving feature relevance analytically from the learning model's objective. Table 1 summarizes some feature selection criteria and algorithms.

4 Stability of Feature Selection Algorithm

The stability of a feature selection algorithm refers to how sensitive it is to changes in feature preferences or rankings. It quantifies how different training set affect the feature preferences [31]. To calculate the stability, we require a similarity measure for feature preferences: Consider two subsets A and B we denote:

| . | The cardinality.

∩ The union.

U The intersection.

4.1 SS Stability

Kalousis et al. [29] define the similarity index between two subsets, A and B, as:

$$s_s = 1 - \frac{|A| + |B| - 2|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|}. \quad (3)$$

The SS stability is a simple adaptation of the Tanimoto, which measures the similarity distance between two sets A and B. SS takes values in [0,1] with 0 meaning that there is no overlap between the two sets, and 1 that the two sets are identical.

4.2 SH Stability

Dunne et al. [30] calculates the similarity between two subsets by comparing the relative Hamming distance of their corresponding masks. In set notation, this method can be described as follows:

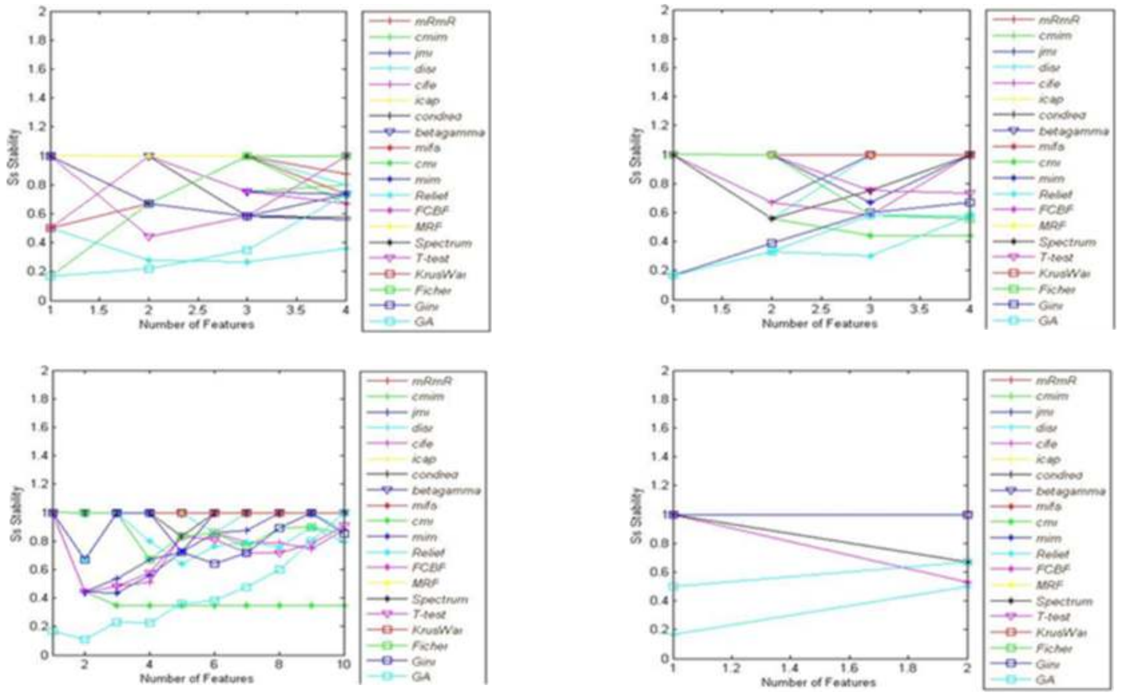

Fig. 8. SS stability over the 4 data sets for each number of selected features

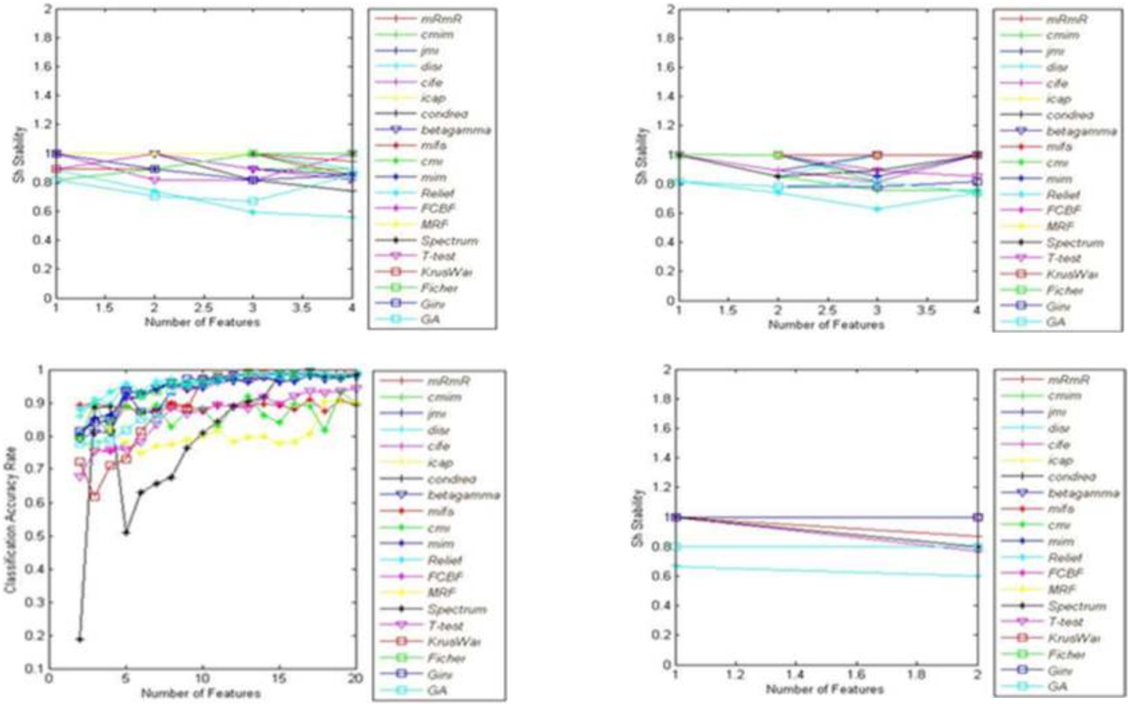

Fig. 9. SH stability over the 4 data sets for each number of selected features

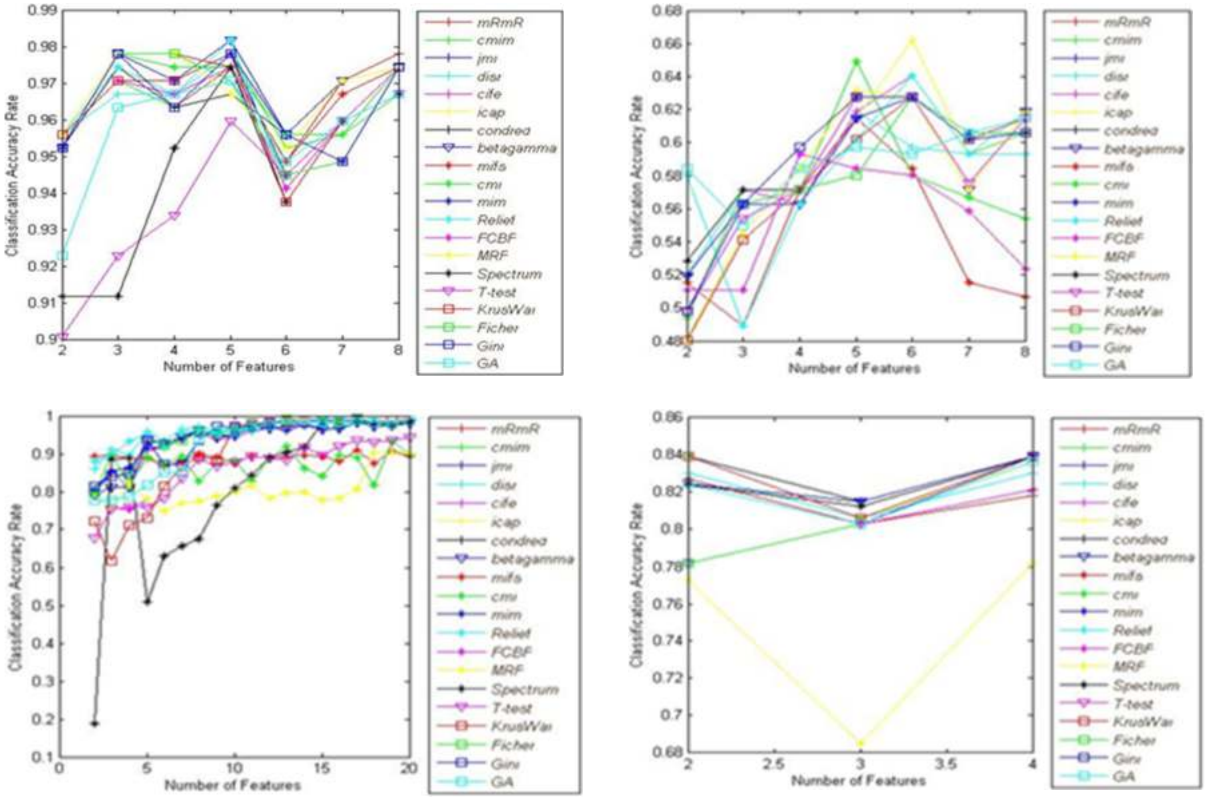

Fig. 10. The classification accuracy rate for each number of features

$$s_H = 1 - \frac{|A \setminus B| + |B \setminus A|}{n}. \quad (4)$$

4.3 Kuncheva Stability

Kuncheva [32] define the consistency index for two subsets with the same cardinality as:

