Supplementary materials

Hui Yu[a,*], QiaoFeng Wang[a], JianYu Shi[b,*]

[a] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

[b] School of Life Sciences, Northwestern Polytechnical University, Xi'an 710072, China

The supplementary material contains the evaluation metrics used in the experiment and the experimental results on the other five datasets.

1 Evaluation Metrics

We evaluate the clustering performance with five standard clustering evaluation metrics, i.e, Accuracy (ACC), F-measure (F1), Jaccard Index (JI), Precsion and Recall. The brief introduction of the five indices is given below.

Accuracy defines the ratio of the number of correctly predicted samples to the total number of predicted samples, expressed as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

where TP, FP, TN and FN are the values of the confusion matrix in classification, which denote the correct instances, false positive examples, correct negative examples and false negative examples, respectively.

Precision refers to the ratio of the number of correctly predicted positive samples to the number of all predicted positive samples, expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall refers to the ratio of the number of positive samples predicted by the model to the actual number of positive samples, expressed as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 combines the two indicators of accuracy and recall rate. Only when the accuracy and recall rate is relatively high can the model get a better F1 value, which is defined as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

The Jaccard index is used to quantify the similarity between two sets, and the value ranges from 0 to 1, where the larger the value, the more similar the two sets are, specifically expressed as:

$$JI(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN} \tag{5}$$

*Corresponding author: huiyu@nwpu.edu.cn, jianyushi@nwpu.edu.cn

2 Performance results

This section provides additional information including the 2 real datasets and 3 artificial datasets (see Table S1) and the performance results of the above five datasets(see Table S2∼Table S6). Each table corresponds to the evaluation results on one dataset, and the best results are highlighted in bold and the minor results are underlined. According to the results, GMM_WGAN performs best compared to other data augmentation methods under all metrics on the five datasets.

Table. S1. Relevant information of the five datasets

Type	Dataset	Data size	Dimension	Cluster
real dataset	Wine	178	13	3
	Seeds	210	7	3
	Thyroid	215	5	3
artificial dataset	Pathbased	300	2	3
	R15	600	2	15

Table. S2. Thyroid dataset

Algorithm	Model	ACC	F1	JI	Precision	Recall
DPC	None	0.76168	0.69556	0.58414	0.82245	0.76168
	SMOTE	0.7757	0.73084	0.61382	0.78573	0.7757
	ADASYN	0.7907	0.72722	0.64382	0.69033	0.7907
	GAN	0.76279	0.69688	0.58577	0.82298	0.76279
	WGAN	0.73488	0.65185	0.54277	0.80789	0.73488
	EWGAN	0.75426	0.67452	0.55733	0.81457	0.75382
	PCWGAN-GP	0.77459	0.69344	0.57248	0.83682	0.7734
	GMM_WGAN	**0.81395**	**0.75937**	**0.66707**	**0.85312**	**0.81395**
k_means	None	0.88785	0.87713	0.79126	0.90341	0.88785
	SMOTE	0.88785	0.87713	0.79126	0.90341	0.88785
	ADASYN	0.88837	0.87768	0.79214	0.90377	0.88837
	GAN	0.87442	0.85732	0.76829	0.89098	0.87442
	WGAN	0.89302	0.88366	0.80027	0.90725	0.89302
	EWGAN	0.89945	0.88945	0.81456	0.9147	0.90345
	PCWGAN-GP	0.90366	0.89525	0.81963	0.91845	0.90734
	GMM_WGAN	**0.91163**	**0.90581**	**0.83319**	**0.92156**	**0.91163**
FCM	None	0.90654	0.89774	0.82414	0.9176	0.90654
	SMOTE	0.90654	0.89774	0.82414	0.9176	0.90654
	ADASYN	0.90698	0.89819	0.8249	0.91792	0.90698
	GAN	0.87907	0.86207	0.77627	0.89303	0.87907
	WGAN	0.89767	0.88859	0.80844	0.91076	0.89767
	EWGAN	0.91456	0.90571	0.83257	0.93642	0.91458
	PCWGAN-GP	0.93118	0.92691	0.85329	0.94657	0.9375
	GMM_WGAN	**0.96279**	**0.96208**	**0.92777**	**0.96377**	**0.96279**
	None	0.87383	0.85454	0.76716	0.89077	0.87383
	SMOTE	0.86449	0.84478	0.75105	0.88656	0.86449

