Context-Aware Feature Selection and Classification Supplementary Material

Juanyan Wang , Mustafa Bilgic
Illinois Institute of Technology
jwang245@hawk.iit.edu, mbilgic@iit.edu

A Parameter Settings

In this section, we provide the setting of the tunable parameters for all methods in detail.

For our approach CFSC, we experimented with $\lambda_a \in$ {0.05, 0.25, 0.5, 0.75, 0.95}.

The ATT-FL model requires one hyper-parameter, λ_{att}, that controls the weight of attention regularization in its loss function. We experimented with $\lambda_{att} \in$ {0.05, 0.25, 0.5, 0.75, 1.0}.

The RNP model has one hyper-parameter, λ_Ω, that controls the regularization on the number of selected features in the loss function. We experimented with $\lambda_\Omega \in$ {1e-2, 1e-3, 1e-4}.

For all neural network based feature selection models and FF, we trained them for 100 epochs and with the learning rate from the set {1e-2, 1e-3, 1e-4} and we chose the best settings via the validation set.

The DT model has four hyper-parameters: the maximum depth of the tree $\phi_{depth} \in$ {None, 1, 5, 10}, the minimum number of samples to split an internal node $\phi_{ss} \in$ {2, 10, 20, 200}, the minimum number of samples for a leaf node $\phi_{sl} \in$ {1, 5, 10, 100}, the maximum number of the total leaf nodes $\phi_l \in$ {None, 10, 50, 100} and the minimum decrease of the impurity $\phi_{imp} \in$ {0, 0.05, 0.1}.

The two rule-based learners (RL-P and RL-N), have five hyper-parameters: the number of optimization iterations $\psi_k \in$ {2, 4, 8}, the maximum number of rules $\psi_r \in$ {None, 5, 10, 20, 40}, the maximum number of conditions per rule $\psi_{cr} \in$ {None, 1, 2, 3}, the maximum number of total conditions $\psi_c \in$ {None, 5, 10, 20, 40}, and the maximum number of bins for fitting numeric attributes into discrete bins $\psi_{bin} \in$ {5, 10, 20}.

The LR model has one hyper-parameter, $C \in$ {0.01, 0.1, 1, 10, 100} for regularization.

The LR-PL model has two hyper-parameters, $C_y \in$ {0.01, 0.1, 1, 10, 100} and $C_a \in$ {0.01, 0.1, 1, 10, 100}, to regularize the two logistic regression models respectively.

B Classification and Feature Selection Performance

In the main paper, we presented the results of the combined measures. In this section, we supplement the main results with the separate classification performance and feature selection performance in Tables 1 and 2.

C Density

In this section, we supplement the density results under the evidence counterfactual strategy in our main paper with the density results under the decision tree strategy in Table 3. The trends for two simulation strategies are similar.

D Ablation Study

In this section, we present additional results of our ablation study using the decision tree strategy in Table 4.

E Significance Test Results

In this section, we present the significance test results comparing our approach to other baselines using the combined metrics in Tables 5 and 6.

Also, we provide the significance test results comparing the performance of the tuned CFSC to the other two settings, (λ_a=0.5, γ_a=0.5) and (λ_a=1, γ_a=1) in Tables 7 and 8.

F Company Feature List

In this section, we provide the full names for all features in Company dataset in Table 9.

		FF	LR	DT	RL-P	RL-N	ATT	RNP	LR-PL	ATT-FL	CFSC
Credit	*Clf.* F_1	**.910**	.856	.903	.811	.816	.731	.889	.806	.898	.886
	$Acc^{\mathbf{on}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$Acc^{\mathbf{off}}$	.000	.150	.992	.997	.998	.618	.806	**1.000**	.999	**1.000**
	$F_1{}^{\mathbf{w}}$	.504	.503	.499	.340	.260	.366	.460	.437	.592	**.684**
Company	*Clf.* F_1	**.824**	.605	.792	.337	.666	.679	.807	.388	.766	.771
	$Acc^{\mathbf{on}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$Acc^{\mathbf{off}}$	.000	.000	.910	**1.000**	**1.000**	.069	.573	**1.000**	.999	**1.000**
	$F_1{}^{\mathbf{w}}$	.355	.355	.120	.011	.005	.332	.326	.267	.450	**.631**
Mobile	*Clf.* F_1	**.988**	.985	.930	.900	.908	.927	.973	.908	.953	.957
	$Acc^{\mathbf{on}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$Acc^{\mathbf{off}}$	.000	.022	**1.000**	.993	.990	.958	.718	**1.000**	.998	**1.000**
	$F_1{}^{\mathbf{w}}$	.719	.719	.774	.531	.522	.630	.710	.662	.852	**.856**
NHIS	*Clf.* F_1	**.870**	.861	.808	.831	.836	.744	.840	.781	.739	.827
	$Acc^{\mathbf{on}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$Acc^{\mathbf{off}}$	.000	.283	.982	.998	.995	.911	.752	**1.000**	.830	**1.000**
	$F_1{}^{\mathbf{w}}$	.342	.335	.191	.065	.117	.250	.259	.588	.452	**.735**
Ride	*Clf.* F_1	**.861**	.852	.855	.720	.754	.824	.841	.736	.841	.832
	$Acc^{\mathbf{on}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$Acc^{\mathbf{off}}$	.000	.036	.969	**1.000**	.998	.997	.759	**1.000**	.998	**1.000**
	$F_1{}^{\mathbf{w}}$	.497	.497	.487	.264	.326	.358	.476	.409	.570	**.697**