$$I_c = \frac{r - \frac{k^2}{n}}{k - \frac{k^2}{n}} = \frac{rn - k^2}{k(n-k)}, \quad (5)$$

where $k = |A| = |B|$ and $r = |A \cap B|$. The maximum value of the index is $IC = 1$.it mean that $r = k$, and the minimum value is $IC = -1$.

4.4 Information Stability

Lei Yu et al. [33, 34] propose the normalized mutual information as a measure of stability of two feature sets:

$$\mathrm{Sim}(x_a,\ x_b) = \frac{I(x_a, x_b)}{H(x_a) + H(x_b)}. \quad (6)$$

The stability of a set of sequences features, $F = \{S_1, S_2, \dots, S_K\}$ is the average of all pairwise.

5 Experimental Results

In this section, we have made a comparison protocol between the several feature selections techniques defined in the literature and shown the performance of each technique.

The experiment is analyzed by using the following performance measures: classification accuracy rate calculated by using the support vector machine. Also, we use the stability criteria: Kuncheva stability, Information stability, SS and SH stability. Table 2 presents a summary of four selected datasets used in the feature selection experiment:

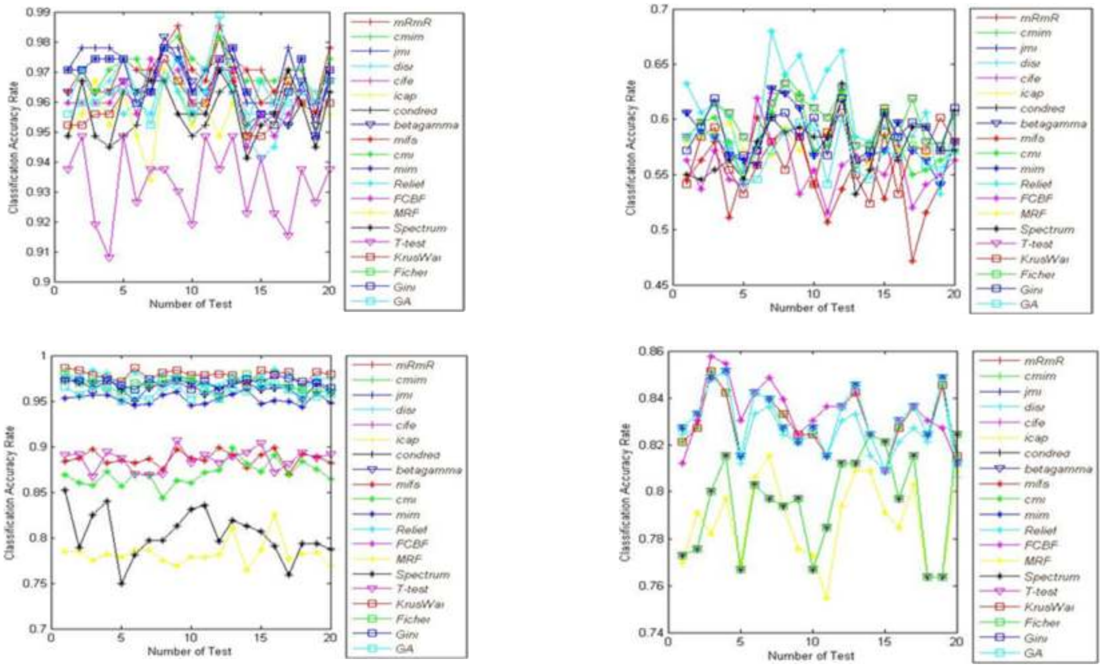

Fig. 11. The classification accuracy rate for each training set over the 4 better features

