Comparative Study of Embedded Feature Selection Methods on Microarray Data

Hind Hamla, Khadoudja Ghanem

Comparative Study of Embedded Feature Selection Methods on Microarray data

Hind Hamla[1][0000−0002−7062−7795] and Khadoudja Ghanem[2]

[1] University of Abdelhamid Mehri constantine2, Constantine, Algeria
[2] Laboratory of Modelling and Implementation of Complex Systems
hind.hamla@univ-constantine2.dz
[3] University of Abdelhamid Mehri constantine2, Constantine, Algeria
[4] Laboratory of Modelling and Implementation of Complex Systems
gkhadoudja@yahoo.fr

Abstract. Microarray data collects information from tissues that could be used in early diagnosis such as cancer. However, the classification of microarray data is a challenging task due to the high number of features and a small number of samples leading to poor classification accuracy. Feature selection is very effective in reducing dimensionality; it eliminates redundant and irrelevant features to enhance the classifier's performance. In order to shed light on the strengths and weaknesses of the existing techniques, we compare the performances of five embedded feature selection methods namely decision tree, random forest, lasso, ridge, and SVM-RFE. Ten well-known microarray datasets are tested. Obtained results show the outperformance of SVM-RFE in term of accuracy, and comes in the second position after decision tree in terms of number of selected features and execution time.

Keywords: Feature selection · Machine learning · Embedded methods · Microarray data · SVM-RFE.

1 Introduction

Artificial inelligence AI based learning is able to substantially "automate discovery" across many domains where classification and prediction tasks play an important role. It hold out the prospect of dramatically lower costs and improve performance especially with DNA microarray data. whithin this specific field, AI based learning can learn to read the genome in ways that human cognition and perception cannot as it was argued by Leung et al 2016.[13]

DNA Microarray technology collects information from tissues and cell samples, analyze this type of data allows research to early diagnose diseases such as cancer [10].Artificial intelligence techniques have made it possible to analyze this data for classification using statistical methods. However, there are often a few samples (often fewer than 100 samples) and a huge number of features in the raw data up to 60,000 which causes the curse of dimensionality [2].
According to [23] only a small number of features are used in classification, hence,

feature selection plays an important role in removing irrelevant features from microarray data. Feature selection is used for dimensionality reduction, it eliminates irrelevant and redundant features and preserves all or most informative features [22]. As mentioned in [20], feature selection improves the performance of machine learning algorithms either by increasing the classification accuracy or by decreasing the learning speed.
Feature selection methods are categorized into three categories namely: filter, wrapper, and embedded [4]. Wrapper methods use the classification model to measure the goodness of features [14], these methods are slower than filter methods but they have better performance. Filter methods, however, select features independently from classification models and use statistical proprieties of data to select features. These methods are more time-saving compared to wrapper methods, but they show poorer performance [14]. Embedded methods are a combination of filters and wrapper methods. They integrate the process of selection features in the model training phase. With these methods, the search for the best features subset is guided by the learning model. Embedded methods perform better than filter methods and have less computational time compared to wrapper methods. This is way we explore in this paper this category of feature selection methods.
In order to shed light on the strengths and weaknesses of the existing techniques, we compare the performances of five embedded feature selection methods namely: decision tree, random forest, lasso, ridge, and SVM-RFE. This work tests ten well-known microarray datasets that suffer from the high dimensional problem. The remainder of the paper is organized as follows: Section 2 presents the related works and Section 3 describes the five embedded methods used in this study. Experimental results are given and discussed in Section 4. Finally, Section 5 concludes the paper.