Table. S2 (continued)

Algorithm	Model	ACC	F1	JI	Precision	Recall
BIRCH	ADASYN	0.86977	0.84809	0.76002	0.89025	0.86977
	GAN	0.84651	0.82621	0.7207	0.87419	0.84651
	WGAN	0.86512	0.84099	0.75226	0.88274	0.86512
	EWGAN	0.86834	0.84357	0.75845	0.88632	0.86457
	PCWGAN-GP	0.86956	0.84123	0.75023	0.88945	0.86835
	GMM_WGAN	**0.87442**	**0.8551**	**0.76799**	**0.89358**	**0.87442**
DBSCAN	None	0.80841	0.77835	0.67708	0.81387	0.80841
	SMOTE	0.78972	0.75346	0.64919	0.78497	0.78972
	ADASYN	0.8093	0.77935	0.67846	0.81462	0.8093
	GAN	0.77674	0.73283	0.62849	0.76322	0.77674
	WGAN	0.73023	0.64734	0.54412	0.7223	0.73023
	EWGAN	0.75346	0.66924	0.5627	0.7472	0.75139
	PCWGAN-GP	0.76328	0.68245	0.58231	0.76248	0.7824
	GMM_WGAN	**0.8093**	**0.78086**	**0.70272**	**0.85277**	**0.8093**

Table. S3. Wine dataset

Algorithm	Model	ACC	F1	JI	Precision	Recall
DPC	None	0.88136	0.87691	0.78389	0.90254	0.88136
	SMOTE	0.96045	0.96013	0.9236	0.96315	0.96045
	ADASYN	0.96629	0.96632	0.93505	0.96745	0.96629
	GAN	0.8427	0.83408	0.72271	0.88001	0.8427
	WGAN	0.86517	0.8586	0.76344	0.89585	0.86517
	EWGAN	0.90135	0.91493	0.82451	0.90354	0.88341
	PCWGAN-GP	0.91495	0.93023	0.83492	0.92958	0.90352
	GMM_WGAN	**0.98315**	**0.9831**	**0.96684**	**0.98369**	**0.98315**
k_means	None	0.9548	0.95455	0.91375	0.95895	0.9548
	SMOTE	0.96045	0.96006	0.92336	0.96291	0.96045
	ADASYN	0.95506	0.9548	0.91423	0.95917	0.95506
	GAN	0.88764	0.88327	0.7931	0.90546	0.88764
	WGAN	0.94944	0.94904	0.90366	0.95191	0.94944
	EWGAN	0.95143	0.95342	0.91443	0.95823	0.95421
	PCWGAN-GP	0.95932	0.96025	0.92054	0.96023	0.96023
	GMM_WGAN	**0.96629**	**0.96609**	**0.93468**	**0.96839**	**0.96629**
FCM	None	0.94915	0.94851	0.90243	0.95328	0.94915
	SMOTE	0.9548	0.95428	0.9128	0.958	0.9548
	ADASYN	0.95506	0.95462	0.91358	0.95852	0.95506
	GAN	0.85955	0.85195	0.74587	0.88632	0.85955
	WGAN	0.94382	0.9432	0.89304	0.94673	0.94382
	EWGAN	0.95421	0.95131	0.90341	0.95316	0.95429
	PCWGAN-GP	0.95854	0.95942	0.90831	0.95941	0.95924

Table. S3 (continued)