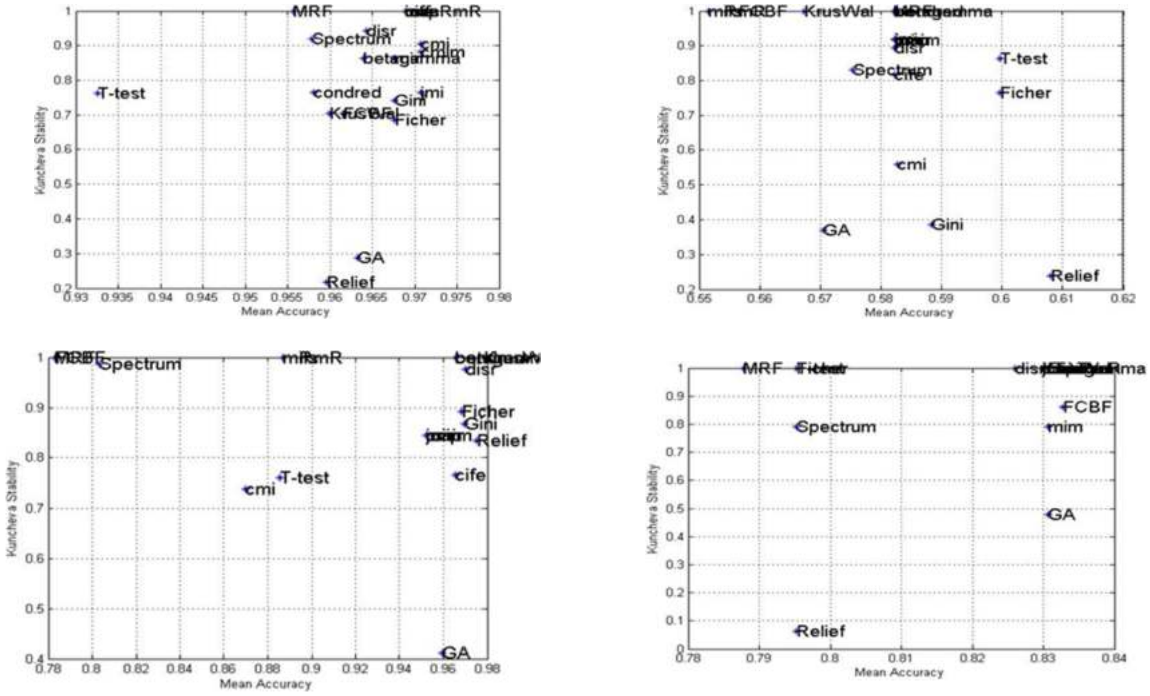

Fig. 12. Kuncheva's stability versus the average classification accuracy rate over 20 different training sets for each dataset

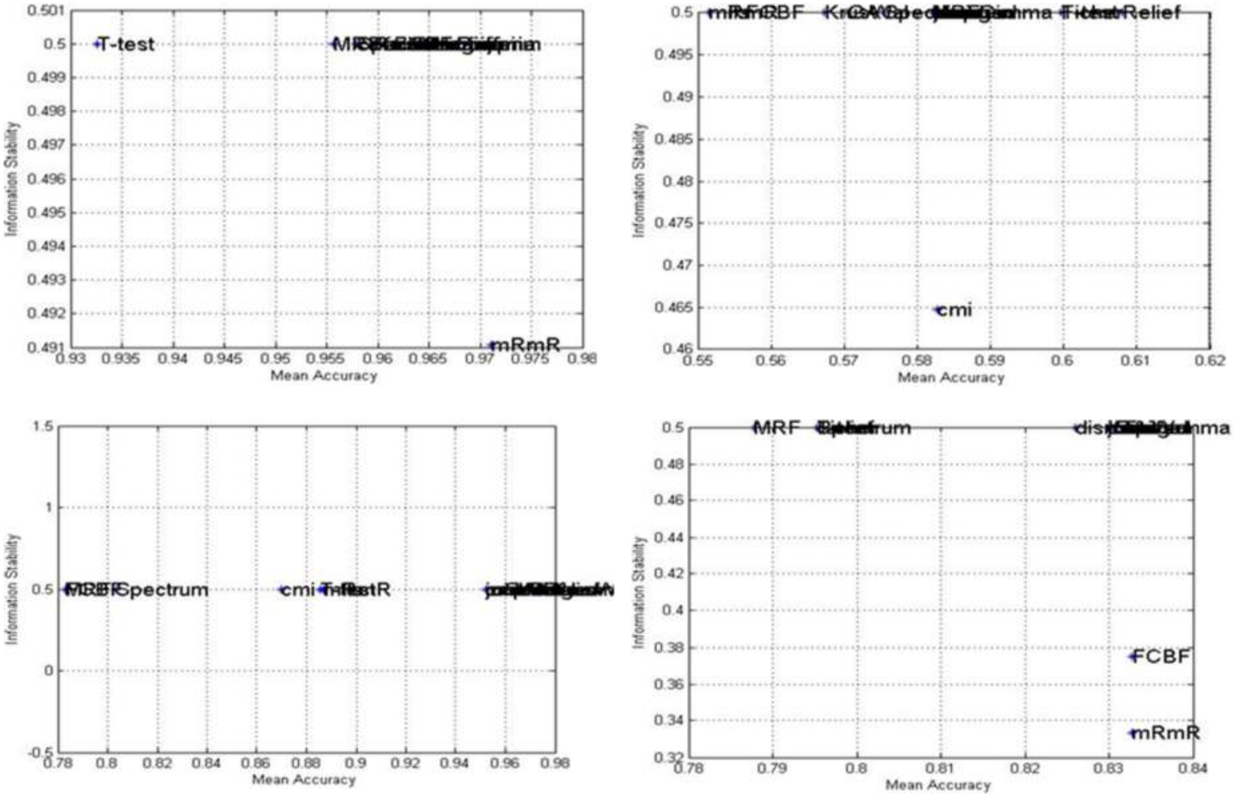

Fig.13. Information stability versus the average classification accuracy rate over 20 different training sets for each data

