The Experiment of Leukemia Classification using SVMs with Feature Selection Methods

2019-01-09

Abstract

Motivation: Feature selection methods are used to optimized SVM method for better classifying the unknown tissue sample which are applied on leukemia cancer datasets.
Problem: The MSFS method claims to have better accuracy than the others feature selection. Since, the paper give only brief explanation about the details, the method could not be confirmed. And also, the lymphoma datasets do not have concrete clarity of the data as diverse as investigation that has been conducted. Thus, we only focused on leukemia datasets.
Approach: We performed the SVM by evaluating several feature-selection with different kernel method, and evaluated the weighted average to compared it with the claimed accuracy.
Results: Got similar accuracy as the original paper, but with significantly less feature number. RFE models predict better than SFS, and far more effective.
Conclusion: RFE with linear classification kernel is the best model for Leukenia dataset.

Introduction

Support vector machines (SVMs) are supervised learning method which broadly used in many domains which has been successfully applied and developed in bioinformatics fields. It is one of the state-of-the-art kernel-based machine learning techniques which is suitable for tissue classification and prediction. For instance, the classification of microarray gene expression data. Microarray technologies are utilized to estimate the survival time and risk of cancer metastasis or recurrence based on the patient's genotypic microarray data. Microarray data typically high dimensionality and relatively few examples characterized in gene expression data, which make it impossible to do it manually. Thus, SVMs are utilized to obtain classification models with such high dimensional data. The main focus is the feature selection in order to reduce the dimensionality of dataset. Such that, the better accuracy results can be obtained.

Problem description

The gene expression analysis for classifying and predicting cancer in tissue samples is an important topic on bioinformatics. This paper [1] claims that SVM-MSFS (modified successive feature selection) is better than the other feature selection methods mentioned in the article on classification accuracy (up to 95 %), but there were no detailed description provided in the article.

Data description

We focused on the Leukemia dataset, which contains 7128 genes taken from 72 samples. Two variants of leukemia, ALL (acute lymphocytic leukemia) and AML (acute myelocytic leukemia), were labeled before training (Table 1). The original data were transposed for easier processing with Python. During training process, the order of data within the dataset was randomized before every iteration to avoid bias.

Table 1: Sample numbers for 2 different variants of leukemia, ALL and AML.

Variant name	Numbers
Acute Lymphoblastic Leukemia (ALL)	47
Acute Myeloid Leukemia (AML)	25

Methods

- SVM-Recursive Feature Elimination (RFE)
  SVM-RFE removes redundant genes based on weight-based saliency analysis. Genes connected to important features receive large absolute values of weight, which will later being conserved, and those genes that receive small absolute values of weight will be removed, and the new gene subset is formed for further analysis.

- SVM-Successive Feature Selection (SFS)
  SVM-SFS processes the features in a set one at a time, and found that the suitable values of x taken due to the memory constraints is equal or less than 10, and the rank of features will be the output of the algorithm. In the successive level, the feature is dropped once at a time, a subset of feature is obtained, and the best subset of the feature is processed to the next level after evaluation.

We applied RFE with linear regression kernel and tried linear, RBF (radial basis function) and polynomial classification kernel. As for SFF we tested knn (k-nearset neighbor) and logistic regression kernel with linear classification kernel (Table 2). The flowchart of how the SVM

feature selection algorithms work is shown in Figure 1.

The suggested feature number for the model to select by [1] is 140, and number of training data and testing data are 38 and 34, suggested by [2].But the feature number should not excess sample number significantly like this setting, so we decrease the selected feature number from 10 feature to 70 features, and compare the performance of the model. Since the total number of the samples are quite small, the combination of the training data acts as major factor to the accuracy of the prediction. In order to eliminate the influence of the training data configuration, we randomized the order all the 72 samples in every iteration, but the number of training data and testing data remain the same, 38 and 34. In each iteration, we calculate the Matthews correlation coefficient (MCC), on both training data and testing data.