2 Related Works

Wrapper methods suffer from high computational cost while filter methods do not interact with classifiers [12]. Embedded methods can solve these two problems by including the feature selection method in the learning process. In [20], some of the popular feature selection methods were reviewed, and the performance of some of these methods was evaluated in medical domain. Authors in [23], presented a comprehensive state of the art of many existing feature selection methods. Below we present an overview of several recent embedded methods.
In [9], the authors proposed an embedded feature selection method named Recursive Feature Addition (RFA), RFA worked in a forward fashion and based on SVM. The authors used five datasets to test the performance of the proposed method, the results showed superior performance of the proposed method over filter, wrapper, and other embedded methods.
A stable feature selection method based on the L1-norm support vector machine (SVM) was proposed in [19], the proposed method combined L1-norm with SVM classifier for the classification of renal clear cell carcinoma stage classification us-

ing backward feature elimination. RNA-seq gene expression dataset was used to evaluate the performance of the proposed method. The results showed the outperformance of the proposed method.
The authors of [5] propose an embedded feature selection method named MAGS for gene selection and classification of microarray data. This method uses SVM not only to evaluate the quality of the subsets but also to provide valuable information about gene relevancy to design specialized crossover and local search operators. The performance of the proposed method was evaluated over eight microarray datasets, the results illustrate that the proposed method improves the classification accuracy.
In [16] the authors proposed an embedded feature selection method to solve the class imbalance problem called GI-FSw. In this approach, the weighted Gini index was used as splitting criteria of the CART decision tree classifier. The proposed method was tested over two datasets and it has achieved high levels of ROC and f-measure.
Markov blanket-embedded genetic algorithm (MBEGA) for gene selection problem was proposed in [24], in this approach, Markov blanket-based deleted or added features from a solution of genetic algorithm in order to improve the final solution, the performance of this approach was evaluated over four synthetic datasets and eleven microarray datasets, the results proved the effectiveness of the proposed method in eliminating redundant and irrelevant features and it outperforms existing methods in terms of accuracy, number of selected features and computational cost.
In [15], the authors proposed an improved version of RFE that used variable step size named (VSSRFE) in which the step size was decreased as the number of features is eliminated. This method was combined with a more efficient implementation of SVM called (LLSVM). Six well-known microarray datasets were used in the experiments, the results showed that the proposed methods have obtained a comparable classification performance and reduced the computational time.
From this overview, we found that most of the embedded methods presented in the literature to solve feature selection problem are: Decision tree, Rondom Forest, SVM associated with different techniques like L1, L2, RFE... Like it has been mentioned. SVM-RFE was compared in [15] with new variants of SVM-RFE and it has been shown that these variants outperform SVM-RFE. In this paper, we compare DT, RF, L1, and L2 with SVM-RFE on specific data which are microarray data.

3 Methods and materials

As stated before, in embedded methods the selection of features and the learning are dependent. In the following section, the principles of the used embedded methods are introduced.

3.1 Decision tree

Decision tree [21] is a well-known machine learning classifier in form of a tree structure that consists of decision nodes and leaves, whereas each leaf represents a class. Decision tree selects features. This classifier is known by its ability to select features that are important in classification. There are several algorithms for constructing a decision tree using different splitting criteria for example ID3 uses impurity measure, C4.5 uses the concept of information entropy, and CART uses Gini index [16]. Decision tree is able of selecting informative features required for classification because the selection of features is inherent in the decision tree classifier [7].

3.2 Random Forest

Random Forest was proposed by [3], it is an embedded feature selection method that uses variable importance to select features. Random forest constructs a large number of classification trees using bootstrapped samples and determines the predictive class using majority voting. This method is well suited for microarray data because it shows excellent performance when the number of samples is much less than the number of features [4].

3.3 Lasso

Least Absolute Shrinkage and Selection Operator (LASSO) is an embedded feature selection method proposed by Robert Tibshirani in 1996, it applies a regularization process that penalizes the coefficients of the regression variables (features) shrinking some of them to zero [6] and preserves features that still have non-zero coefficient after the regularization process. Lasso is an effective feature selection method considering high-dimensional data due to its ability to produce sparse models in a reasonable time [17].