Table 1: Classification and feature selection performance comparison. Feature labels are generated via the evidence counterfactual strategy.

		FF	LR	RL-P	RL-N	ATT	RNP	LR-PL	ATT-FL	CFSC
Credit	*Clf.* F_1	.908	.860	.835	.814	.652	.895	.810	.908	**.911**
	$Acc^{\mathbf{on}}$	**1.000**	**1.000**	.353	.377	.759	.904	**1.000**	**1.000**	.998
	$Acc^{\mathbf{off}}$	.000	.315	.994	.994	.244	.680	**1.000**	**1.000**	**1.000**
	$F_1{}^{\mathbf{w}}$	.684	.684	.382	.437	.609	.708	.925	.749	**.979**
Company	*Clf.* F_1	**.824**	.605	.339	.669	.676	.763	.616	.732	.800
	$Acc^{\mathbf{on}}$	**1.000**	**1.000**	.000	.000	.818	.756	**1.000**	.973	.993
	$Acc^{\mathbf{off}}$	.000	.052	.997	.999	.093	.653	**1.000**	.431	**1.000**
	$F_1{}^{\mathbf{w}}$	.896	.896	.000	.000	.861	.853	**.995**	.910	.989
Mobile	*Clf.* F_1	**.988**	.985	.899	.899	.949	.961	.908	.945	.939
	$Acc^{\mathbf{on}}$	**1.000**	**1.000**	.539	.499	.527	.999	**1.000**	**1.000**	.999
	$Acc^{\mathbf{off}}$	.000	.087	.998	.991	.954	.744	**1.000**	.999	**1.000**
	$F_1{}^{\mathbf{w}}$	.262	.262	.409	.447	.098	.279	.000	.256	**.867**
NHIS	*Clf.* F_1	**.870**	.861	.826	.821	.762	.846	.830	.657	.847
	$Acc^{\mathbf{on}}$	**1.000**	**1.000**	.054	.010	.122	.737	**1.000**	.567	**1.000**
	$Acc^{\mathbf{off}}$	.000	.423	.994	.994	.841	.843	**1.000**	.698	**1.000**
	$F_1{}^{\mathbf{w}}$	.646	.646	.015	.254	.272	.659	**.979**	.547	.971
Ride	*Clf.* F_1	**.860**	.856	.710	.757	.765	.847	.796	.850	.822
	$Acc^{\mathbf{on}}$	**1.000**	**1.000**	.261	.477	.311	.973	**1.000**	**1.000**	.997
	$Acc^{\mathbf{off}}$	.000	.481	.999	.998	.975	.841	**1.000**	**1.000**	**1.000**
	$F_1{}^{\mathbf{w}}$	.715	.715	.298	.493	.626	.715	**.946**	.837	.939
Synthetic1	*Clf.* F_1	**.979**	.837	.935	.787	.807	.944	.883	**.979**	.974
	$Acc^{\mathbf{on}}$	**1.000**	**1.000**	.120	.163	.621	.578	**1.000**	.978	.999
	$Acc^{\mathbf{off}}$	.000	.400	**1.000**	.988	.359	.680	**1.000**	.941	**1.000**
	$F_1{}^{\mathbf{w}}$	.826	.826	.330	.484	.731	.807	.958	.907	**.985**
Synthetic2	*Clf.* F_1	.869	.754	.735	.820	.662	.899	.745	.918	**.925**
	$Acc^{\mathbf{on}}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$Acc^{\mathbf{off}}$	.000	.233	.966	.963	.158	.641	**1.000**	.978	**1.000**
	$F_1{}^{\mathbf{w}}$	.639	.639	.317	.379	.606	.806	.877	.862	**.966**
Synthetic3	*Clf.* F_1	.955	.912	-	-	.617	.911	.928	.865	**.957**
	$Acc^{\mathbf{on}}$	**1.000**	**1.000**	-	-	.128	.468	**1.000**	.800	**1.000**
	$Acc^{\mathbf{off}}$	.000	.307	-	-	.866	.811	**1.000**	.985	**1.000**
	$F_1{}^{\mathbf{w}}$	.672	.757	-	-	.434	.779	.888	.389	**.971**

Table 2: Classification and feature selection performance comparison. Feature labels are generated using decision trees.

