FTA: a novel feature training approach for classification

Wanwan Zheng
Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu, Japan
teiwanwan@gmail.com

Mingzhe Jin
Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu, Japan
mjin@mail.doshisha.ac.jp

Abstract

Several studies have been conducted to find the best classification algorithm. Random Forest (RF) and Support Vector Machine (SVM) have been successfully introduced in various prediction models and served as the major data analysis tools that outperform many standard methods. However, RF has difficulties in achieving high accuracy when handling datasets with few instances or variables, and SVM is hard to produce good models if datasets have numerous variables. In this study, the Feature Training Approach (FTA) was proposed, which overcomes the weaknesses of RF and SVM by two trials, namely feature selection and training, SVM ensemble. According to the results of experiments, FTA is quite robust to the two types of data that cannot be well classified, i.e. data with few instances and variables, data with few instances and numerous variables. In most cases, even with different data from different domains, FTA could achieve better performance than RF and SVM.

1 Introduction

Machine learning has become a hot topic in various fields, and classification is a prominent task in machine learning. Data used for classification consists of instance and variable, which can fall into four cases: (1) data with few instances and variables; (2) data with numerous instances and few variables; (3) data with few instances and numerous variables; (4) data with numerous instances and variables. Because enough information is required to complete a statistical description of each class, it is well known that the training of classifiers requires considerable amount of training data (Zhu et al., 2016; Halevy et al., 2009; Mathur and Foody, 2008). However, even if there is considerable amount of data, the classification accuracy of classifiers is not necessarily high. Support Vector Machine (SVM, Cortes and Vapnik, 1995) is an example.

SVM as an effective data analysis tool has been successfully applied to various prediction models. Thanh and Kappas (2018) using Sentinel-2 image data examined and compared the performances of the RF, *k*-Nearest Neighbor (kNN), and SVM for land use/cover classification. According to their findings, SVM produces the highest accuracy with the least sensitivity to the training sample sizes. Kremic and Subasi (2016) applied RF and SVM in facial recognition. As a result, SVM achieves accuracy of 97.94% to the greatest, and RF is 97.17%. Chevalier et al. (2011) compared the performance of SVM with that of Neural Network (NN) in determining air temperature values, and they confirmed the superiority of SVM. Besides, some hybrid methods have been proposed based on SVM. Yong et al. (2015) developed a method based on the combination of Wavelet Transforms (WT) and SVM, which is optimized for those special cases where the real signals contain numerous events in the analyzed temporal window. Tests and trainings were performed using real complex signals, and the results showed the proposed methodology highly efficient. Zheng et al. (2014) proposed to combine *k*-means and SVM to increase the classification accuracy on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to 97.38%.

However, SVM's classification accuracy is affected by the noise involved in datasets for it uses all variables in tuning models. Thus, the accuracy is relatively low when dealing with high dimensional datasets.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 316-323, Hakodate, Japan, September 13-15, 2019

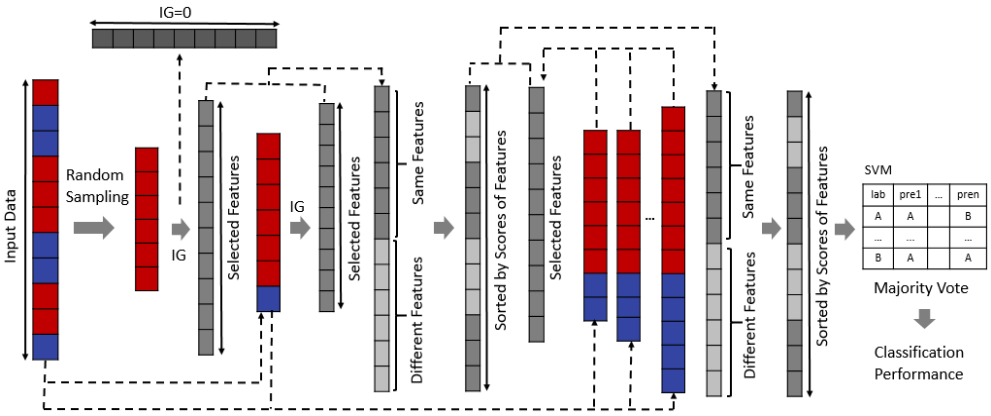

Figure 1: The overall procedure of FTA.