WDBC (Wisconsin Dataset Breast Cancer), Cardiotocography, ILPD (Indian Liver Patient Dataset), and Mammographic Mass. The performance evaluation of feature selection techniques requires the determination of the training and testing set.

In this study, we split randomly the initial dataset by using the hold out method which is a king of cross validation. In this experiment, less than one-third of the initial data is allocated for testing purposes. Specifically, 60% of the instances are designated for training, while the remaining 40% are reserved for testing.

Table 3 outlines the number of instances utilized during both the training and testing phases for each dataset. To compare the feature selection criteria defined above, we proceed as follows: for each data set, we select different training set and we take a set of features for each training set by using each feature selection criterion.

The following figures 2,3,4,5 show the Kuncheva's Stability, Information Stability, SS Stability and SH Stability measures over 4 datasets for each feature selection criterion. For each data set we calculate the stability for different training set obtained by using the hold out method which selects randomly a training set.

10 training sets are selected for each data set, we use this principle to better exploit each dataset. The results show that for all the training set which are selected randomly for each data sets, all the methods are stable except GA, CMI, T-test, Fisher, Gini, and relief.

The stability for JMI, MRMR, Disr, Condred, Mifs, FCBF, MRF and Kruskal-Wallis is equal to 1 for all the datasets, this means that these methods have select the same subset of feature for each training set of the four datasets. Therefore, theses feature selection criterions have selected the relevant subset of feature.

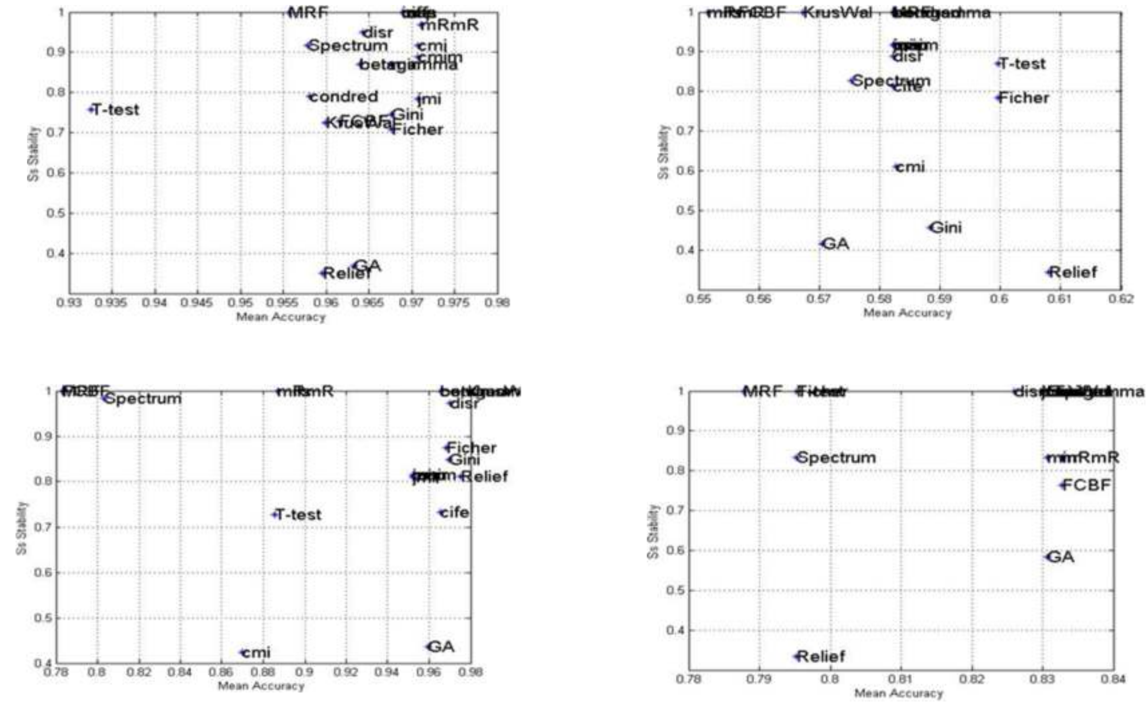

Fig. 14. SS stability versus the average classification accuracy rate over 20 different training sets for each data set