3.4 Ridge

Ridge is another shrinkage-based embedded method proposed by (Hoerl and Kennard, 1988) [11], it penalizes the square root of the sum of the squared weights (L2 norm). This method applies the regularization process by penalizing the feature coefficients toward zero but never attempt exactly zero [18]. Ridge does not reduce the number of variables since it never leads to a zero coefficient.

3.5 SVM-RFE

SVM-RFE was introduced by [8], it is an iteration backward elimination method that trains SVM with the current set of features and removes the least performing features indicated with SVM recursively until finding the optimal subset [20]. SVM-RFE steps for feature set selection are shown as follows.

– Train the classifier (SVM) with the current dataset.
– Calculates the ranking criterion for all features.
– Delete the least important features (has the smallest ranking criterion).

More than one feature can be removed in one iteration [1].

4 Experimental Results and Discussion

4.1 Datasets

To analyze and test the embedded feature selection performance, ten well-known microarray datasets [24] for various cancer diagnoses were used. Table 1 shows datasets details. The above datasets are partitioned into training and testing sets,

Table 1. Microarray Datasets Details.

Datasets	Number of features	Number of instances	Number of classes
Colon tumor	2000	60	2
Central Nervous System	7129	60	2
Leukemia	7129	72	2
Breast cancer	24481	97	2
Lung_cancer	12533	181	2
Ovarian cancer	15154	253	2
Leukemia 3 classes	7129	72	3
Luekemia 4 classes	7129	72	4
Lymphoma	4026	62	3
MLL	12582	72	3

eight parts were used to train the classifiers and the remaining two were used to test the classifiers. We replace missing values with the mean of the observed values for the corresponding feature. The 10-fold cross-validation method was used in this study.

4.2 Experimental Setup

The five methods used in this study were implemented in python using scikit-learn library. All experiments are performed on a Personal Computer (PC) with an Intel Core i7 processor, 2.9 GHz, and 8 GB of RAM. The parameters values selected to perform the experiments are recorded in tables 2 and 3. It worth to mention that we perform several runs to select the suited parameters.

Table 2. List of classification algorithms with parameter tuning details.

Classifiers Name	Parameter Tuning
Decision tree (DT)	Criterion = 'gini', splitter = 'best', min_samples_split= 2 min_samples_leaf= 1, presort= 'deprecated'
Random Forest (RF)	n_estimators= 100, criterion= 'gini', min_samples_split= 2 min_samples_leaf= 1, bootstrap = True
Lasso	dual= false, tol= 1e-4 C= 1.0, fit_intercept= True
Ridge	Alpha = 1.0, fit_intercept= True normalize= False, tol= 1e-3
SVM-RFE	C =1.0, Kernal= 'linear', degree= 3, gamma= 'scale' shrinking= True, tol= 1e-3, step= 1

In addition, we set the max-depth parameter of decision tree classifier max-depth = 10 for all the data sets except for colon and leukemia-4C max-depth=20 and breast max-depth =100.
Table 3 presents the classification accuracy and execution time in milliseconds without using any feature selection methods, whereas, table 4 presents classification accuracy and execution time (in millisecond) and the number of selected features using embedded methods. The results are the average of five runs.

4.3 Discussion

From the tables above it is evident that using embedded methods significantly improves classification accuracy and reduces execution time. SVM-RFE outperforms all methods on the ten used datasets in term of accuracy using less than twenty features, it is worth to mention that SVM-RFE reaches 100% accuracy in the case of seven datasets, and it comes in the second place after decision tree in term of execution time and the number of selected features. Decision tree selects the least number of features and has the best execution time on nine of the ten datasets, it also has the best performance in term of accuracy after SVM-RFE in the case of five datasets (colon, CNS, leukemia 4-C, breast, and lung), with regard to the other datasets, the accuracy remains quite important. Random forest has the lowest classification accuracy and requires high execution time, however, the number of selected features is reasonable. Thus, when considering these three parameters (accuracy, Execution time and number of selected features), it is suggested to avoid this method in the presence of this type of datasets. But, lasso has average results when compared to the other methods and ridge provides sufficient classification accuracy however it requires the longest execution time on almost all the datasets and it selects the largest number of features (over one thousand features for nine datasets). Finally, we think that it remains preferable to leave the choice to the user, who will use the method that suits his needs in terms of precision (very Sensitive applications), time (Reel time applications) or number of features (Mobile applications) ...

