Simplicial SMOTE: Oversampling Solution to the Imbalanced Learning Problem

Oleg Kachan[1,2], Andrey Savchenko [1,2], Gleb Gusev[1]

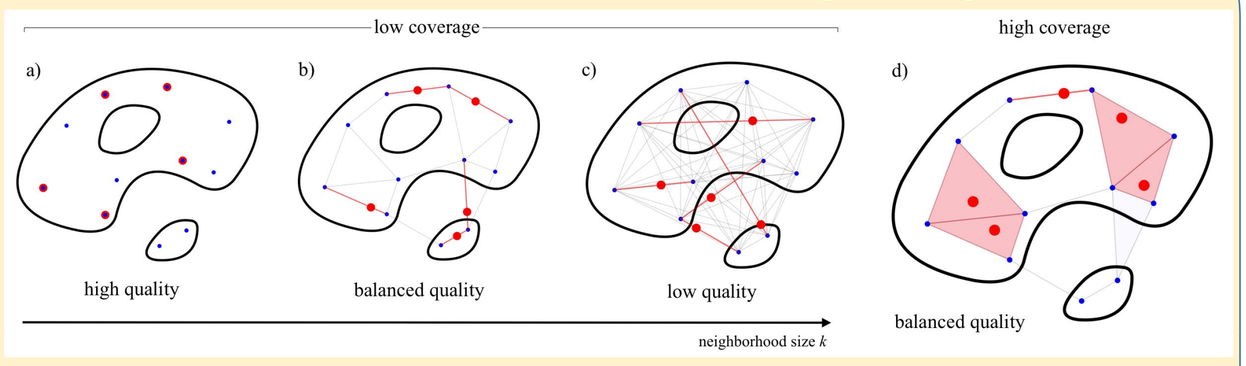
1 Sber AI Lab. 2 HSE University

Contact: avsavchenko@hse.ru

Abstract

SMOTE (Synthetic Minority Oversampling Technique) is the established geometric approach to random oversampling to balance classes in the imbalanced learning problem, followed by many extensions. Its idea is to introduce synthetic data points of the minor class, with each new point being the convex combination of an existing data point and one of its *k*-nearest neighbors. In this paper, by viewing SMOTE as sampling from the edges of a geometric neighborhood graph and borrowing tools from the topological data analysis, we propose a novel technique, Simplicial SMOTE, that samples from the simplices of a geometric neighborhood simplicial complex. A new synthetic point is defined by the barycentric coordinates w.r.t. a simplex spanned by an arbitrary number of data points being sufficiently close rather than a pair. Such a replacement of the geometric data model results in better coverage of the underlying data distribution compared to existing geometric sampling methods and allows the generation of synthetic points of the minority class closer to the majority class on the decision boundary. We experimentally demonstrate that our Simplicial SMOTE outperforms several popular geometric sampling methods, including the original SMOTE. Moreover, we show that simplicial sampling can be easily integrated into existing SMOTE extensions. We generalize and evaluate simplicial extensions of the classic Borderline SMOTE, Safe-level SMOTE, and ADASYN algorithms, all of which outperform their graph-based counterparts.

Geometric oversampling



a) random oversampling, b) SMOTE, c) global sampling, d) Simplicial SMOTE.

Random oversampling just duplicates existing points. Assuming that synthetic data points lie within a convex hull of existing points, global methods do not respect the intrinsic properties of data such as clusters and holes, resulting in low sample quality. While SMOTE, being a local method, improves on this, it still models the data with a union of one-dimensional segments, unable to sample all of the data support. Simplicial SMOTE, by modeling data with a union of higher-dimensional simplices, samples dense areas of the data support while avoiding sampling from topological holes, effectively improving coverage of the data distribution.

Proposed algorithm

Algorithm 1: Simplicial SMOTE

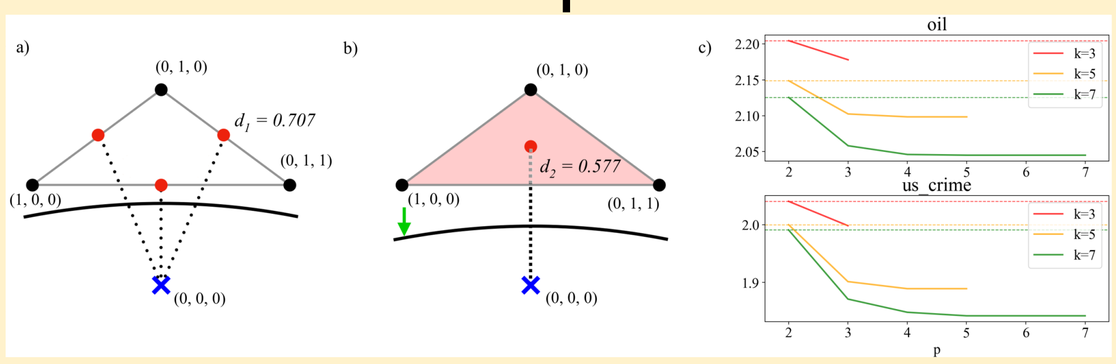
Input : Minority class points $\mathbf{X}^+$.
Parameters : Neighborhood size k, maximal relation arity $p \geq k$,
Output : Synthetic minority class points $\hat{\mathbf{X}}^+$.