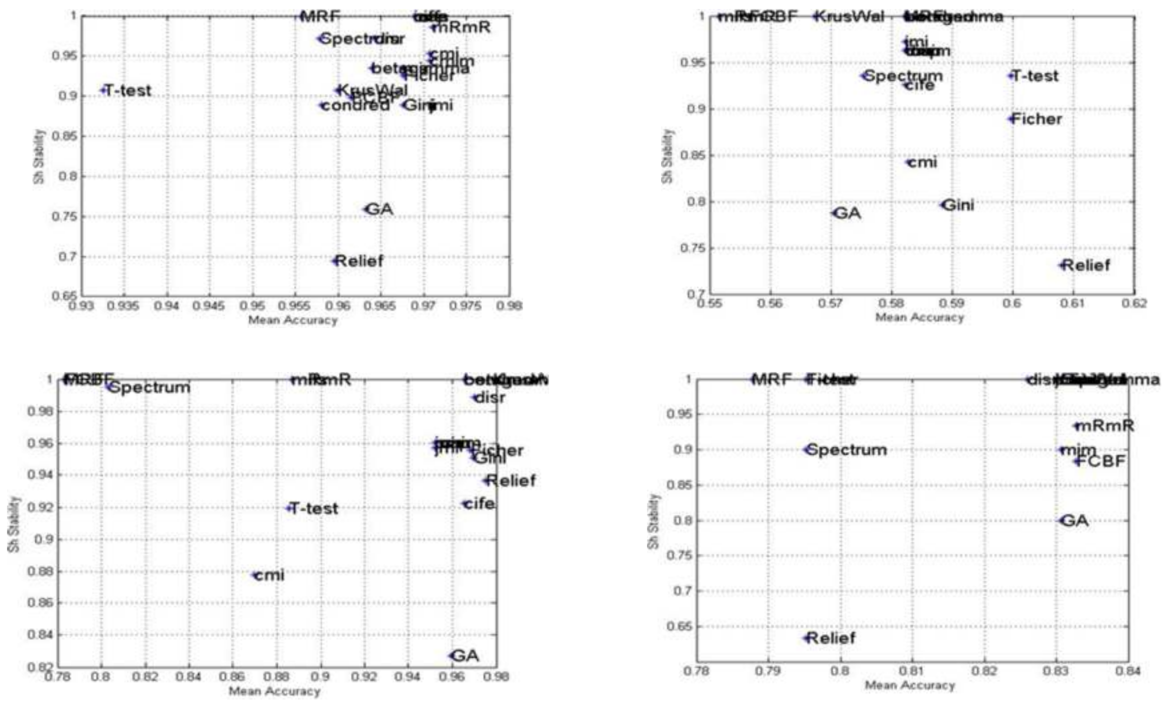

Fig. 15. SH stability versus the average classification accuracy rate over 20 different training sets for each dataset

Table 4. Average classification accuracy rate for each data set. Filled cell represents the higher accuracy rate

Methods	Average classification accuracy rate (%)			
	WDBC	ILPD	Cardio	Mammo
MRMR	97,11	55,15	88,68	83,27
CMIM	97,07	58,24	95,21	82,98
JMI	97,07	58,22	95,21	82,98
DISR	96,43	58,22	97,00	82,60
CIFE	96,89	58,22	96,55	82,98
ICAP	96,89	58,24	95,21	82,98
CONDRED	95,80	58,24	96,55	82,98
BETAGAMMA	96,36	58,24	96,55	83,67
MIFS	96,89	55,15	88,68	83,27
CMI	97,07	58,26	87,00	82,98
MIM	96,76	58,24	95,21	83,07
RELIEF	95,95	60,80	97,50	79,53
FCBF	96,15	55,64	78,28	83,27
MRF	95,56	58,20	78,28	78,77
SPEC	95,79	57,53	80,33	79,53
T-TEST	93,25	59,95	88,56	79,53
KRUSKAL-WALLIS	95,99	56,73	97,85	82,98
FISHER	96,76	59,95	96,82	79,53
GINI	96,76	58,83	96,95	83,07
GA	96,32	57,05	95,96	83,07

5.1 Comparison and Discussion

The figures 2,3,4,5 show the stability criteria for each feature selection techniques over the four datasets. The results show that MRMR, JMI, DISR, CIFE, ICAP, CONDRED, KRUSKAL-WALLIS and MRF have a stability value around 1.

This means that these feature selection criterions have selected the same feature selection for all the training sets in each data set. Figures 6,7,8,9 illustrate the stability criteria versus the number of features.