RF has also been extensively used since it's introduced in 2001 (Breiman, 2001). It also has become a standard classification approach in many fields. Couronne et al. (2017) presented a large-scale benchmarking experiment based on 260 real datasets to compare the performance of RF and logistic regression (LR) in prediction. As a result, all measures suggest a significantly better performance of RF. Chelgani et al., (2016) employed RF as a sensible tool for variable importance measurements using various coal properties to predict coke quality. According to the result, RF can further be a reliable and accurate technique to determine complex relationship by fuel and energy investigations. Liu et al. (2013) introduced and investigated RF, Back Propagation Neural Network (BPNN) and SVM to deal with electronic tongue data, and RF is proven to outperform BPNN and SVM.

RF has several advantages over other statistical modeling techniques: (1) capable of dealing with missing values and high-dimensional data; (2) capable of identifying complex interactions between variables and the most important variables; (3) high prediction accuracy; (4) robust against over-fitting. However, from the perspective of random sampling of instances and variables, RF is usually not very accurate when the numbers of samples or variables are small.

In this study, Feature Training Approach (FTA) is proposed as a new classification model which trains features and improves the weaknesses of RF and SVM.

The rest of this article is organized as follows. Sections 2 proposes the approach followed by experiments reported in Section 3. Section 4 gives an explanation why FTA works, and Section 5 draws the conclusions and states limitations and further work.

2 Feature Training Approach

The FTA refers to a two-phase hybrid approach. In the first phase, it performs feature selection and feature training alternately to make a list of selected features. In the second phase, SVM is used to make predictions. The same process is performed K times, and labels are finally determined for test data by majority vote. Figure 1 summarizes the overall procedure of FTA.

Information gain (IG) serves as base feature selection method, which has been validated as a representative feature selection method (Geurts et al., 2018; Chinnaswamy et al., 2017; Wosaiak and Dziomdziora, 2015; Adel et al., 2014). IG measures the reduction in entropy (impurity in an arbitrary collection of examples). With the entropy of Y defined as:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)) \quad (1)$$

Where $p(y)$ is the marginal probability density function for the random variable Y.

IG is defined as:

$$\Delta H = H - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R \quad (2)$$

Where m is the total number of instances, with m_k instances belonging to class k (k=1,2,…,k).

IG, a supervised feature selection method, is more independent on the number of training

Split D_{all} into D_{tra} and D_{pre}

Input: (training) data $D_{tra} = \left\{\left(I_i, V_i\right)\right\}_{i=1}^{N}$ I: instance V: variable

1: **For** t=1 to K **do**:

Feature selection and feature training phase

2: Split D_{tra} into N-fold

3: **Input:** Define subspace s by extracting n fold ($n < N$) randomly

$$s = \bigcup_{j \in n_1} s_j \,,\ s' = \bigcup_{m \in n_2} s_m \,,\ n_1 + n_2 = N \,,\ s \cap s' = \varnothing$$

4: Perform feature selection using IG for s

5: Delete $V_i \left(IG_{V_i} = 0\right)$; Record $V_j \left(IG_{V_j} \neq 0\right)$

6: **For** m=1 to n_2 **do**:

7: $s = s + s'_m$

8: Perform feature selection using IG for s

9: Delete $V_{im} \left(IG_{V_i m} = 0\right)$; Record $V_{jm} \left(IG_{V_{jm}} \neq 0\right)$

10: $V_j = V_j \bigcup V_{jm}$, $V_{same} = V_j \bigcap V_{jm}$, $V_{diff} = V'_{same}$

11: $IG_{V_{same}} = mean(V_j, V_{jr})$, $IG_{V_{diff}} = IG_{V_{jm}}$

12: Sort V_j by the scores of IG_{V_j}

13: **End for;**

Output: Extract the top t features as the final list of selected features

Prediction phase

14: **Input:** (testing) data $D_{pre} = \left\{\left(I_i, V_i\right)\right\}_{i=1}^{M}$

15: Apply SVM using the final list of selected features

16: Build training model

17: **Output:** predictions for every instance in D_{pre}

18: **End For;**

Output: majority vote

Figure 2: Pseudo-code FTA.