Algorithm	Model	ACC	F1	JI	Precision	Recall
	GMM_WGAN	**0.96067**	**0.96031**	**0.92389**	**0.96322**	**0.96067**
BIRCH	None	0.9774	0.97742	0.95587	0.97785	0.9774
	SMOTE	0.96045	0.96063	0.92465	0.96262	0.96045
	ADASYN	0.97753	0.9776	0.95669	0.97804	0.97753
	GAN	0.97753	0.97753	0.95657	0.97753	0.97753
	WGAN	0.97191	0.97203	0.94594	0.97339	0.97191
	EWGAN	0.97831	0.97431	0.93851	0.97932	0.97002
	PCWGAN-GP	0.97934	0.97923	0.94025	0.97314	9.97423
	GMM_WGAN	**0.98315**	**0.98318**	**0.96698**	**0.98352**	**0.98315**
DBSCAN	None	0.70621	0.67906	0.54586	0.76971	0.70621
	SMOTE	0.70621	0.69548	0.55479	0.75243	0.70621
	ADASYN	0.69663	0.66638	0.5364	0.76486	0.69663
	GAN	0.70787	0.70725	0.54863	0.75509	0.70787
	WGAN	0.71348	0.70285	0.56245	0.75159	0.71348
	EWGAN	0.71495	0.70534	0.56331	0.76421	0.71954
	PCWGAN-GP	0.71007	0.70834	0.56539	0.76902	0.71494
	GMM_WGAN	**0.72958**	**0.71021**	**0.56823**	**0.83866**	**0.72958**

Table. S4. Seeds dataset

Algorithm	Model	ACC	F1	JI	Precision	Recall
DPC	None	0.90431	0.9031	0.82698	0.90856	0.90431
	SMOTE	0.92344	0.92288	0.85914	0.92457	0.92344
	ADASYN	0.92381	0.92223	0.85783	0.92915	0.92381
	GAN	0.90952	0.9082	0.83421	0.91235	0.90952
	WGAN	0.9	0.90001	0.82045	0.90584	0.9
	EWGAN	0.91345	0.91464	0.84236	0.91356	0.91694
	PCWGAN-GP	0.91405	0.91851	0.84813	0.91742	0.91853
	GMM_WGAN	**0.93333**	**0.93305**	**0.87553**	**0.93302**	**0.93333**
k_means	None	0.88995	0.88996	0.80414	0.89263	0.88995
	SMOTE	0.88995	0.88996	0.80414	0.89263	0.88995
	ADASYN	0.89048	0.89049	0.80489	0.89315	0.89048
	GAN	0.89524	0.89586	0.8131	0.8989	0.89524
	WGAN	0.89048	0.89049	0.80489	0.89315	0.89048
	EWGAN	0.90235	0.90356	0.82464	0.90452	0.90453
	PCWGAN-GP	0.91452	0.90934	0.8253	0.90843	0.91043
	GMM_WGAN	**0.92857**	**0.92861**	**0.86775**	**0.92951**	**0.92857**
FCM	None	0.89952	0.89972	0.81959	0.90193	0.89952
	SMOTE	0.89952	0.89972	0.81959	0.90193	0.89952
	ADASYN	0.9	0.90019	0.82028	0.90238	0.9
	GAN	0.90476	0.9058	0.82917	0.91027	0.90476
	WGAN	0.88571	0.88581	0.79735	0.88899	0.88571

Table. S4 (continued)