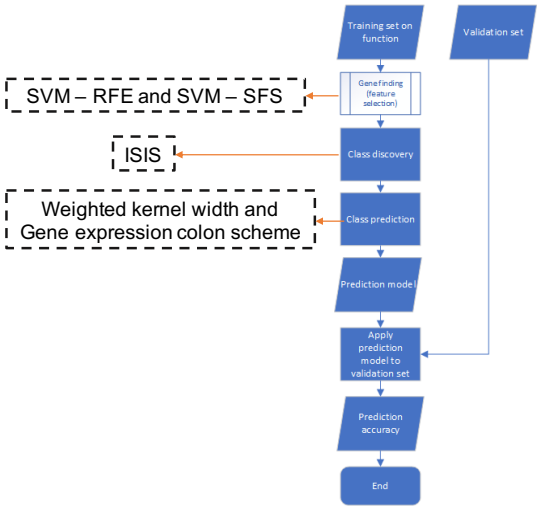

Figure 1: Flowchart of the SVM feature selection algorithm.

Results and Discussions

For RFE method, the performance of polynomial classification kernel is the worst among all classification kernels. No matter how many features and how many iterations, the model can only recognize one class. In this case, the model can not make any prediction at all, so we omit polynomial classification kernel in further comparison. As for linear kernel, the MCC of training data are 1.0 (Figure 3), and MCC of testing data rise with respect to feature number (Figure 4). With RBF kernel, both the MCC of training data (Figure 5) and testing data (Figure 6) vary, but with all of them approach to 1 with increasing feature number.

For SFS method, the time required for computation is far more than RFE, and the MCC of trainind data are all very close to 1.0 (Figure 7, Figure 8). Though KNN regression method takes even more time than logistic regression method, the MCC of KNN on testing data (Figure 9) are a bit less than logistic one (Figure 10).

Conclusion

In Figure 2, we compare the weighted accuracy of our four method with the author's result. The selected feature number are 140 in every method, and the accuracy is the average of the 50 iteration of each method. The accuracy of RFE in our results are a little bit better than the author's. But the accuracy of SFS, which the author claimed is higher than RFE, turned out to be worse than RFE in our result. Especially for SFS with KNN regression kernel, the accuracy is 10% less than the paper's.

The feature number effects accuracy of RFE with RBF kernel considerably. In Figure 11, only when feature number exceeded 40 can the average accuracy got better than 90%. But the RFE accuracy with linear kernel seems similar in every feature number. Also in Figure 12, both Knn and logistic regression kernel show no relationship between accuracy and feature number.

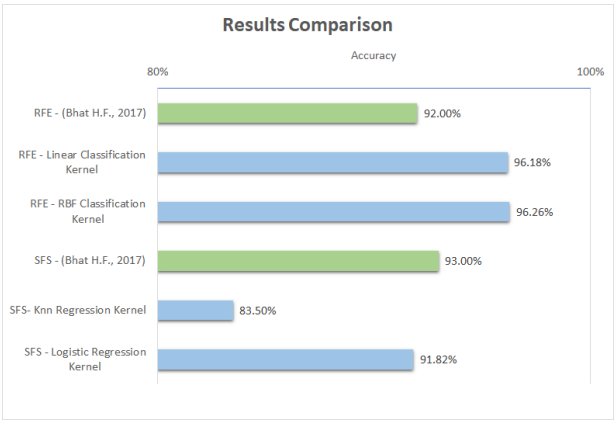

Figure 2: Accuracy comparison between our result and the author's result

References

[1] Heena Farooq Bhat. Evaluating SVM algorithms for bioinformatics gene expression analysis. *International Journal of Computer Science Engineering*, 6(2):42–52, 2017.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999. doi: 10.1126/science.286.5439.531.

Appendix

Table 2: The methods we apply to conduct feature selection and class prediction.