Data type	#Samples	#Variables	#Datasets
Data with few instances and variables	5×n	13–40	30×10
Data with numerous instances and few variables	40×n–100×n	13–40	30×10
Data with few variables and numerous instances	5×n	294–3,645	30×10
Data with numerous instances and variables	100×n	294–3,645	30×10

n is the number of classes, in the experiment, n=2, 3

Table 1: Information of data.

samples than the unsupervised feature selection method (e.g. principal component analysis, PCA) and distance-based feature selection method (e.g. chi-squared) (Zheng and Jin, 2018). Inspired by this recognition, feature training as the core mechanism in the FTA gradually increases the amount of training samples and updates the list of selected features. In such a way, the same effect as repeated learning with different training data can be obtained. The pseudo code of FTA is shown in Figure 2.

	Reduction of dimension (%)		Mean			Win			p-value		
	min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Leukemia	75	88	0.8073	0.7954	0.9817	0	0	10		***	***
Bioresponse	93	99	0.5428	0.5695	0.9266	0	0	10		***	***
Gina agnostic	92	96	0.6132	0.5968	0.9229	0	0	10		***	***
Scene	63	91	0.7363	0.6744	0.9450	2	1	10		**	***
Isolet	61	77	0.9588	0.9497	0.9817	7	5	9			
Speech	97	99	0.7700	0.7346	0.8182	0	0	10		*	***
Robert	89	94	0.5949	0.5153	0.9489	0	0	10		***	***
Christine	86	93	0.5332	0.5013	0.9541	0	0	10		***	***
Madelon	89	98	0.4066	0.4176	0.9908	0	0	10		***	***
Arcene	78	98	0.4308	0.4576	0.8052	1	1	10		***	***
Character Font_ARIAL	84	96	0.5613	0.5206	0.8477	0	0	10		***	***
Character Font_CALIBRI	91	98	0.4827	0.3871	0.8484	0	0	10		***	***
Character Font_COURIER	81	99	0.5146	0.5300	0.8861	0	0	10		***	**
Character Font_LUCIDA	93	98	0.4028	0.3755	0.7773	0	0	10		***	***
Character Font_NIRMALA	84	96	0.5793	0.5862	0.9450	0	0	10		***	***
cifar-10-small(0,1,2)	88	97	0.4519	0.4936	0.7871	0	0	10		***	***
cifar-10-small(3,4,5)	80	98	0.3446	0.3341	0.7559	0	0	10		***	***
cifar-10-small(6,7)	79	97	0.6066	0.6017	0.9033	1	0	10		**	**
cifar-10-small(8,9)	76	92	0.6545	0.7112	0.9337	0	0	10		**	*
Eating(1,2,3)	91	95	0.3875	0.4239	0.5829	1	1	8		**	*
Eating(4,5)	84	92	0.5830	0.6878	0.8138	2	1	9		*	
Eating(6,7)	82	92	0.6021	0.7751	0.8545	2	6	8	†	**	
Fashion_Mnist(0,1,2)	65	91	0.8873	0.8762	0.9276	5	4	7			
Fashion_Mnist(3,4,5)	59	82	0.8255	0.8528	0.8785	4	4	8			
har(1,2,3)	65	91	0.7512	0.7912	0.9469	0	0	10		**	**
har(4,5,6)	63	94	0.7167	0.6079	0.8865	0	1	10	†	**	***
svhn(1,2,3)	85	97	0.4645	0.3294	0.7591	0	0	10		**	***
svhn(4,5,6)	93	96	0.3468	0.3183	0.7801	0	0	10		***	***
svhn(7,8)	81	95	0.5131	0.5110	0.8861	0	0	10		***	***
svhn(9,10)	82	96	0.5385	0.5013	0.8888	0	0	10		***	***
mean	**81.0349**	**94.0384**	**0.5869**	**0.5809**	**0.8721**	**0.8333**	**0.8000**	**9.6333**			