Table 3: Classification accuracy and execution time using all features.

Datasets		DT	RF	Lasso	Ridge	RFE-SVM
Colon	Time	16.18	10.60	10.20	7.32	8414.06
	ACC	75.46%	75.63%	85.76%	78.16%	75.16%
CNS	Time	71.03	13.86	49.57	8.82	181601.87
	ACC	44.66%	62.83%	70.83%	72.5%	72.0%
Leukemia	Time	73.31	41.11	88.24	82.25	169159.77
	ACC	87.44%	91.10%	97.71%	96.33%	93.14%
Leukemia 3-C	Time	85.84	39.79	168.67	77.07	159072.89
	ACC	84.25%	84.62%	96.57%	94.57%	96.0%
Leukemia 4-C	Time	131.84	93.22	326.56	95.02	255629.21
	ACC	69.90%	77.33%	92.92%	91.21%	94.64%
Breast	Time	443.38	34.45	182.76	35.85	4758377.28
	ACC	52.49%	61.05%	72.43%	66.58%	64.44%
Lung cancer	Time	673.86	63.49	965.80	49.48	2016552.5
	ACC	85.40%	88.55%	92.63%	93.28%	90.42
Ovarian	Time	639.34	225.18	352.96	435.79	2099057.88
	ACC	96.27%	97.09%	100.0%	100.0%	100.0%
Lymphoma	Time	62.29	50.37	79.70	107.13	37563.19
	ACC	85.1%	93.32%	94.16%	100.0%	96.66%
MLL	Time	147.90	51.48	291.64	115.85	444714.36
	ACC	80.54%	90.95%	98.57%	96.07%	90.23%

Table 4: Classification accuracy and execution time and the number of selected features.

Datasets		DT	RF	L1	L2	SVM-RFE
ColonTumor	Time	0.55	12.74	4.85	3.45	1.33
	ACC	91.9%	80.56%	84.16%	89.49%	96.0%
	Nb FS	4	42	204.6	590	10
CNS	Time	0.34	9.43	2.07	2.77	1.09
	ACC	89.89%	76.09%	79.33%	81.0%	98.00%
	Nb FS	4	44.2	290.6	1822	15
Leukemia	Time	0.79	9.72	3.08	12.96	1.31
	ACC	96.33%	96.40%	97.8%	98.00%	100.0%
	Nb FS	2	28.8	167.8	1741	10
Leukemia 3-C	Time	0.50	9.51	8.68	15.96	1.26
	ACC	93.24%	90.96%	97.77%	94.57%	100.0%
	Nb FS	3	46.6	349.2	1748	10
Leukemia 4-C	Time	0.50	9.32	12.92	13.98	0.94
	ACC	94.54%	88.26%	93.63%	91.21%	100.0%
	Nb FS	5	57.2	449.6	1771	8

Table 4 continued from previous page

Breastcancer	ACC	0.66	10.41	2.67	13.79	1.44
	Time	96.58%	65.5%	76.19%	82.81%	98.57%
	Nb FS	5	61.6	75.8	9255	15
Lung cancer	Time	0.98	14.13	62.08	15.46	3.86
	ACC	95.17%	91.46%	92.58%	95.17%	100.0%
	Nb FS	8	101.6	665.6	3437	20
Ovarian	Time	0.59	10.90	4.05	62.09	1.58
	ACC	98.81%	97.53%	100.0%	100.0%	100.0%
	Nb FS	3	71.8	42.6	5003	3
Lymphoma	Time	0.49	7.51	1.68	24.99	0.54
	ACC	95.03%	96.66%	97.5%	100.0%	100.0%
	Nb FS	2	25.4	41.2	1574	2
MLL	Time	0.97	8.81	8.92	20.56	0.59
	ACC	95.99%	94.85%	98.92%	96.07%	100.0%
	Nb FS	3	44.4	512.6	3303	4

