Geometric SMOTENC

A geometrically enhanced drop-in replacement for SMOTENC

Joao Fonseca[1*], Georgios Douzas[1], Fernando Bacao[1]

[1]NOVA Information Management School, Universidade Nova de Lisboa

*Corresponding Author

Postal Address: NOVA Information Management School, Campus de Campolide, 1070–312 Lisboa, Portugal

Telephone: +351 21 382 8610

This is an abstract.

1. Introduction

This is text [1].

Table 1: Description of the datasets collected after data preprocessing. The sampling strategy is similar across datasets. Legend: (IR) Imbalance Ratio

Dataset	Metric	Non-Metric	Obs.	Min. Obs.	Maj. Obs.	IR	Classes
ABALONE	1	7	4139	15	689	45.93	18
ADULT	8	6	5000	1268	3732	2.94	2
ADULT (10)	8	6	5000	451	4549	10.09	2
ANNEALING	4	6	790	34	608	17.88	4
CENSUS	24	7	5000	337	4663	13.84	2
CONTRACEPTIVE	4	5	1473	333	629	1.89	3
CONTRACEPTIVE (10)	4	5	1036	62	629	10.15	3
CONTRACEPTIVE (20)	4	5	990	31	629	20.29	3
CONTRACEPTIVE (31)	4	5	973	20	629	31.45	3
CONTRACEPTIVE (41)	4	5	966	15	629	41.93	3
COVERTYPE	2	10	5000	20	2449	122.45	7
CREDIT APPROVAL	9	6	653	296	357	1.21	2
GERMAN CREDIT	13	7	1000	300	700	2.33	2
GERMAN CREDIT (10)	13	7	770	70	700	10.00	2
GERMAN CREDIT (20)	13	7	735	35	700	20.00	2
GERMAN CREDIT (30)	13	7	723	23	700	30.43	2
GERMAN CREDIT (41)	13	7	717	17	700	41.18	2
HEART DISEASE	5	5	740	22	357	16.23	5
HEART DISEASE (21)	5	5	735	17	357	21.00	5
THYROID	22	6	5000	1376	3624	2.63	2

Continued on next page

Table 1: Description of the datasets collected after data preprocessing. The sampling strategy is similar across datasets. Legend: (IR) Imbalance Ratio

Dataset	Metric	Non-Metric	Obs.	Min. Obs.	Maj. Obs.	IR	Classes
THYROID (10)	22	6	4584	416	4168	10.02	2
THYROID (101)	22	6	4209	41	4168	101.66	2
THYROID (20)	22	6	4376	208	4168	20.04	2
THYROID (30)	22	6	4306	138	4168	30.20	2
THYROID (40)	22	6	4272	104	4168	40.08	2
THYROID (50)	22	6	4251	83	4168	50.22	2
THYROID (60)	22	6	4237	69	4168	60.41	2
THYROID (70)	22	6	4227	59	4168	70.64	2
THYROID (80)	22	6	4220	52	4168	80.15	2
THYROID (90)	22	6	4214	46	4168	90.61	2

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

A. Appendix

Table 2: Wide optimal results

Dataset	Classifier	Metric	G-SMOTE	NONE	RAND-OVER	RAND-UNDER	SMOTENC
Abalone	DT	OA	0.221	**0.256**	0.203	0.207	0.190
Abalone	DT	F-Score	0.168	**0.170**	0.154	0.132	0.156
Abalone	DT	G-Mean	**0.460**	0.413	0.457	0.421	0.445
Abalone	KNN	OA	0.215	**0.237**	0.197	0.188	0.186
Abalone	KNN	F-Score	**0.167**	0.157	0.151	0.140	0.150
Abalone	KNN	G-Mean	**0.429**	0.391	0.397	0.421	0.409
Abalone	LR	OA	0.235	**0.272**	0.229	0.195	0.228
Abalone	LR	F-Score	**0.189**	0.180	0.179	0.166	0.186
Abalone	LR	G-Mean	**0.473**	0.415	0.456	0.441	0.466
Abalone	RF	OA	0.237	**0.276**	0.224	0.197	0.221
Abalone	RF	F-Score	**0.194**	0.174	0.184	0.162	0.180
Abalone	RF	G-Mean	**0.486**	0.416	0.465	0.448	0.461
Adult	DT	OA	0.830	**0.835**	0.800	0.785	0.785
Adult	DT	F-Score	**0.767**	0.763	0.755	0.744	0.754
Adult	DT	G-Mean	**0.809**	0.747	0.806	0.801	0.808
Adult	KNN	OA	0.786	**0.805**	0.763	0.761	0.781
Adult	KNN	F-Score	**0.738**	0.732	0.718	0.728	0.735
Adult	KNN	G-Mean	0.766	0.724	0.757	**0.780**	0.762
Adult	LR	OA	0.803	**0.839**	0.804	0.801	0.803
Adult	LR	F-Score	0.768	**0.773**	0.771	0.769	0.767