1 Construct a k-NN neighborhood graph $G_k(X^+)$.
2 Compute a p-skeleton $(K_p \circ G_k)(X^+)$ of a clique complex $(K \circ G_k)(X^+)$, get its maximal simplices Σ_p^{MAX}
3 Sample $m = n^- - n^+$ simplices $\sigma_i^{(p_i)}$ of dimension p_i, $\Sigma = \{\sigma_i^{(p_i)}\}_{i \in 1,\cdots,m}$ from Σ_p^{MAX}.
4 **for** $i \in 1, \ldots, m$ **do**
5 Sample barycentric coordinates $\boldsymbol{\lambda}_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (1, \ldots, 1) \in \mathbb{R}^{(p+1)}$.
6 Compute Euclidean coordinates $\hat{\mathbf{x}}_i = \boldsymbol{\lambda}_i^T \mathbf{X}_i$ w.r.t. a simplex $\sigma_i^{(p_i)} = (\mathbf{x}_0, \ldots, \mathbf{x}_{p_i}) \in \Sigma$ of dimension p_i.
7 **return** $\{\hat{\mathbf{x}}_i\}_{i \in 1,\ldots,m}$

Neighborhood graphs

$$R^{kNN}(k) = \{(x,y) \mid d(x,y) \leq \min_k d(x,z),\ z \in X\} \qquad R^{\varepsilon}(\varepsilon) = \{(x,y) \mid d(x,y) \leq \varepsilon\}$$

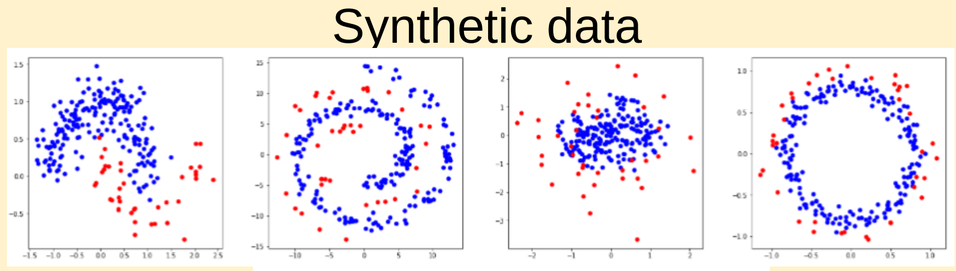
SMOTE vs Simplicial SMOTE



For the configuration of three points of the minor class (black circles) equidistant to a point of the major class (blue cross) b) Simplicial SMOTE will generate synthetic points of the minor class (red circles) closer to the point of the major class (projection distance to the 2-simplex $d_2 = 0.577$), than a) SMOTE (projection distance to any edge $d_1 = 0.707$), effectively moving the local decision boundary. c) Mean projection distance to the geometric model of minority class gets smaller with increasing maximal relation arity parameter p. Distance to the simplicial model is shown as solid lines for different values of neighborhood size parameter k, distance to the graph model is shown as a dashed line of the same color.

Experimental results

Synthetic data



	Imbalanced	Gaussian	Random	Global	SMOTE	Simplicial
moons	0.9511	0.8830	0.9485	0.9348	**0.9694**	**0.9694**
swiss_rolls	0.5317	0.6673	0.7168	0.6774	**0.7208**	0.6823
g_circle	0.7129	0.6750	0.7089	0.6542	0.6937	**0.7269**
circles	0.6541	0.7060	0.6777	0.6356	0.7005	**0.7139**
rank	4.0000	4.5000	3.2500	5.2500	2.3750	**1.6250**