	Regression kernel	Classification kernel
RFE	Linear	Linear
	Linear	RBF
	Linear	Polynomial
SFS	KNN	Linear
	Logistic	Linear

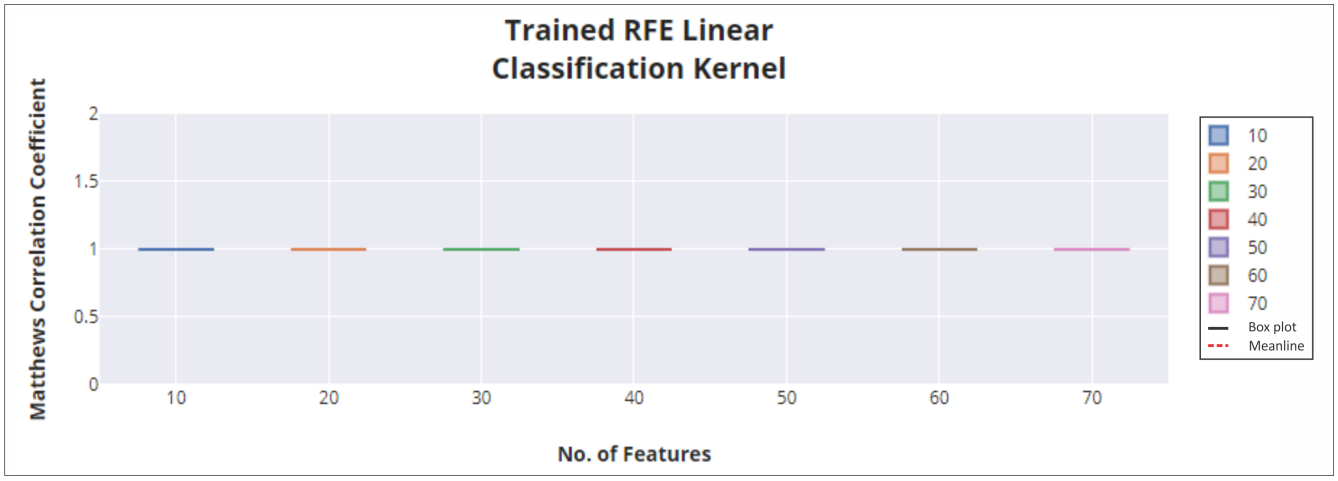

Figure 3: Training data: RFE linear classification kernel

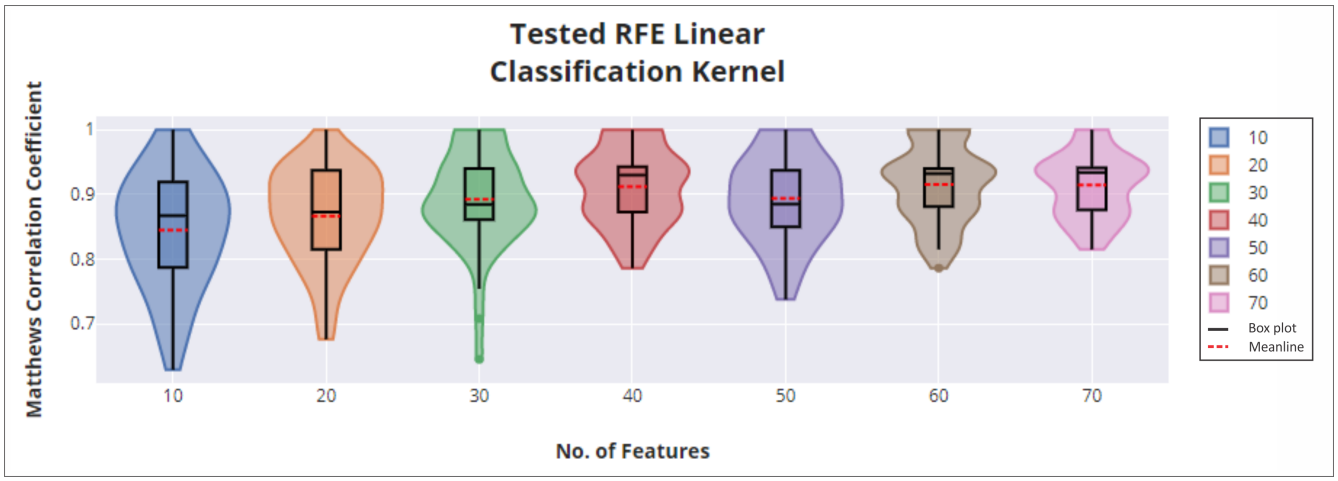

Figure 4: Testing data: RFE linear classification kernel

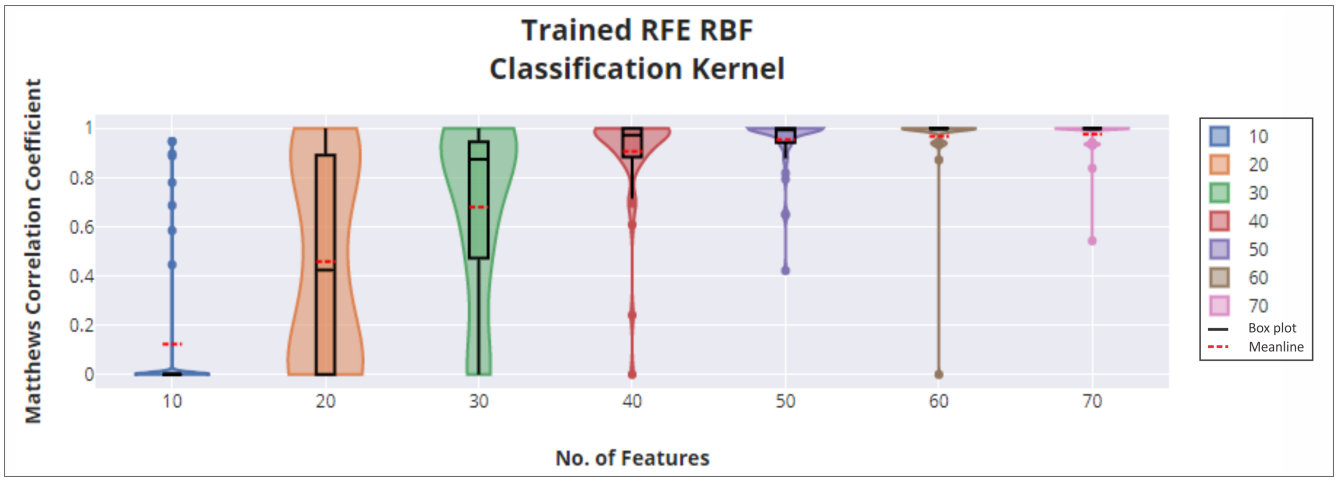

Figure 5: Training data: RFE RBF classification kernel

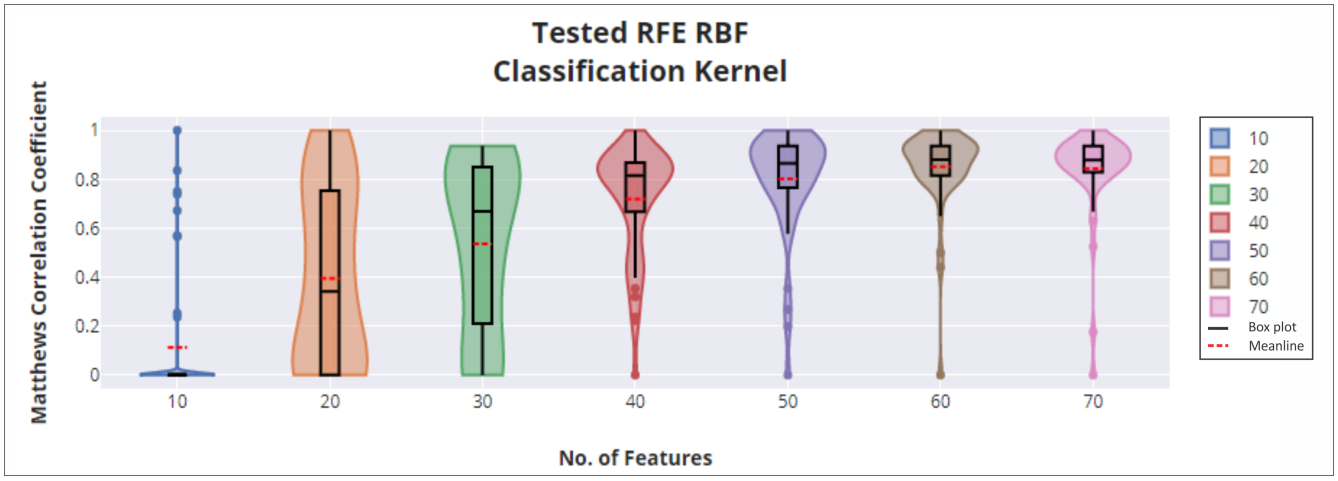

Figure 6: Testing data: RFE RBF classification kernel