Continued on next page

Table 2: Wide optimal results

Dataset	Classifier	Metric	G-SMOTE	NONE	RAND-OVER	RAND-UNDER	SMOTENC
Adult	LR	G-Mean	0.813	0.758	**0.815**	**0.815**	0.805
Adult	RF	OA	0.820	**0.832**	0.755	0.753	0.757
Adult	RF	F-Score	**0.769**	0.739	0.729	0.728	0.727
Adult	RF	G-Mean	0.796	0.711	**0.797**	**0.797**	0.787
Annealing	DT	OA	0.828	**0.843**	0.742	0.676	0.742
Annealing	DT	F-Score	**0.741**	0.643	0.731	0.665	0.732
Annealing	DT	G-Mean	**0.915**	0.738	0.905	0.874	0.909
Annealing	KNN	OA	0.849	0.848	**0.856**	0.477	0.829
Annealing	KNN	F-Score	0.778	0.726	**0.785**	0.453	0.747
Annealing	KNN	G-Mean	0.899	0.783	**0.909**	0.797	0.867
Annealing	LR	OA	0.576	**0.814**	0.570	0.489	0.573
Annealing	LR	F-Score	0.618	0.540	**0.623**	0.481	0.617
Annealing	LR	G-Mean	**0.850**	0.663	0.848	0.810	0.843
Annealing	RF	OA	**0.870**	0.868	0.716	0.633	0.737
Annealing	RF	F-Score	**0.796**	0.644	0.723	0.637	0.732
Annealing	RF	G-Mean	**0.914**	0.727	0.906	0.881	0.906
Census	DT	OA	0.942	**0.943**	0.844	0.795	0.894
Census	DT	F-Score	**0.733**	0.731	0.652	0.617	0.693
Census	DT	G-Mean	0.813	0.698	0.814	**0.817**	0.800
Census	KNN	OA	0.874	**0.933**	0.878	0.731	0.867
Census	KNN	F-Score	0.652	0.648	0.640	0.567	**0.655**
Census	KNN	G-Mean	0.767	0.620	0.733	**0.794**	0.768
Census	LR	OA	0.940	**0.949**	0.940	0.815	0.938
Census	LR	F-Score	0.760	0.743	**0.762**	0.639	0.760
Census	LR	G-Mean	0.807	0.707	0.801	**0.837**	0.782
Census	RF	OA	0.876	**0.933**	0.740	0.714	0.819
Census	RF	F-Score	**0.679**	0.483	0.580	0.562	0.636
Census	RF	G-Mean	**0.827**	0.500	0.822	0.814	0.818
Contraceptive	DT	OA	**0.563**	0.538	0.512	0.525	0.537
Contraceptive	DT	F-Score	**0.549**	0.518	0.507	0.520	0.529
Contraceptive	DT	G-Mean	**0.661**	0.630	0.630	0.641	0.646
Contraceptive	KNN	OA	0.465	**0.478**	0.435	0.468	0.455
Contraceptive	KNN	F-Score	0.460	**0.462**	0.432	0.461	0.450
Contraceptive	KNN	G-Mean	0.588	0.580	0.566	**0.590**	0.579
Contraceptive	LR	OA	**0.515**	0.514	0.510	0.510	0.514
Contraceptive	LR	F-Score	**0.512**	0.492	0.505	0.506	0.509
Contraceptive	LR	G-Mean	**0.635**	0.604	0.628	0.627	0.631
Contraceptive	RF	OA	0.553	**0.557**	0.534	0.526	0.540
Contraceptive	RF	F-Score	**0.545**	0.524	0.529	0.522	0.535
Contraceptive	RF	G-Mean	**0.659**	0.634	0.649	0.643	0.653
Covertype	DT	OA	0.580	**0.705**	0.567	0.450	0.587
Covertype	DT	F-Score	0.484	**0.490**	0.475	0.361	0.481
Covertype	DT	G-Mean	**0.769**	0.671	0.758	0.700	0.758
Covertype	KNN	OA	0.690	**0.700**	0.699	0.454	0.683
Covertype	KNN	F-Score	0.532	0.457	**0.561**	0.367	0.535
Covertype	KNN	G-Mean	0.745	0.642	**0.763**	0.691	0.753
Covertype	LR	OA	0.637	**0.721**	0.611	0.472	0.640
Covertype	LR	F-Score	0.516	0.507	0.492	0.353	**0.526**