Real data, k-NN, F1-score

	Imbalanced	Random	Global	SMOTE	Border.	Safelevel	ADASYN	MWMOTE	DBSMOTE	LVQ	Simplicial	S-Border.	S-Safe.	S-ADASYN
ecoli	0.5628	0.5735	0.6048	0.5965	0.5696	0.5718	0.5875	0.6024	0.5893	0.5767	0.6282	0.6003	0.5839	0.6230
optical_digits	0.5586	0.6698	0.7381	0.7193	0.6779	0.7178	0.6824	0.7269	0.6717	0.6627	0.7551	0.6876	0.7624	0.7339
pen_digits	0.6719	0.8017	0.6951	0.8110	0.7006	0.8118	0.6857	0.7270	0.8044	0.8005	0.8290	0.6916	0.8278	0.7268
abalone	0.0000	0.3700	0.3983	0.3769	0.3792	0.3799	0.3716	0.3879	0.3785	0.3747	0.3883	0.3950	0.3865	0.3847
sick_euthyroid	0.8494	0.8243	0.8214	0.8288	0.8247	0.7334	0.8273	0.8297	0.8397	0.8109	0.8382	0.8321	0.8401	0.8310
spectrometer	0.6129	0.7237	0.6315	0.7186	0.7453	0.7697	0.7025	0.6878	0.7828	0.6183	0.8068	0.7456	0.7426	0.7792
car_eval_34	0.2588	0.6426	0.7485	0.7058	0.7120	0.6743	0.7187	0.6990	0.6429	0.5837	0.7278	0.7131	0.7278	0.7019
us_crime	0.4243	0.4639	0.4692	0.4702	0.4787	0.4753	0.4623	0.4557	0.4652	0.4455	0.4723	0.4814	0.4560	0.4575
yeast_ml8	0.0000	0.1320	0.1560	0.1484	0.1502	0.1565	0.1445	0.1423	0.1386	0.1271	0.1527	0.1477	0.1538	0.1451
scene	0.0109	0.2549	0.2528	0.2617	0.2578	0.2543	0.2552	0.2580	0.0705	0.0477	0.2352	0.2490	0.2368	0.2408
libras_move	0.4906	0.6951	0.6548	0.6638	0.6678	0.6398	0.6333	0.6510	0.6802	0.6834	0.7003	0.6769	0.6512	0.6878
thyroid_sick	0.8334	0.7835	0.7323	0.7920	0.7857	0.6269	0.7846	0.7381	0.8075	0.7323	0.7916	0.7840	0.7845	0.7854
coil_2000	0.0000	0.2120	0.2248	0.2184	0.2166	0.2074	0.2150	0.2199	0.0811	0.0101	0.2092	0.2165	0.2095	0.2092
solar_flare_m0	0.0164	0.1959	0.2459	0.1828	0.1918	0.1917	0.1754	0.1923	0.0659	0.1572	0.1712	0.1807	0.1701	0.1708
oil	0.3640	0.3905	0.3752	0.3659	0.3993	0.3904	0.3585	0.3487	0.3782	0.3906	0.4600	0.4522	0.4064	0.4418
car_eval_4	0.0000	0.4061	0.5011	0.4387	0.4383	0.4290	0.4213	0.4280	0.4034	0.5403	0.4696	0.4750	0.4696	0.4403
wine_quality	0.0764	0.2317	0.1821	0.2091	0.2246	0.2191	0.1949	0.2006	0.1765	0.2753	0.2015	0.2241	0.2104	0.1842
letter_img	0.6064	0.4611	0.5567	0.5507	0.4252	0.5268	0.4499	0.4832	0.5624	0.5206	0.6195	0.4331	0.6236	0.5385
yeast_me2	0.0972	0.2700	0.2836	0.2768	0.3272	0.2999	0.2610	0.2935	0.2727	0.3121	0.3071	0.3366	0.2858	0.2930
ozone_level	0.0528	0.2384	0.2198	0.2354	0.2633	0.2240	0.2280	0.2402	0.2393	0.2395	0.2846	0.2823	0.2559	0.2775
abalone_19	0.0000	0.0471	0.0367	0.0439	0.0565	0.0448	0.0448	0.0591	0.0415	0.0390	0.0442	0.0522	0.0499	0.0460
mean	0.3089	0.4470	0.4537	0.4578	0.4520	0.4450	0.4383	0.4462	0.4330	0.4261	**0.4806**	0.4599	**0.4683**	0.4618
rank	12.1429	8.4286	6.9524	6.8095	6.2381	7.8095	9.4286	7.3810	8.5238	9.4762	**4.1429**	**5.3810**	5.8095	6.4762