*** p < 0.001, ** p < 0.01, * p < 0.05, † p < 0.1

Table 2: Results of the benchmarking experiment using the macro-averaged F-measure. (Data with few instances and numerous variables)

3 Experiments

3.1 Analysis data

Experiments were run on a total of 60 benchmark datasets (30 datasets with numerous variables, 30 datasets with few variables), covering biological data, image data, voice recognition data, physical data and artificial data. Furthermore, to generate the four types of data mentioned in Section 1, random sampling was performed 10 times respectively. The information of data is listed in Table 1.

For data with few instances and variables, the number of instances is set as $5\times n$ (n is the number of classes, each of which has 5 instances), the number of variables is between 13 and 40, and the number of datasets is 30×10 (30 datasets, random sampling was performed 10 times for each dataset). For data with numerous instances, the number of instances is set as 100. The classifications include binary classification and 3-class classification. Macro-averaged F-measure serves as the evaluation metric.

3.2 Experimental results

In this study, for datasets with 5×n instances and 40×n–100×n instances, leave-one-out cross validation (LOOCV) and 10-fold cross validation was conducted, respectively. Furthermore, all the features selected after training were applied as the final list of selected features.

For classifiers, because the probability of overfitting increases with the increase in the number of variables, one of the most challenging tasks is to make correct prediction of data with few instances and numerous variables. The classifier requires the ability to create a learning model that describes the characteristics of data with few instances. Table 2 shows the result of data with few instances and numerous variables.

FTA reduced the dimension of data to the minimum 81.0349% and the maximum of 94.0384% by average. The average of macro-averaged F-measure of FTA, RF, and SVM are 0.8721, 0.5809, 0.5869, respectively, and the average numbers of wins of FTA, RF, and SVM

	Reduction of dimension (%)		Mean			Win			p-value		
	min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Cardiotocography	78	92	0.4040	0.5424	0.6874	0	3	7		**	
WDBC	27	67	0.8061	0.8644	0.8308	6	7	7			
Vehicle	21	92	0.4485	0.5420	0.6505	0	1	9		**	
Waveform	69	86	0.7152	0.6754	0.8155	3	2	8			†
Software	13	88	0.6742	0.7652	0.8324	3	5	8		*	
Climate	75	86	0.7065	0.7104	0.8733	2	2	8			
HallofFame	14	71	0.5902	0.6106	0.6596	2	5	5			
Fri	80	90	0.5149	0.6064	0.8361	0	2	10		**	*
analcatdata_authorship	66	85	0.9424	0.9526	0.9630	7	7	8			
zernike(1,2,3)	29	61	0.9878	0.8828	0.9878	10	0	10	***		***
zernike(4,5,6)	63	83	0.7702	0.6846	0.8737	0	1	10	†	*	***
zernike(7,8)	36	77	0.8629	0.8821	0.9433	4	5	8			
zernike(9,10)	51	81	0.9056	0.8561	0.9908	4	2	10	†		**
first-order-theorem(1,2,3)	80	100	0.3878	0.4756	0.6544	0	0	10		***	**
first-order-theorem(4,5,6)	78	94	0.3450	0.4416	0.6196	0	0	10		**	†
gesturehaseSegmentation (DHP)	80	97	0.4449	0.4389	0.5033	2	1	7			
gesturehaseSegmentation(RS)	79	94	0.5342	0.5723	0.7395	1	2	9		**	*
hillVally	3	50	0.4666	0.4381	0.5879	2	2	10			
kc1	9	85	0.5328	0.5626	0.6551	2	4	8			
musk	74	94	0.5840	0.5631	0.7876	2	2	9		*	**
ozone-level-8hr	53	94	0.6117	0.7144	0.8569	0	3	9		**	
qsar-biodeg	59	90	0.6517	0.7721	0.8541	3	3	9		*	
semeion(1,2,3)	86	96	0.9175	0.9167	1.0000	2	2	10		**	**
semeion(4,5,6)	87	99	0.9052	0.9049	0.9744	2	3	9		*	*
semeion(7,8)	89	98	0.8861	0.9233	0.9908	2	4	10		**	†
semeion(9,10)	84	96	0.9317	0.8669	0.9800	7	2	10			*
spambase	73	93	0.5727	0.7789	0.7940	1	5	5	*	**	
steel-plateds-fault	44	90	0.5662	0.5756	0.7503	3	0	9		*	*
wall-robot-navigation (1,2)	58	92	0.5008	0.6284	0.7248	1	3	7		*	
wall-robot-navigation (3,4)	33	58	0.9417	0.9817	0.9541	6	9	7			
mean	**56.3667**	**85.9667**	**0.6703**	**0.7043**	**0.8124**	**2.5667**	**2.9000**	**8.5333**			