Continued on next page

Table 2: Wide optimal results

Dataset	Classifier	Metric	G-SMOTE	NONE	RAND-OVER	RAND-UNDER	SMOTENC
Covertype	LR	G-Mean	**0.792**	0.678	0.790	0.697	0.786
Covertype	RF	OA	0.598	**0.704**	0.587	0.485	0.583
Covertype	RF	F-Score	0.517	0.360	**0.519**	0.394	0.507
Covertype	RF	G-Mean	0.800	0.572	**0.804**	0.737	0.799
Credit Approval	DT	OA	**0.867**	0.847	0.861	0.865	0.862
Credit Approval	DT	F-Score	**0.867**	0.845	0.861	0.865	0.862
Credit Approval	DT	G-Mean	**0.874**	0.848	0.867	0.872	0.869
Credit Approval	KNN	OA	**0.870**	0.865	**0.870**	0.865	0.868
Credit Approval	KNN	F-Score	**0.869**	0.864	**0.869**	0.864	0.867
Credit Approval	KNN	G-Mean	**0.871**	0.865	**0.871**	0.866	0.868
Credit Approval	LR	OA	0.873	0.868	**0.874**	0.873	0.871
Credit Approval	LR	F-Score	0.873	0.868	**0.874**	0.873	0.871
Credit Approval	LR	G-Mean	0.877	0.873	**0.879**	0.878	0.877
Credit Approval	RF	OA	0.876	**0.877**	0.868	0.868	0.871
Credit Approval	RF	F-Score	0.876	**0.877**	0.868	0.868	0.871
Credit Approval	RF	G-Mean	**0.879**	**0.879**	0.872	0.873	0.876
German Credit	DT	OA	0.704	**0.713**	0.660	0.644	0.702
German Credit	DT	F-Score	**0.662**	0.608	0.633	0.623	0.654
German Credit	DT	G-Mean	**0.681**	0.608	0.663	0.660	0.667
German Credit	KNN	OA	0.681	**0.718**	0.670	0.641	0.682
German Credit	KNN	F-Score	**0.653**	0.628	0.636	0.616	0.650
German Credit	KNN	G-Mean	**0.675**	0.621	0.656	0.642	0.668
German Credit	LR	OA	0.727	**0.751**	0.724	0.712	0.729
German Credit	LR	F-Score	0.695	0.681	**0.697**	0.686	**0.697**
German Credit	LR	G-Mean	**0.722**	0.672	0.720	0.713	0.713
German Credit	RF	OA	**0.760**	0.741	0.737	0.700	0.739
German Credit	RF	F-Score	0.701	0.580	**0.709**	0.680	0.702
German Credit	RF	G-Mean	0.715	0.588	**0.730**	0.719	0.716
Heart Disease	DT	OA	0.532	**0.566**	0.473	0.430	0.509
Heart Disease	DT	F-Score	**0.371**	0.322	0.331	0.295	0.342
Heart Disease	DT	G-Mean	**0.588**	0.534	0.545	0.515	0.563
Heart Disease	KNN	OA	0.538	**0.564**	0.534	0.504	0.535
Heart Disease	KNN	F-Score	**0.363**	0.287	0.352	0.341	0.360
Heart Disease	KNN	G-Mean	**0.571**	0.509	0.560	0.557	**0.571**
Heart Disease	LR	OA	0.558	**0.584**	0.536	0.480	0.557
Heart Disease	LR	F-Score	**0.397**	0.329	0.374	0.333	0.395
Heart Disease	LR	G-Mean	0.601	0.539	**0.603**	0.567	0.601
Heart Disease	RF	OA	0.553	**0.601**	0.539	0.480	0.546
Heart Disease	RF	F-Score	**0.385**	0.314	0.360	0.326	0.366
Heart Disease	RF	G-Mean	**0.600**	0.531	0.569	0.566	0.580
Thyroid	DT	OA	0.952	**0.953**	0.946	0.948	0.952
Thyroid	DT	F-Score	**0.942**	0.941	0.935	0.936	0.941
Thyroid	DT	G-Mean	0.953	0.940	0.948	0.950	**0.955**
Thyroid	KNN	OA	0.836	**0.840**	0.830	0.810	0.831
Thyroid	KNN	F-Score	**0.791**	0.778	0.790	0.769	0.785
Thyroid	KNN	G-Mean	**0.795**	0.755	0.794	0.776	0.781
Thyroid	LR	OA	0.776	**0.818**	0.775	0.775	0.771
Thyroid	LR	F-Score	**0.733**	0.721	0.732	**0.733**	0.728

Continued on next page

Table 2: Wide optimal results

Dataset	Classifier	Metric	G-SMOTE	NONE	RAND-OVER	RAND-UNDER	SMOTENC
Thyroid	LR	G-Mean	0.748	0.693	0.747	**0.749**	0.742
Thyroid	RF	OA	0.942	0.925	**0.944**	0.939	0.939
Thyroid	RF	F-Score	0.928	0.902	**0.931**	0.925	0.925
Thyroid	RF	G-Mean	0.936	0.886	**0.938**	0.934	0.932