Figure 7: Training data: SFS Logistic Regression kernel

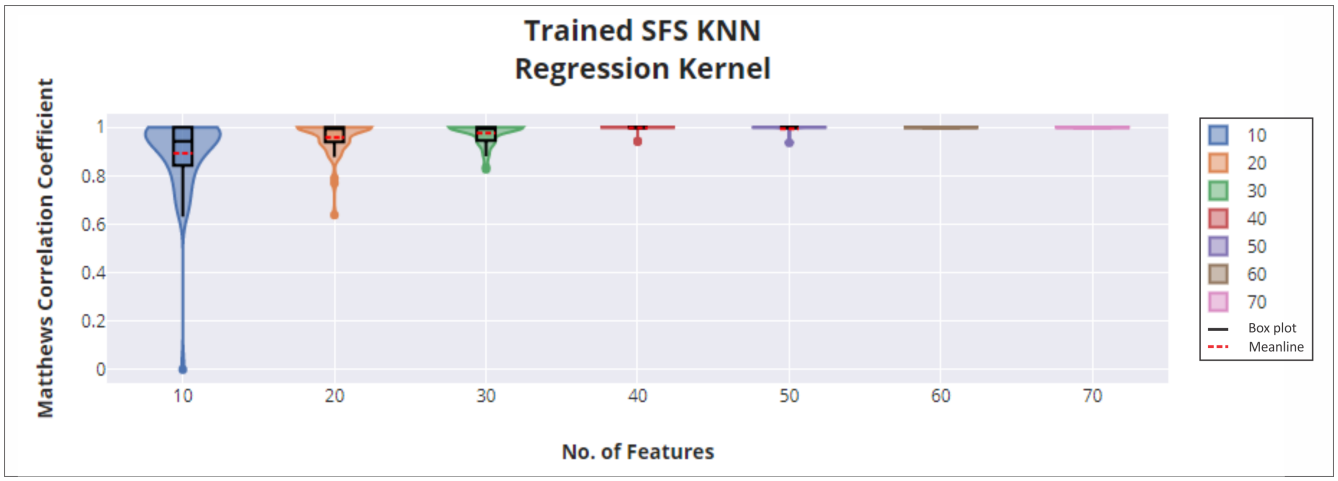

Figure 8: Training data: SFS KNN Regression kernel

Figure 9: Testing data: SFS KNN Regression kernel

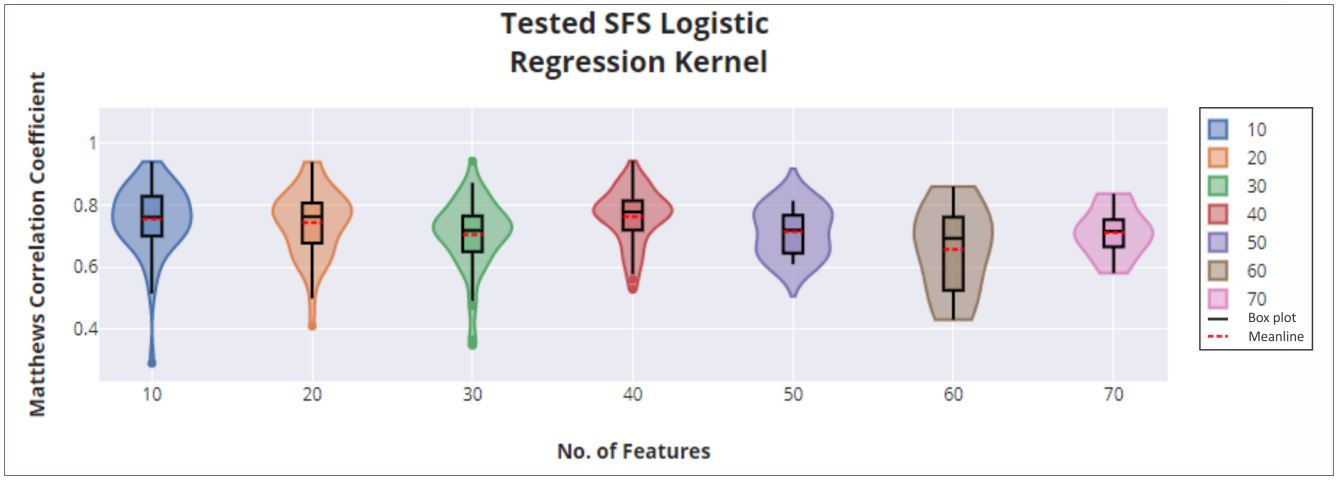

Figure 10: Testing data: SFS Logistic Regression kernel

Table 3: The average Matthew Correlation Coefficient (MCC) and average weighted accuracy of RFE and SFS with different kernel in different feature numbers