*** p < 0.001, ** p < 0.01, * p < 0.05,† p < 0.1

Table 3: Results of the benchmarking experiment using the macro-averaged F-measure. (Data with few instances and variables)

are 9.6333, 0.8000, 0.8333, respectively. These average numbers of wins were calculated based on the number of wins of every classifier in terms of macro-averaged F-measure per dataset after 10 times of random sampling. Furthermore, the Tukey's honest significant difference method was employed to verify whether there exists significant difference between any two classifiers. According to the result, significant difference was found between FTA and RF, SVM in most cases.

For classifiers, another challenging task is to correctly predict data with few instances and variables. Halevy et al. (2009) reported that even very complex problems in artificial intelligence may be solved by simple statistical models trained on massive datasets. And numerous research have shown that classification accuracy tends to be positively related to training dataset size (Zhu et al., 2015, Mathur and Foody., 2008, Foody and Mathur, 2004, Pal and Mather, 2003). Because classifiers require enough training data to complete the statistical description of each class, few instances and variables mean that the information used to tune model may probably be insufficient. Table 3 shows the result of data with few instances and variables.

FTA reduced the dimension of data to the minimum of 56.3667% and to the maximum of 85.9667% by average. The average of macro-averaged F-measure of FTA, RF, and SVM are 0.8124, 0.7043, 0.6703, respectively, and the average numbers of wins of FTA, RF, and SVM are 8.5333, 2.9000, 2.5667, respectively. Furthermore, according to the results of Tukey's honest significant difference method, there exists significant difference between FTA and RF, SVM in most cases.

The results of the other two types of datasets are summarized in Table 4 and Table 5, respectively.

For data with numerous instances and variables, it is considered that RF should be good at dealing with such datasets. The average of macro-averaged F-measure of FTA, RF, and SVM are 0.7572, 0.7593, 0.7140, respectively. FTA performs as well as RF does. The average numbers of wins of FTA, RF, and SVM are 5.9333, 3.7333, 1.5667, respectively. Moreover, there exists significant difference between FTA and RF, SVM in almost half cases according to