	FF	LR	RL-P	RL-N	ATT	RNP	LR-PL	ATT-FL	CFSC	*True*
Credit	37.0	26.6	1.1	1.2	28.1	13.7	2.4	2.2	2.5	2.5
Company	64.0	60.8	.2	.1	58.0	23.7	2.6	37.4	2.6	2.6
Mobile	20.0	18.6	.8	.9	1.7	7.1	1.0	1.2	1.4	1.4
NHIS	144.0	85.5	1.0	1.4	23.6	26.2	3.3	45.8	3.3	3.3
Ride	46.0	25.8	.6	1.1	2.8	10.4	2.4	2.3	2.4	2.4
Synthetic1	5.0	4.2	.5	.7	3.4	2.7	2.4	2.6	2.4	2.4
Synthetic2	10.0	8.6	.9	.9	8.5	4.1	1.6	1.8	1.8	1.8
Synthetic3	20.0	14.5	-	-	3.5	5.3	2.7	1.9	2.7	2.8

Table 3: Comparison on density measures. Feature labels are generated via decision trees. CFSC often had the closest density to the true values. ATT-FL and LR-PL had lower density than the true values in most cases.

		(λ_a=λ_a^*, γ_a=0.5)	(λ_a=0.5, γ_a=0.5)	(λ_a=1, γ_a=1)
Credit	*Clf.* F_1	.911	.913	.416
	Fea. F_1	.979	.957	.988
	Comb. F_1	.945	.935	.702
Company	*Clf.* F_1	.800	.805	.495
	Fea. F_1	.989	.979	.990
	Comb. F_1	.895	.892	.743
Mobile	*Clf.* F_1	.939	.943	.396
	Fea. F_1	.867	.872	.875
	Comb. F_1	.903	.908	.636
NHIS	*Clf.* F_1	.847	.849	.485
	Fea . F_1	.971	.944	.978
	Comb. F_1	.909	.897	.732
Ride	*Clf.* F_1	.822	.820	.462
	Fea. F_1	.939	.935	.947
	Comb. F_1	.881	.878	.705
Synthetic1	*Clf.* F_1	.974	.971	.516
	Fea . F_1	.985	.983	.988
	Comb. F_1	.980	.977	.752
Synthetic2	*Clf.* F_1	.925	.916	.558
	Fea. F_1	.966	.962	.972
	Comb. F_1	.946	.939	.765
Synthetic3	*Clf.* F_1	.957	.958	.257
	Fea. F_1	.971	.967	.976
	Comb. F_1	.964	.963	.617

Table 4: Comparison between different sets of λ_a and γ_a for CFSC. Feature labels are generated via decision trees.

	FF	LR	DT	RL-P	RL-N	ATT	RNP	LR-PL	ATT-FL
Credit	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
Company	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
Mobile	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	.205
NHIS	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
Ride	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**

Table 5: Statistical significance results on all datasets comparing CFSC to other baselines using the combined measure. Feature labels are generated using the evidence counterfactual strategy. Results where CFSC is statistically significantly better than other baselines (with p values less than 0.05) are boldfaced

	FF	LR	RL-P	RL-N	ATT	RNP	LR-PL	ATT-FL
Credit	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
Company	**.000**	**.000**	**.000**	**.000**	**.000**	**.001**	**.000**	**.020**
Mobile	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
NHIS	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.020**	**.000**
Ride	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.004**	**.000**
Synthetic1	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.043**
Synthetic2	**.000**	**.000**	**.000**	**.000**	**.000**	**.002**	**.000**	.057
Synthetic3	**.000**	**.000**	-	-	**.000**	**.000**	**.000**	**.000**

Table 6: Statistical significance results on all datasets comparing CFSC to other baselines using the combined measure. Feature labels are generated using the decision tree strategy. Results where CFSC is statistically significantly better than other baselines (with p values less than 0.05) are boldfaced

		(λ_a=0.5, γ_a=0.5)	(λ_a=1, γ_a=1)
Credit	*Clf.* F_1	.779	**.025**
	Fea. F_1	.114	<u>.959</u>
	Comb. F_1	.149	**.039**
Company	*Clf.* F_1	<u>.952</u>	**.000**
	Fea. F_1	**.000**	<u>.998</u>
	Comb. F_1	**.000**	**.000**
Mobile	*Clf.* F_1	.121	.068
	Fea. F_1	.566	.534
	Comb. F_1	.375	.064
NHIS	*Clf.* F_1	.156	**.004**
	Fea. F_1	**.010**	<u>.987</u>
	Comb. F_1	**.007**	**.006**
Ride	*Clf.* F_1	.845	**.009**
	Fea. F_1	**.037**	.304
	Comb. F_1	.322	**.008**

Table 7: Statistical significance results comparing tuned CFSC to the CFSC methods with fixed γ_a and λ_a values. Feature labels are generated using the evidence counterfactual strategy. The results where tuned CFSC performs significantly better than other methods (with p values less than 0.05) are bold-faced. The results where the other method is significantly better are underlined.