4.4 Comparison with other works

In this section, we give a comparison of the results obtained by SVM-RFE and some other works presented in the literature. The results are presented in table 5. From table 5, we can see that the method proposed in [15] has achieved better classification accuracy with colon dataset and has selected less number of features. With leukemia, lung, ovarian, and lymphoma, both SVM-RFE and the method in [15] achieved 100% accuracy. With CNS and breast, SVM-RFE outperforms the method proposed in [15] in terms of classification accuracy and the number of selected features. SVM-RFE has better performance than the proposed method in [18] considering the only tested dataset (colon). SVM-RFE produced better performances than the method in [20] in terms of classification accuracy and the number of selected features for the ten datasets.

Table 5: Comparison with other works.

Datasets		DT	RF	Lasso	Ridge	RFE-SVM
Colon	Time	16.18	10.60	10.20	7.32	8414.06
	ACC	75.46%	75.63%	85.76%	78.16%	75.16%
CNS	Time	71.03	13.86	49.57	8.82 72.5% ‘	181601.87
	ACC	44.66%	62.83%	70.83%		72.0%
Leukemia	Time	73.31	41.11	88.24	82.25 96.33%	169159.77
	ACC	87.44%	91.10%	97.71%		93.14%
Leukemia 3-C	Time	85.84	39.79	168.67	77.07	159072.89
	ACC	84.25%	84.62%	96.57%	94.57%	96.0%
Leukemia 4-C	Time	131.84	93.22	326.56	95.02 91.21%	255629.21
	ACC	69.90%	77.33%	92.92%		94.64%

Table 5 continued from previous page

Breast	Time	443.38	34.45	182.76	35.85 66.58%	4758377.28
	ACC	52.49%	61.05%	72.43%		64.44%
Lung cancer	Time	673.86	63.49	965.80	49.48	2016552.5
	ACC	85.40%	88.55%	92.63%	93.28%	90.42
Ovarian	Time	639.34	225.18	352.96	435.79	2099057.88
	ACC	96.27%	97.09%	100.0%	100.0%	100.0%
Lymphoma	Time	62.29 85.1%	50.37	79.70	107.13	37563.19
	ACC		93.32%	94.16%	100.0%	96.66%
MLL	Time	147.90	51.48	291.64	115.85	444714.36
	ACC	80.54%	90.95%	98.57%	96.07%	90.23%

5 Conclusion and Future Work

In this paper we evaluate the performances of five embedded feature selection methods namely decision Tree, random forest, lasso, ridge, and SVM-RFE on ten cancer microarray gene expression datasets which are (Colon, CNS, Leukemia, Leukemia 3-C, Leukemia 4-C, breast, lung, ovarian, lymphoma, and MLL). The experiments show that SVM-RFE gives the highest accuracy among the five tested methods, using a reduced number of features. Decision tree provides high classification accuracy and selects the smallest number of features, it is the fastest method among the other methods discussed in this work and random forest produces the lowest classification accuracy. These are mathematical and algorithmic results, we believe that an expert opinion about the pertinence and the relevance of a specific feature in the final decision making is crucial especially in this domain because he is the only one who can ensure the stability criterion. This is the limitation of this study. As future work, we aim to consider more feature selection methods as well as use other evaluation metrics such as ROC. A main drawback of SVM-RFE is that the calculation time is very high. Therefore, we aim to develop a faster version of SVM-RFE in the future.