	Reduction of dimension (%)		Mean			Win			p-value		
	min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Leukemia	73	84	0.9731	0.9679	0.9854	6	4	10		†	*
Bioresponse	98	100	0.6399	0.6911	0.7418	0	1	9	*	***	*
Gina agnostic	92	95	0.8268	0.8484	0.8700	0	0	10	*	***	*
Scene	52	68	0.8103	0.8177	0.8209	3	6	5			
Isolet	39	44	0.9960	0.9850	0.9957	10	2	10	***		***
Speech	88	92	0.5588	0.5407	1.0000	0	0	10		***	***
Robert	58	75	0.7026	0.7386	0.7131	1	7	2	*		
Christine	84	98	0.6762	0.6668	0.7037	1	0	10			
Madelon	98	100	0.5753	0.5535	0.6458	0	0	10		**	***
Arcene	92	100	0.7242	0.7605	0.8319	0	0	10	*	***	***
Character Font_ARIAL	15	30	0.7650	0.7797	0.7668	1	8	1			
Character Font_CALIBRI	93	99	0.6554	0.6671	0.6772	0	3	7			
Character Font_COURIER	75	96	0.7088	0.7724	0.7045	0	10	0	**		**
Character Font_LUCIDA	85	98	0.5502	0.5804	0.5430	2	7	1			
Character Font_NIRMALA	95	99	0.5781	0.5806	0.6405	0	0	10		***	**
cifar-10-small(0,1,2)	43	63	0.6933	0.6696	0.6900	5	0	5			
cifar-10-small(3,4,5)	84	99	0.5061	0.5163	0.5281	3	1	6			
cifar-10-small(6,7)	61	85	0.7727	0.7857	0.7958	1	2	8			
cifar-10-small(8,9)	66	88	0.7647	0.7678	0.7790	2	1	8			
Eating(1,2,3)	87	92	0.3558	0.6871	0.3826	0	10	0	***	*	***
Eating(4,5)	75	79	0.6679	0.9521	0.6511	0	10	0	***		***
Eating(6,7)	73	82	0.6594	0.9226	0.6620	0	10	0	***		***
Fashion_Mnist(0,1,2)	20	35	0.9513	0.9482	0.9522	3	3	6			
Fashion_Mnist(3,4,5)	9	14	0.9405	0.9399	0.9408	5	5	5			
har(1,2,3)	23	29	0.9640	0.9357	0.9643	4	0	7	***		***
har(4,5,6)	67	77	0.9041	0.9328	0.9119	0	9	1	***		**
svhn(123)	96	99	0.5610	0.6483	0.6564	0	3	7	***	***	
svhn(456)	95	99	0.5544	0.6429	0.6259	0	6	4	***	***	
svhn(78)	93	99	0.7361	0.7829	0.7899	0	3	7	†	*	
svhn(910)	96	100	0.6480	0.6962	0.7461	0	1	9		***	**
mean	**70.8333**	**80.6000**	**0.7140**	**0.7593**	**0.7572**	**1.5667**	**3.7333**	**5.9333**			

*** p < 0.001, ** p < 0.01, * p < 0.05,† p < 0.1

Table 4: Results of the benchmarking experiment using the macro-averaged F-measure.
(Data with numerous instances and variables)

	Reduction of dimension (%)		Mean			Win			p-value		
	Min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Cardiotocography	66	81	0.4462	0.8958	0.6585	0	10	0	***	***	***
WDBC	10	23	0.9131	0.9471	0.9166	1	10	1	***		***
Vehicle	4	36	0.5709	0.6891	0.5801	0	10	0	***		***
Waveform	55	57	0.8500	0.8436	0.8560	2	3	7			
Software	37	82	0.6109	0.7492	0.6672	0	10	0	***	*	**
Climate	81	88	0.7670	0.7791	0.8135	0	2	8			
HallofFame	0	7	0.7353	0.7507	0.7374	3	8	2			
Fri	84	90	0.6390	0.8368	0.8852	0	0	10	***	***	**
analcatdata_authorship	17	23	0.9897	0.9880	0.9887	5	3	5			
zernike(1,2,3)	2	12	0.9950	0.9910	0.9957	9	0	10	*		**
zernike(4,5,6)	18	27	0.9459	0.9177	0.9442	6	0	5	***		***
zernike(7,8)	10	23	0.9831	0.9696	0.9831	6	0	8	**		**
zernike(9,10)	29	36	0.9821	0.9841	0.9856	3	6	5			
first-order-theorem(1,2,3)	70	94	0.5057	0.5311	0.5127	2	4	5			
first-order-theorem(4,5,6)	56	69	0.5692	0.5914	0.5708	2	6	2			
gesturehaseSegmentation(DHP)	21	49	0.5513	0.6141	0.5532	1	8	1	***		***
gesturehaseSegmentation(RS)	83	93	0.6663	0.6942	0.6448	1	8	1			*
hillVally	23	45	0.5056	0.5291	0.6173	0	0	10		***	***
kc1	5	18	0.7031	0.6963	0.7029	2	3	6			
musk	37	71	0.8371	0.8547	0.8323	2	6	2			
ozone-level-8hr	16	34	0.7439	0.8203	0.7670	0	10	0	***		**
qsar-biodeg	33	57	0.7807	0.8317	0.8140	0	8	2	**	*	
semeion(1,2,3)	26	31	0.9781	0.9707	0.9771	6	0	4	**		*
semeion(4,5,6)	30	36	0.9864	0.9747	0.9864	5	0	6	***		***
semeion(7,8)	67	73	0.9861	0.9867	0.9891	3	3	9			
semeion(9,10)	56	58	0.9920	0.9910	0.9935	6	4	7			
spambase	38	56	0.7010	0.8979	0.7040	0	10	0	***		***
steel-plateds-fault	33	70	0.6311	0.9189	0.6280	0	10	0	***		***
wall-robot-navigation 12	33	58	0.8124	0.9773	0.8184	0	10	0	***		***
wall-robot-navigation 34	4	4	0.9801	0.9925	0.9796	1	10	1	***		***
mean	**34.8000**	**50.0333**	**0.7786**	**0.8405**	**0.8034**	**2.2000**	**5.4000**	**3.9000**			

