Experiments vs. EFS

This paper compares NRS-SFSF with the eight EFS feature selection algorithms. EFS is an integrated feature selection tool implemented as an R package and web application, including eight different algorithms (Median, Pearson Correlation, Spearman Correlation, Logistic Regression, algorithms based on random forests with four different metrics) for processing static datasets. In the experiment, we use the online version of these algorithms to calculate the ranking results of each feature. For the online streaming feature selection method NRS-SFSF, it determines the number of selected features automatically. Thus, to compare NRS-SFSF with these eight algorithms, we make them select the same number of top-ranking features. Besides, since we use the online version of EFS and the running time is very long. Thus, we cannot compare the running time between these competing algorithms.

1. Experimental Setup

Since the EFS algorithm is only suitable for binary classification tasks, we select three of them, Leukemia, Colon, Dlbcl, and five low-dimensional datasets. We applied the proposed algorithm to 8 real-world datasets from cDNA microarrays and UCI datasets, as shown in Table 1.

Table 1: Real-world Data sets

Index	Data Set	Instances/Features	Feature Characteristics	Classes
1	Leukemia	72/7,129	Real	2
2	Colon	62/2,000	Real	2
3	Dlbcl	77/7,129	Integer	2
4	Sonar	208/60	Real	2
5	Wdbc	569/30	Real	2
6	Pima	768/8	Real,Integer	2
7	German	1000/20	Real,nominal	2
8	Heart	270/13	Real,nominal	2

We use three MATLAB build-in classifiers: KNN(k=3), SVM(with the linear kernel), and CART to evaluate a selected feature subset in our experiments. We perform 5-fold cross-validation on each data set, and all competing algorithms use the same training and testing data for each fold. The order of streaming features is random, and we run five times for each data set. All experimental results are conducted on a PC with AMD(R) Core i5-7500, 3.4 GHz CPU, and 16 GB memory. We conduct the Friedman test at a 95\% significance level to validate whether these competing algorithms have a significant difference and use the Nemenyi test as a post-hoc test. Besides, the win/tie/loss (W/T/L for short) counts are summarized in the statistical performance.

2. RS-SFSF(k) vs EFS Feature Selection Methods

We compare NRS-SFSF(k=7) with EFS-Median, EFS-Pcor, EFS-Scor, EFS-LogReg, EFS-ERRF, EFS-GiniRF, EFS-AUCCF, and EFS-ERCF. Meanwhile, EFS feature selection algorithms adopt default parameters. Table 2-4 summarizes the average prediction accuracy and the mean number of selected features of these competing algorithms. The p-values of the Friedman test of SVM is 0.0794. Thus, there is no significant difference in predictive accuracy in cases of SVM. The p-values of the Friedman test for KNN and CART are 0.0202 and 0.0351, respectively. Therefore, these competing algorithms have significant differences in both cases of KNN and CART. According to the Nemenyi test, the value of CD is 4.2506.

Table 2: Average prediction accuracy using KNN (N=3) classifier

Data Set	EFS-Median	EFS-Pcor	EFS-Scor	EFS-LogReg	EFS-ERRF	EFS-GiniRF	EFS-AUCCF	EFS-ERCF	NRS-SFSF(k)
Leukemia	0.7029	0.7314	0.7086	0.6543	0.6486	0.6514	0.6543	0.6514	**0.9343**
Colon	0.6867	0.6467	0.6933	0.7	0.6467	0.6633	0.6467	0.65	**0.8267**
Dlbcl	0.7387	0.7093	0.736	0.6933	0.7627	0.7467	0.6773	0.744	**0.9093**
Sonar	0.6933	0.6633	0.6691	0.7077	0.6994	0.6896	0.6965	0.6965	**0.7632**
Wdbc	0.9244	0.9149	0.93	0.9052	0.9276	0.9388	0.9388	**0.9395**	0.9325
Pima	0.6912	0.6907	0.6912	0.6768	0.6867	0.7065	0.6907	0.6907	**0.7266**
German	0.6446	0.6552	0.6446	0.6486	0.6576	0.6508	0.657	0.6582	**0.693**
heart	0.7252	0.737	0.7207	0.7081	0.7274	0.7259	0.7311	0.7297	**0.7541**
W/T/L	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/0/7	7/0/1
AVG.	0.7259	0.7186	0.7242	0.7118	0.7196	0.7216	0.7116	0.72	**0.8175**
AVG. RANKS	2.625	5.875	5.625	6.4375	5.5	4.875	5.3125	4.375	**1.375**