Algorithm	Model	ACC	F1	JI	Precision	Recall
	EWGAN	0.90351	0.90352	0.81344	0.89341	0.89344
	PCWGAN-GP	0.90842	0.90824	0.82049	0.90452	0.90342
	GMM_WGAN	**0.92381**	**0.92346**	**0.85926**	**0.92336**	**0.92381**
BIRCH	None	0.88995	0.8869	0.80184	0.89641	0.88995
	SMOTE	0.90431	0.90159	0.82496	0.91428	0.90431
	ADASYN	0.90476	0.90376	0.82732	0.90448	0.90476
	GAN	0.88571	0.88127	0.79368	0.89997	0.88571
	WGAN	0.9	0.89675	0.81693	0.91008	0.9
	EWGAN	0.90352	0.90034	0.81945	0.90942	0.90301
	PCWGAN-GP	0.90934	0.90523	0.82493	0.90923	0.9094
	GMM_WGAN	**0.91429**	**0.91349**	**0.84255**	**0.9151**	**0.91429**
DBSCAN	None	0.34928	0.19829	0.12859	0.66779	0.34928
	SMOTE	0.4067	0.31332	0.19784	0.66655	0.4067
	ADASYN	0.4	0.31044	0.19436	0.65711	0.4
	GAN	0.58095	0.58014	0.42094	0.65554	0.58095
	WGAN	0.36667	0.2484	0.15634	0.58906	0.36667
	EWGAN	0.48392	0.2753	0.18439	0.61894	0.40282
	PCWGAN-GP	0.50283	0.3053	0.20451	0.6239	0.4394
	GMM_WGAN	**0.64286**	**0.63494**	**0.48877**	**0.67192**	**0.64286**

Table. S5. Pathbased dataset

Algorithm	Model	ACC	F1	JI	Precision	Recall
DPC	None	0.72	0.67056	0.53086	0.80189	0.72
	SMOTE	0.72575	0.67929	0.53975	0.80637	0.72575
	ADASYN	0.70667	0.64987	0.51309	0.79665	0.70667
	GAN	0.75333	0.72877	0.58558	0.80388	0.75333
	WGAN	0.69333	0.62746	0.49255	0.79095	0.69333
	EWGAN	0.70341	0.65331	0.51492	0.81344	0.71342
	PCWGAN-GP	0.72492	0.67321	0.53482	0.84239	0.73943
	GMM_WGAN	**0.85667**	**0.85584**	**0.75462**	**0.8917**	**0.85667**
k_means	None	0.73244	0.68897	0.5488	0.80782	0.73244
	SMOTE	0.73244	0.68897	0.5488	0.80782	0.73244
	ADASYN	0.73	0.68591	0.54541	0.80678	0.73
	GAN	0.72333	0.68034	0.53807	0.79332	0.72333
	WGAN	0.72667	0.68525	0.54347	0.78737	0.72667
	EWGAN	0.7284	0.68493	0.54925	0.78349	0.72941
	PCWGAN-GP	0.72945	0.68831	0.54831	0.78942	0.7284
	GMM_WGAN	**0.73333**	**0.691**	**0.55056**	**0.80848**	**0.73333**
	None	0.73579	0.69606	0.5553	0.80042	0.73579
	SMOTE	0.73579	0.69606	0.5553	0.80042	0.73579
	ADASYN	0.73	0.68804	0.54692	0.79733	0.73

Table. S5 (continued)

Algorithm	Model	ACC	F1	JI	Precision	Recall
FCM	GAN	0.58667	0.56564	0.40774	0.58562	0.58667
	WGAN	0.72333	0.68446	0.54157	0.77097	0.72333
	EWGAN	0.74832	0.69298	0.56924	0.79342	0.75492
	PCWGAN-GP	0.75392	0.71982	0.58231	0.81331	0.76942
	GMM_WGAN	**0.78195**	**0.79188**	**0.71133**	**0.85695**	**0.78195**
BIRCH	None	0.70903	0.65337	0.51759	0.79857	0.70903
	SMOTE	0.73244	0.69643	0.55552	0.79258	0.73244
	ADASYN	0.74333	0.7059	0.5658	0.81372	0.74333
	GAN	0.72667	0.68123	0.54102	0.80661	0.72667
	WGAN	0.72	0.6971	0.54425	0.79168	0.72
	EWGAN	0.7432	0.70392	0.5621	0.80214	0.7394
	PCWGAN-GP	0.7549	0.71394	0.57491	0.80984	0.7293
	GMM_WGAN	**0.76**	**0.72952**	**0.59027**	**0.82226**	**0.76**
DBSCAN	None	0.36455	0.23495	0.14989	0.4743	0.36455
	SMOTE	0.36	0.22651	0.14411	0.47411	0.36
	ADASYN	0.34	0.33304	0.2663	0.40767	0.34
	GAN	0.42333	0.33357	0.22172	0.48839	0.42333
	WGAN	0.46667	0.3861	0.26805	0.49139	0.46667
	EWGAN	0.4665	0.3814	0.26354	0.48953	0.46667
	PCWGAN-GP	0.46667	0.3777	0.26891	0.48995	0.46667
	GMM_WGAN	**0.46667**	**0.38717**	**0.26909**	**0.49242**	**0.46667**