*** p < 0.001, ** p < 0.01, * p < 0.05,† p < 0.1

Table 5: Results of the benchmarking experiment using the macro-averaged F-measure.
(Data with numerous instances and few variables)

the results of Tukey's honest significant difference method.

For data with numerous instances and few variables, rising the number of instances will bring advantages to RF and SVM. According to the result, the average of macro-averaged F-measure of FTA, RF, and SVM are 0.8034, 0.8405, 0.7786, respectively. The average numbers of wins of FTA, RF, and SVM are 3.9000, 5.4000, 2.2000, respectively. Besides, there exists significant difference between RF and FTA, SVM in almost half cases according to the results of Tukey's honest significant difference method. Therefore, RF is considered the best, followed by FTA and SVM.

4 The reasons why FTA works

The reasons why FTA works are concluded as follows:

1. Because FTA has feature selection process, FTA is expected to work better than SVM in dealing with data with numerous variables.
2. By gradually increasing the amount of training samples and updating the list of selected features, the same effect as repeated learning with different training data is obtained. FTA is expected to be superior to RF in handling data with few instances.
3. Introduction of the majority vote can ensure the high accuracy to a certain extend.

5 Conclusion

This study proposed FTA as a variable choice which is based on feature training. As proven in this benchmark study, FTA (1) provides more accurate models than RF and SVM in handling two types of challenging data which is difficult to make correct prediction for classifiers (i.e. data with few instances and variables, data with few instances and numerous variables), and data with numerous instances and variables; (2) For data with numerous instances and few variables, FTA ranks in the middle of RF and SVM; (3) This time only the well-balanced data was used, whereas, FTA may also work with data with high skew if IG is converted to BNS (Forman, 2003), which was previously shown to substantially improve classification accuracy, especially when dealing with tasks with high skew.

For the limitations of FTA, we do note that FTA is time-consuming especially when dealing with data with numerous instances or variables. This is considered primarily coming from SVM, the feature training and the number of runs in order to make the majority vote. The number of runs was set to 101 in this study, however it might be possible to further improve the model by automatically stopping FTA when a certain great model is made. Furthermore, all the features selected after training were used as the final list of selected features this time, the model may be further improved with a well set of the top *t* features as the final list of selected features.

Caigny et al. (2018) proposed the logit leaf model (LLM), which is constructed based on logistic regression and decision trees. In their experiment, LLM provides more accurate models than logistic regression and decision trees, and performs at least as well as RF and logistic model trees (LMT). As a feature work, the comparison will be made between LLM and FTA. Furthermore, the combination of SVM and IG will also be added as a comparison task in the future work.

References

Aisha Adel, Nazlia Omar, and Adel Al-Shabi. 2014. A comparative study of combined feature selection methods for arabic text classification. Journal of Computer Science, 10(11):2232–2239. https://thescipub.com/abstract/10.3844/jcssp.2014.2232.2239.