References

1. Adorada, A., Permatasari, R., Wirawan, P.W., Wibowo, A., Sujiwo, A.: Support vector machine-recursive feature elimination (svm-rfe) for selection of microrna expression features of breast cancer. In: 2018 2nd International Conference on Informatics and Computational Sciences (ICICoS). pp. 1–4. IEEE (2018)
2. Bolón-Canedo, V., Alonso-Betanzos, A.: Microarray Bioinformatics. Springer (2019)
3. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
4. Díaz-Uriarte, R., De Andres, S.A.: Gene selection and classification of microarray data using random forest. BMC bioinformatics **7**(1), 3 (2006)
5. Duval, B., Hao, J.K., Hernandez Hernandez, J.C.: A memetic algorithm for gene selection and molecular classification of cancer. In: Proceedings of the 11th Annual conference on Genetic and evolutionary computation. pp. 201–208 (2009)

6. Fonti, V., Belitser, E.: Feature selection using lasso. VU Amsterdam Research Paper in Business Analytics **30**, 1–25 (2017)
7. Grabczewski, K., Jankowski, N.: Feature selection with decision tree criterion. in null (2005)
8. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine learning **46**(1-3), 389–422 (2002)
9. Hamed, T., Dara, R., Kremer, S.C.: An accurate, fast embedded feature selection for svms. In: 2014 13th International Conference on Machine Learning and Applications. pp. 135–140. IEEE (2014)
10. Hameed, S.S., Muhammad, F.F., Hassan, R., Saeed, F.: Gene selection and classification in microarray datasets using a hybrid approach of pcc-bpso/ga with multi classifiers. J. Comput. Sci. **14**(6), 868–880 (2018)
11. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
12. Kumar, C.A., Sooraj, M., Ramakrishnan, S.: A comparative performance evaluation of supervised feature selection algorithms on microarray datasets. Procedia computer science **115**, 209–217 (2017)
13. Leung, M.K., Delong, A., Alipanahi, B., Frey, B.J.: Machine learning in genomic medicine: a review of computational problems and data sets. Proceedings of the IEEE **104**(1), 176–197 (2015)
14. Li, H., Guo, W., Wu, G., Li, Y.: A rf-pso based hybrid feature selection model in intrusion detection system. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). pp. 795–802. IEEE (2018)
15. Li, Z., Xie, W., Liu, T.: Efficient feature selection and classification for microarray data. PloS one **13**(8), e0202167 (2018)
16. Liu, H., Zhou, M., Liu, Q.: An embedded feature selection method for imbalanced data classification. IEEE/CAA Journal of Automatica Sinica **6**(3), 703–715 (2019)
17. Ma, S., Song, X., Huang, J.: Supervised group lasso with applications to microarray data analysis. BMC bioinformatics **8**(1), 60 (2007)
18. Marafino, B.J., Boscardin, W.J., Dudley, R.A.: Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to icu risk stratification from nursing notes. Journal of biomedical informatics **54**, 114–120 (2015)
19. Moon, M., Nakai, K.: Stable feature selection based on the ensemble l 1-norm support vector machine for biomarker discovery. BMC genomics **17**(13), 1026 (2016)
20. Remeseiro, B., Bolon-Canedo, V.: A review of feature selection methods in medical applications. Computers in biology and medicine **112**, 103375 (2019)
21. Tahir, N.M., Hussain, A., Samad, S.A., Ishak, K.A., Halim, R.A.: Feature selection for classification using decision tree. In: 2006 4th Student Conference on Research and Development. pp. 99–102. IEEE (2006)
22. Zhang, X., Shi, Z., Liu, X., Li, X.: A hybrid feature selection algorithm for classification unbalanced data processsing. In: 2018 IEEE International Conference on Smart Internet of Things (SmartIoT). pp. 269–275. IEEE (2018)
23. Zheng, Y., Li, Y., Wang, G., Chen, Y., Xu, Q., Fan, J., Cui, X.: Retracted: A hybrid feature selection algorithm for microarray data. Concurrency and Computation: Practice and Experience **31**(12), e4716 (2019)
24. Zhu, Z., Ong, Y.S., Dash, M.: Markov blanket-embedded genetic algorithm for gene selection. Pattern Recognition **40**(11), 3236–3248 (2007)