		(λ_a=0.5, γ_a=0.5)	(λ_a=1, γ_a=1)
Credit	*Clf.* F_1	.872	**.001**
	Fea. F_1	**.013**	.948
	Comb. F_1	**.012**	**.001**
Company	*Clf.* F_1	.701	**.000**
	Fea. F_1	**.006**	.813
	Comb. F_1	.322	**.000**
Mobile	*Clf.* F_1	.809	**.030**
	Fea. F_1	.701	.695
	Comb. F_1	.781	**.029**
NHIS	*Clf.* F_1	.587	**.003**
	Fea. F_1	**.025**	.940
	Comb. F_1	**.014**	**.003**
Ride	*Clf.* F_1	.383	**.002**
	Fea. F_1	.123	<u>.986</u>
	Comb. F_1	.111	**.002**
Synthetic1	*Clf.* F_1	.227	**.006**
	Fea. F_1	.075	.864
	Comb. F_1	.117	**.006**
Synthetic2	*Clf.* F_1	.118	**.008**
	Fea. F_1	.222	<u>.981</u>
	Comb. F_1	.133	**.008**
Synthetic3	*Clf.* F_1	.895	**.000**
	Fea. F_1	.104	.934
	Comb. F_1	.102	**.000**

Table 8: Statistical significance results comparing tuned CFSC to the CFSC methods with fixed γ_a and λ_a values. Feature labels are generated using decision trees. The results where tuned CFSC performs significantly better than other methods (with p values less than 0.05) are bold-faced. The results where the other method is significantly better are underlined.

Feature	Full Name
Attr1	net profit / total assets
Attr2	total liabilities / total assets
Attr3	working capital / total assets
Attr4	current assets / short-term liabilities
Attr5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
Attr6	retained earnings / total assets
Attr7	EBIT / total assets
Attr8	book value of equity / total liabilities
Attr9	sales / total assets
Attr10	equity / total assets
Attr11	(gross profit + extraordinary items + financial expenses) / total assets
Attr12	gross profit / short-term liabilities
Attr13	(gross profit + depreciation) / sales
Attr14	(gross profit + interest) / total assets
Attr15	(total liabilities * 365) / (gross profit + depreciation)
Attr16	(gross profit + depreciation) / total liabilities
Attr17	total assets / total liabilities
Attr18	gross profit / total assets
Attr19	gross profit / sales
Attr20	(inventory * 365) / sales
Attr21	sales (n) / sales (n-1)
Attr22	profit on operating activities / total assets
Attr23	net profit / sales
Attr24	gross profit (in 3 years) / total assets
Attr25	(equity - share capital) / total assets
Attr26	(net profit + depreciation) / total liabilities
Attr27	profit on operating activities / financial expenses
Attr28	working capital / fixed assets
Attr29	logarithm of total assets
Attr30	(total liabilities - cash) / sales
Attr31	(gross profit + interest) / sales
Attr32	(current liabilities * 365) / cost of products sold
Attr33	operating expenses / short-term liabilities
Attr34	operating expenses / total liabilities
Attr35	profit on sales / total assets
Attr36	total sales / total assets
Attr37	(current assets - inventories) / long-term liabilities
Attr38	constant capital / total assets
Attr39	profit on sales / sales
Attr40	(current assets - inventory - receivables) / short-term liabilities
Attr41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
Attr42	profit on operating activities / sales
Attr43	rotation receivables + inventory turnover in days
Attr44	(receivables * 365) / sales
Attr45	net profit / inventory
Attr46	(current assets - inventory) / short-term liabilities
Attr47	(inventory * 365) / cost of products sold
Attr48	EBITDA (profit on operating activities - depreciation) / total assets
Attr49	EBITDA (profit on operating activities - depreciation) / sales
Attr50	current assets / total liabilities
Attr51	short-term liabilities / total assets
Attr52	(short-term liabilities * 365) / cost of products sold)
Attr53	equity / fixed assets
Attr54	constant capital / fixed assets
Attr55	working capital
Attr56	(sales - cost of products sold) / sales
Attr57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
Attr58	total costs /total sales
Attr59	long-term liabilities / equity
Attr60	sales / inventory
Attr61	sales / receivables
Attr62	(short-term liabilities *365) / sales
Attr63	sales / short-term liabilities
Attr64	sales / fixed assets

Table 9: Feature list of the Company dataset.