Number of features	RFE Linear Classification Kernel		RFE RBF Classification Kernel		SFS KNN Regression Kernel		SFS Logistic Regression Kernel	
	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC	Accuracy
10	0.84	92.80%	0.11	44.18%	0.64	84.46%	0.75	89.52%
20	0.86	94.02%	0.39	68.78%	0.55	79.51%	0.74	89.16%
30	0.89	95.42%	0.53	78.26%	0.57	81.74%	0.70	86.36%
40	0.91	95.57%	0.71	89.85%	0.64	84.06%	0.76	89.94%
50	0.89	95.55%	0.80	92.38%	0.50	76.93%	0.71	86.39%
60	0.91	96.24%	0.85	93.41%	0.50	77.50%	0.65	84.57%
70	0.91	96.16%	0.84	92.86%	0.64	83.96%	0.71	86.35%
80	0.91	96.24%	0.87	94.81%	0.62	83.36%	0.64	83.93%
90	0.92	96.61%	0.89	95.76%	0.67	84.82%	0.71	87.07%
100	0.91	95.78%	0.91	95.98%	0.58	81.43%	0.73	88.79%
110	0.92	96.46%	0.88	95.28%	0.72	87.00%	0.73	88.04%
120	0.91	96.01&	0.90	95.74%	0.68	86.07%	0.72	87.75%
130	0.92	96.28%	0.91	96.18%	0.71	87.86%	0.73	88.75%
140	0.91	96.18%	0.92	96.26%	0.66	83.50%	0.80	91.82%

Figure 11: RFE prediction accuracy in different feature number

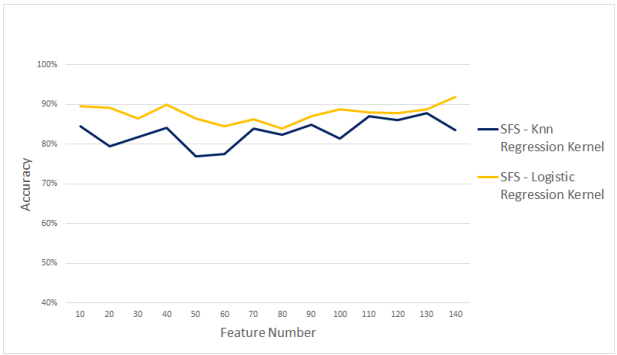

Figure 12: SFS prediction accuracy in different feature number