The analysis of the results indicates that the Relief, GA, and Ficher methods exhibit lower stability (measured by Kurcheva's, Information, SS, and SH metrics) across all datasets. Therefore, we conclude that these techniques are instable compared to the MRMR, JMI, DISR, CIFE, ICAP, CONDRED, KRUSKAL-WALLIS and MRF which have given an average stability close to 1.

The classification accuracy rate represents an important term to evaluate the performance of feature selection techniques.

In the figure 10 describes the classification accuracy rate for each number of features obtained by each feature selection criterions.

In term of classification accuracy rate, we show clearly that the both Spectrum and MRF methods have provided the lower classification accuracy rate. The higher accuracy rate for the WDBC data set is reached by the both JMI and MIM methods with 5 features.

For the ILPD dataset, we record the high accuracy for the CMIM and CIFE methods with 6 features. In the Cardiotocography data set, the high classification accuracy rate is achieved with Fisher score by using 13 features. For the Mammographic Mass data set, we record high accuracy for the CMIM and JMI methods with 2 features.

The figure 11 illustrates the classification accuracy rate obtained by the four better features selected by these methods in each test. We use the hold out method to generate 20 training sets for each data sets and we calculate the classification accuracy rate for each training sets by using the four better features.

There is different interpretation; each feature selection method is adapted to a special data set. We calculate the average classification accuracy rate obtained in each test and we summarize the results in the following table.

The goal of feature selection is to achieve a balance between the stability of a criterion and the classification accuracy rate (Gulgezen et al. 2009). This is why, experimental protocol was to take the average classification accuracy rate obtained by the 20 training sets plotted with the Kuncheva's Stability, Information Stability, SS Stability and SH Stability. Figures 12, 13, 14, 15 show the stability criterions versus the means accuracy rate. The goal is the find the set of feature selection criterions which the higher classification accuracy rate and the higher stability, this set is called the Pareto-Optimal Set.

The criteria which belonging the Pareto-Optimal set is said to be non-dominated [18]. Hence, it is evident from each subplot of Figures 12, 13, 14, and 15 that feature selection techniques positioned towards the top right of the space dominate over those towards the bottom left. Given this observation, there is no justification for selecting techniques located at the bottom left [18].

6 Conclusion

This paper introduces a comparison protocol evaluating twenty feature selection techniques across four datasets sourced from the UCI machine learning repository. The experimentation assesses stability criteria and classification accuracy rates calculated using SVM-SMO. Based on this research, we have concluded that each feature selection method can be tailored to suit specific datasets, considering factors such as the number of features and their distribution in the feature space.

The classification accuracy rate and the Stability provide a good experimentation and perfect information of features, the better feature selection method is one that has the both higher accuracy rate and stability. It is very interesting to evaluate the performance of these feature selection techniques in the analysing DNA Microarrays, where there are many features and comparatively few samples.

References

1. **Medjahed, S. A., Ouali, M., Benyettou, A., Ait-Saaid, T. (2015).** An optimization-based framework for feature selection and parameters determination of SVMs. International Journal of Information Technology and Computer Science, Vol. 7, No. 5, pp. 1–9. DOI: 10.5815/ijitcs.2015.05.01.
2. **Guyon, I., Elisseeff, A. (2003).** An introduction to variable and feature selection. International Journal of Machine Learning Research, Vol. 3, pp. 1157–1182.
3. **Goswami, S., Chakrabarti, A. (2014).** Feature Selection: A Practitioner View. International Journal of Information Technology and Computer Science, Vol. 6, No. 11, p. 66. DOI: 10.5815/ijitcs.2014.11.10.
4. **Lin, S. W., Lee, Z. J., Chen, S. C., Tseng, T. Y. (2008).** Parameter determination of support vector machine and feature selection using simulated annealing approach. Applied soft computing, Vol. 8, No. 4, pp. 1505–1512**.** DOI: 10.1016/j.asoc.2007.10.012.
5. **Kittler, J. (1978).** Feature selection and extraction, Academic Press, New York, https://api.semanticscholar.org/CorpusID:60554370.
6. **Rendell, L., Seshu, R. (1990).** Learning hard concepts through constructive induction: Framework and rationale. Computational Intelligence, Vol. 6, No. 4, pp. 247–270. DOI: 10.1111/j.1467-8640.1990.tb00298.x.
7. **Cortes, C., Vapnik, V. (1995).** Support-vector networks. Machine learning, Vol. 20, pp. 273–297. DOI: 10.1007/BF00994018.
8. **Bartlett, P., Shawe-Taylor, J. (1999).** Generalization performance of support vector machines and other pattern classifiers. Advances in Kernel methods—support vector learning, pp. 43–54.
9. **Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., Murthy, K. R. K. (2001).** Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation, Vol. 13, No. 3, pp. 637–649. DOI: 10.1162/089976601300014493.