Table. S6. R15 dataset

Algorithm	Model	ACC	F1	JI	Precision	Recall
DPC	None	0.99666	0.99666	0.9934	0.99674	0.99666
	SMOTE	0.99666	0.99666	0.9934	0.99674	0.99666
	ADASYN	0.99667	0.99667	0.99341	0.99675	0.99667
	GAN	0.99333	0.99333	0.98691	0.99357	0.99333
	WGAN	0.99167	0.99166	0.98365	0.99191	0.99167
	EWGAN	0.99667	0.99667	0.99341	0.99675	0.99667
	PCWGAN-GP	0.99667	0.99667	0.99341	0.99675	0.99667
	GMM_WGAN	**0.99667**	**0.99667**	**0.99341**	**0.99675**	**0.99667**
k_means	None	0.99667	0.99667	0.99341	0.99675	0.99667
	SMOTE	0.99666	0.99666	0.9934	0.99674	0.99666
	ADASYN	0.99667	0.99667	0.99341	0.99675	0.99667
	GAN	0.99333	0.99331	0.98691	0.99341	0.99333
	WGAN	0.99667	0.99667	0.99341	0.99675	0.99667
	EWGAN	0.99667	0.99664	0.99341	0.99675	0.99662
	PCWGAN-GP	0.99667	0.99667	0.99341	0.99679	0.99667
	GMM_WGAN	**0.99667**	**0.99669**	**0.99349**	**0.99683**	**0.99667**
	None	0.99667	0.99667	0.99341	0.99675	0.99667

Table. S6 (continued)

Algorithm	Model	ACC	F1	JI	Precision	Recall
FCM	SMOTE	0.99666	0.99666	0.9934	0.99674	0.99666
	ADASYN	0.99667	0.99667	0.99341	0.99675	0.99667
	GAN	0.99333	0.99331	0.98691	0.99341	0.99333
	WGAN	0.99667	0.99667	0.99341	0.99675	0.99667
	EWGAN	0.99667	0.99667	0.99341	0.99675	0.99667
	PCWGAN-GP	0.99667	0.99667	0.99341	0.99675	0.99667
	GMM_WGAN	**0.99667**	**0.99669**	**0.99349**	**0.99683**	**0.99667**
BIRCH	None	0.98998	0.99	0.98044	0.99046	0.98998
	SMOTE	0.99165	0.99156	0.98366	0.99201	0.99165
	ADASYN	0.98833	0.98844	0.97771	0.9893	0.98833
	GAN	0.985	0.98499	0.97122	0.98544	0.985
	WGAN	0.98667	0.98663	0.97412	0.9873	0.98667
	EWGAN	0.9923	0.99281	0.98999	0.99381	0.98994
	PCWGAN-GP	0.9932	0.99641	0.98983	0.99183	0.99283
	GMM_WGAN	**0.995**	**0.995**	**0.99016**	**0.99508**	**0.995**
DBSCAN	None	0.90651	0.89845	0.87226	0.90213	0.90651
	SMOTE	0.90484	0.89667	0.86905	0.90042	0.90484
	ADASYN	0.91667	0.90174	0.88171	0.89837	0.91667
	GAN	0.91333	0.90064	0.87611	0.9011	0.91333
	WGAN	0.89	0.88634	0.8529	0.89457	0.89
	EWGAN	0.90328	0.90129	0.86342	0.89993	0.90138
	PCWGAN-GP	0.91384	0.9039	0.86932	0.9012	0.9083
	GMM_WGAN	**0.9241**	**0.9068**	**0.88922**	**0.90453**	**0.9241**