Real data, Gradient boosting, F1-score

	Imbalanced	Random	Global	SMOTE	Border.	Safelevel	ADASYN	MWMOTE	DBSMOTE	LVQ	Simplicial	S-Border.	S-Safe.	S-ADASYN
ecoli	0.5780	0.5501	0.5864	0.5822	0.5853	0.5781	0.5705	0.5980	0.6336	0.5827	0.6275	0.6151	0.5688	0.6280
optical_digits	0.9670	0.9491	0.9427	0.9415	0.9559	0.9439	0.9442	0.9376	0.9491	0.9618	0.9443	0.9557	0.9425	0.9423
pen_digits	0.9927	0.9906	0.9895	0.9906	0.9925	0.9900	0.9917	0.9907	0.9922	0.9928	0.9915	0.9925	0.9912	0.9911
abalone	0.1808	0.3326	0.3842	0.3501	0.3586	0.3727	0.3448	0.3753	0.3281	0.3078	0.3698	0.3725	0.3614	0.3594
sick_euthyroid	0.5565	0.5694	0.5872	0.5708	0.5684	0.5246	0.5650	0.5654	0.6081	0.5800	0.6049	0.5981	0.5962	0.6047
spectrometer	0.7618	0.8493	0.8382	0.8430	0.8543	0.8274	0.8423	0.8551	0.7968	0.8209	0.8548	0.8485	0.8386	0.8465
car_eval_34	0.6018	0.5830	0.5718	0.5774	0.5913	0.5886	0.5851	0.6490	0.5830	0.7165	0.6296	0.6081	0.6321	0.6341
us_crime	0.3676	0.4429	0.4404	0.4188	0.4567	0.4346	0.4144	0.4062	0.4429	0.4634	0.4313	0.4687	0.4210	0.4236
yeast_ml8	0.0375	0.1426	0.1613	0.1592	0.1659	0.1542	0.1578	0.1603	0.1426	0.1643	0.1598	0.1658	0.1601	0.1586
scene	0.1004	0.2500	0.2499	0.2386	0.2471	0.2452	0.2309	0.2364	0.1101	0.2579	0.2251	0.2482	0.2219	0.2259
libras_move	0.6997	0.8031	0.7702	0.7672	0.7661	0.7362	0.7519	0.7874	0.8031	0.8029	0.7560	0.7568	0.7640	0.7525
thyroid_sick	0.4966	0.5239	0.5246	0.5255	0.5307	0.4620	0.5214	0.5245	0.5095	0.5031	0.5504	0.5459	0.5271	0.5579
coil_2000	0.0457	0.1745	0.1726	0.1709	0.1759	0.1677	0.1711	0.1748	0.0533	0.1168	0.1714	0.1748	0.1703	0.1743
solar_flare_m0	0.0510	0.2192	0.2043	0.2077	0.2262	0.2280	0.2188	0.2120	0.0486	0.2126	0.2189	0.2337	0.2340	0.2171
oil	0.3156	0.4467	0.4592	0.4428	0.4674	0.3784	0.4240	0.4191	0.4467	0.4626	0.5074	0.5062	0.4267	0.4777
car_eval_4	0.1294	0.3815	0.4810	0.4443	0.4472	0.4121	0.4371	0.5240	0.3815	0.6771	0.5749	0.5773	0.6052	0.5676
wine_quality	0.1292	0.2983	0.2097	0.2558	0.2774	0.2460	0.2537	0.2163	0.1487	0.2244	0.2533	0.2741	0.2731	0.2534
letter_img	0.9722	0.9526	0.9086	0.9410	0.9608	0.9293	0.9531	0.9128	0.9661	0.9673	0.9556	0.9610	0.9547	0.9491
yeast_me2	0.2296	0.3192	0.2705	0.3043	0.3621	0.2943	0.3018	0.3227	0.2952	0.2890	0.3364	0.3759	0.3057	0.3279
ozone_level	0.1608	0.2528	0.2086	0.2087	0.2270	0.2516	0.2095	0.2108	0.2528	0.2352	0.2111	0.2443	0.2056	0.2046
abalone_19	0.0000	0.0277	0.0500	0.0377	0.0482	0.0289	0.0408	0.0366	0.0212	0.0312	0.0512	0.0452	0.0434	0.0498
mean	0.3988	0.4790	0.4767	0.4751	0.4888	0.4664	0.4729	0.4817	0.4530	0.4938	**0.4964**	**0.5032**	0.4878	0.4927
rank	11.5714	7.3095	8.0476	9.0000	**4.6905**	9.7143	9.2381	7.5714	8.5476	6.2381	5.3333	**3.6429**	7.4286	6.6667

Paper

Code