10. **Platt, J. C., Schölkopf, B., Burges, C., Smola, A. (1999).** Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods - Support Vector Learning. DOI: 10.7551/ mitpress/1130.003.0016.

11. **Flake, G. W., Lawrence, S. (2002).** Efficient SVM regression training with SMO. Machine learning, Vol. 46, pp. 271–290. DOI: 10.1023/ A:1012474916001.

12. **Lewis, D. D. (1992).** Feature selection and feature extraction for text categorization. Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, pp. 23–26.

13. **Fleuret, F. (2004).** Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research, Vol. 5, No. 9, pp. 1531–1555.

14. **Yang, H. H., Moody, J. (1999).** Data visualization and feature selection: New algorithms for non-gaussian data. Advances in Neural Information Processing Systems, pp. 687–693.

15. **Meyer, P., Bontempi, G. (2006).** On the use of variable complementarity for feature selection in cancer classification. Evolutionary Computation and Machine Learning in Bioinformatics, pp. 91–102. DOI: /10.1007/117 32242_9.

16. **Lin, D., Tang, X. (2006).** Conditional infomax learning: An integrated framework for feature extraction and fusion. In: Leonardis, A., Bischof, H., Pinz, A. (eds). Computer Vision ECCV 2006, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. Vol 3951. DOI: 10.1007/11744023_6.

17. **Jakulin, A. (2005).** Machine learning based on attribute interactions. PhD thesis, University of Ljubljana, Slovenia.

18. **Brown, G., Pocock, A., Zhao, M. J., Luján, M. (2012).** Conditional likelihood maximization: a unifying framework for information theoretic feature selection. The journal of machine learning research, Vol. 13, No. 1, pp. 27–66.

19. **Battiti, R. (1994).** Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks, Vol. 5, No. 4, pp. 537–550. DOI: 10.1109/72.298224.

20. **Peng, H., Long, F., Ding, C. (2005).** Feature selection based on mutual information: Criteria of max dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238. DOI: 10.1109/TPAMI.20 05.159.

21. **Yu, L., Liu, H. (2004).** Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, Vol. 5, pp. 1205–1224.

22. **Zhao, Z., Liu, H. (2007).** Spectral feature selection for supervised and unsupervised learning. Proceedings of the 24th international conference on Machine learning, pp. 1151–1157. DOI: 10.1145/1273496.1273641.

23. **Wei, L. J. (1981).** Asymptotic conservativeness and efficiency of Kruskal-Wallis test for k dependent samples. Journal of the American Statistical Association, Vol. 76, No. 376, pp. 1006–1009. DOI: 10.1080/0162 1459.1981.10477756.

24. **Duda, R. O., Hart, P. E., Stork, D. G. (2001).** Patter Classification, Jhon Wiley and Sons, New York.

25. **Cover, T. M., Thomas, J. A. (1991).** Elements of Information Theory, Wiley.

26. **Cheng, Q., Zhou, H., Cheng, J. (2011).** The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 6, pp. 1217–1233. DOI: 10.1109/ TPAMI.2010.195.

27. **Zhao Z., Morstatter F., Sharma S., Alelyani S., Anaud A., Liu, H. (2010).** Advancing feature selection research. Feature Selection Repository Arizona State University.

28. **Oh, I. S., Lee, J. S., Moon, B. R. (2004).** Hybrid genetic algorithms for feature selection. IEEE Transactions on pattern analysis and machine intelligence, Vol. 26, No. 11, pp. 1424–1437. DOI: 10.1109/TPAMI.2004.105.

29. **Kalousis, A., Prados, J., Hilario, M. (2005).** Stability of feature selection algorithms. Fifth

IEEE International Conference on Data Mining ICDM'05, IEEE, pp. 8. DOI: 10.1109/ICDM.2005.135.

30. **Dunne, K., Cunningham, P., Azuaje, F. (2002).** Solutions to instability problems with sequential wrapper-based approaches to feature selection. Journal of Machine Learning Research, Vol. 1, pp. 22.

31. **Kuncheva L. I. (2007),** A stability index for feature selection. Proceedings of the IASTED International Multi-Conference on Artificial Intelligence and Applications, pp. 390–395.

32. **Yu, L., Liu, H. (2004).** Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research, Vol. 5, pp. 1205–1224.

33. **Yu, L., Ding, C., Loscalzo, S. (2008).** Stable feature selection via dense feature groups. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 803–811. DOI: 10.1145/1401890.1401986.

34. **Fonseca, C. M., Fleming, P. J. (1996).** On the performance assessment and comparison of stochastic multiobjective optimizers. International conference on parallel problem solving from nature, Springer Berlin Heidelberg. pp. 584–593. DOI: 10.1007/3-540-61723-X_1022.

Article received on 20/12/2023; accepted on 16/04/2024.
[*]Corresponding author is Seyyid Ahmed Medjahed.