Table 3: Average prediction accuracy using SVM classifier

Data Set	EFS-Median	EFS-Pcor	EFS-Scor	EFS-LogReg	EFS-ERRF	EFS-GiniRF	EFS-AUCCF	EFS-ERCF	NRS-SFSF(k)
Leukemia	0.72	0.7257	0.6943	0.6629	0.7457	0.7543	0.7429	0.7143	**0.9343**
Colon	0.77	0.73	0.7067	0.74	0.7133	0.7167	0.71	0.6967	**0.8033**
Dlbcl	0.7653	0.7813	0.7866	0.736	0.7546	0.7386	0.7707	0.7707	**0.9173**
Sonar	0.677	0.624	0.6435	0.6993	0.6629	0.6763	0.662	0.662	**0.7518**
Wdbc	0.9378	0.9276	0.9459	0.922	0.9416	0.9455	0.9476	**0.948**	0.9445
Pima	0.7227	0.7214	0.7227	0.7107	0.7132	0.7414	0.7214	0.7214	**0.7674**
German	0.7	0.7	0.7	0.7006	0.7	**0.7012**	0.7	0.7	0.7004
heart	0.7519	0.7615	0.7526	0.7511	0.7563	0.7481	0.7667	0.7659	**0.7874**
W/T/L	0/0/8	0/0/8	0/0/8	0/0/8	1/0/7	0/0/8	0/0/8	1/0/7	6/0/2
AVG.	0.7556	0.7464	0.744	0.7403	0.7485	0.7528	0.7527	0.7474	**0.8258**
AVG. RANKS	5.125	5.6875	5.625	6.375	5.8125	4.375	4.8125	5.4375	**1.75**

Table 4: Average prediction accuracy using CART classifier

Data Set	EFS-Median	EFS-Pcor	EFS-Scor	EFS-LogReg	EFS-ERRF	EFS-GiniRF	EFS-AUCCF	EFS-ERCF	NRS-SFSF(k)
Leukemia	0.6457	0.6971	0.6257	0.6314	0.6457	0.6486	0.6941	0.6343	**0.9229**
Colon	0.6133	0.6867	0.6267	0.62	0.6167	0.58	0.63	0.6067	**0.7433**
Dlbcl	0.7227	0.6933	0.7093	0.6373	0.696	0.7147	0.7093	0.744	**0.8027**
Sonar	0.629	0.5864	0.5982	0.6607	0.6326	0.6463	0.6278	0.6278	**0.7143**
Wdbc	0.9058	0.8931	0.9048	0.8903	0.9103	0.9121	0.9114	0.9114	**0.9174**
Pima	0.6519	0.6509	0.6519	0.678	0.6425	0.6903	0.6508	0.6508	**0.7039**
German	0.644	0.6446	0.644	0.6632	0.6402	0.6614	0.6414	0.6428	**0.6754**
heart	0.7067	0.72	0.7007	0.7037	0.7141	0.7089	**0.7281**	0.7245	0.7044
W/T/L	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	0/0/8	1/0/7	0/0/8	7/0/1
AVG.	0.6899	0.6965	0.6827	0.6856	0.6873	0.6953	0.6991	0.6928	**0.773**
AVG. RANKS	5.3125	5.25	6.5625	5.75	6.1875	4	4.75	5.4375	**1.75**

Table 5: Average number of selected features

Data Set	Raw Features	EFS-Median	EFS-Pcor	EFS-Scor	EFS-LogReg	EFS-ERRF	EFS-GiniRF	EFS-AUCCF	EFS-ERCF	NRS-SFSF(k)
Leukemia	7129	10.2	10.2	10.2	10.2	10.2	10.2	10.2	10.2	10.2
Colon	2000	13.3	13.3	13.3	13.3	13.3	13.3	13.3	13.3	13.3
Dlbcl	7129	13.8	13.8	13.8	13.8	13.8	13.8	13.8	13.8	13.8
Sonar	60	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6	6.6
Wdbc	30	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7
Pima	8	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1	4.1
German	20	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
heart	13	4.8	4.8	4.8	4.8	4.8	4.8	4.8	4.8	4.8

From Tables 2-5, we have the following observations. NRS-SFSF(k) achieves the best performance in both cases of KNN and CART. In the case of SVM, all competing algorithms have no significant difference. However, NRS-SFSF(k) gets the highest predictive accuracy and the smallest average ranks.

In sum, in terms of our RS-SFSF framework, the Neighborhood Rough Set based algorithm NRS-SFSF(k) can get better performance in the predictive accuracy with fewer selected features.