Leo Breiman. 2001. Random forests. Machine Learning, 45(1):5–32. https://link.springer.com/article/10.1023/A:1010933404324.

Arno De Caigny, Kristof Coussement, and Koen W. De Bock. 2018. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2):760-772. https://www.sciencedirect.com/science/article/pii/S0377221718301243.

Arunkumar Chinnaswamy and Ramakrishnan Srinivasan. 2017. Hybrid information gain based fuzzy roughest feature selection in cancer microarray data. In Proceedings of International Conference on Innovations in Power and Advanced Computing Technologies, pages 1–6, Canberra, Australia. https://ieeexplore.ieee.org/document/8244875.

Corinna Cortes and Vladimir Vapnik. 1995. Support vector networks. Machine Learning, 20(3):273– 297.

https://link.springer.com/article/10.1007/BF00994018.

Robert F. Chevalier, Gerrit Hoogenboom, Ronald W. McClendon, and Joel A. Paz. 2011. Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks. Neural Computing and Applications, 20(1):151–159. https://link.springer.com/article/10.1007/s00521-010-0363-y.

Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. 2018. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 19(270). https://doi.org/10.1186/s12859-018-2264-5.

S. Chehreh Chelgania and S. S. Matinb and James C.Howerc. 2016. Explaining relationships between coke quality index and coal properties by random forest method. Fuel, 182(15):754–760. https://www.sciencedirect.com/science/article/pii/S0016236116304860.

Giles M. Foody and Ajay Mathur. 2004. A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on Geoscience and Remote Sensing, pages 1335–1343, https://ieeexplore.ieee.org/abstract/document/1304900/authors#authors.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research. 3(2003):1289-1305. http://www.jmlr.org/papers/volume3/forman03a/forman03a.pdf.

Renate Geurts, Karl Ask, Pär Anders Granhag, and Aldert Vrij. 2018. Interviewing to manage threats: Exploring the effects of interview style on information gain and threateners' counter-interview strategies. Journal of Threat Assessment and Management, 5(4):189–204. https://psycnet.apa.org/record/2018-63621-001.

Alon Halevy, Peter Borvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2):8–12. https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/35179.pdf.

Emir Kremic and Abdulhamit Subasi. 2016. Performance of random forest and svm in face recognition. The International Arab Journal of Information Technology, 13(2):287–293. http://www.ccis2k.org/iajit/PDF/Vol.13,%20No.2/8468.pdf.

Miao Liu, Mingjun Wang, Jun Wanga, and Duo Li. 2013. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical, 177:970–980. https://www.sciencedirect.com/science/article/pii/S0925400512012671.

Ajay Mathur and Giles M. Foody. 2008. Crop classification by a support vector machine with intelligently selected training data for operational application. Computing Reviews, 24(11):503–512. https://www.tandfonline.com/doi/abs/10.1080/01431160701395203.

Phan Thanh Noi and Martin Kappas. 2018. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. Sensors, 18(1). https://doi.org/10.3390/s18010018.

Mahesh Pal and Paul M. Mather. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment. 86(4):554-564. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1526874.

Agnieszka Wosiak and Agata Dziomdziora. 2015. Feature selection and classification pairwise combinations for high-dimensional tumour biomedical datasets. Schedae Informaticae, 24:53–62. http://www.ejournals.eu/sj/index.php/SI/article/view/6334.

D. De Yong, S. Bhowmik, and F. Magnago. 2015. An effective power quality classifier using wavelet transform and support vector machines. Expert Systems with Applications, 42(15):6075–6081. https://www.sciencedirect.com/science/article/pii/S0957417415002328.

Bichen Zheng, Sang Won Yoon, and Sarah S.Lam. 2014. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Systems with Applications, 41(4):1476–1482. https://www.sciencedirect.com/science/article/pii/S0957417413006659.

Zhihua Zhou and Ji Feng. 2018. Deep forest. National Science Review, 6(1):74–86. https://arxiv.org/abs/1702.08835.