Manifold Distance-Based Over-Sampling Technique for Class Imbalance Learning

Lingkai Yang, Yinan Guo, Jian Cheng
School of Information and Control Engineering, China University of
Mining and Technology, Xuzhou, 221008, Jiangsu, China, +86-15150024213
yanglk@cumt.edu.cn, nanfly@126.com, chengjian@cumt.edu.cn

Abstract

Over-sampling technology for handling the class imbalanced problem generates more minority samples to balance the dataset size of different classes. However, sampling in original data space is ineffective as the data in different classes is overlapped or disjunct. Based on this, a new minority sample is presented in terms of the manifold distance rather than Euclidean distance. The overlapped majority and minority samples apt to distribute in fully disjunct subspaces from the view of manifold learning. Moreover, it can avoid generating samples between the minority data locating far away in manifold space. Experiments on 23 UCI datasets show that the proposed method has the better classification accuracy.

Introduction

The datasets in many real classification problems are imbalanced. Dealing with class imbalanced problem normally employs data-, algorithm- or hybrid-level approaches. SMOTE(Chawla et al. 2002), as a data-level method, created new minority examples along the line between a minority class sample and its neighbors. ADASYN(He et al. 2008) generated more data for the hard to learn minority class examples. However, over-sampling is easy to generate a wrong-labeled new sample due to the overlapped or disjunct data in different classes. Many data-level approaches employed the clustering methods to identify and preserve original data space for class imbalanced problem with small disjunct samples. MWMOTE(Barua et al. 2014) created the new samples within clusters of datapoints. ECO-Ensemble(Lim, Goh, and Tan 2017) combined over-sampling strategies with the clustering methods to generate synthetic samples in each minority clustering. ACOSampling(Yu, Ni, and Zhao 2013) is an under-sampling algorithm to retain important majority samples. Without loss of generality, manifold learning extracts the essential structure of the original dataset by the manifold distance, and maps them to a low-dimensional manifold easy to be classified. Based on this, a new minority sample is generated in terms of the manifold distance rather than Euclidean distance. Especially, the overlapped minority samples may be separable in manifold space. For disjunct minority samples, a new sample is produced around a original minority sample in terms of the neighbors found in manifold space.

The Framework of MDOTE

The key issue of manifold distance-based over-sampling technique(MDOTE) is to generate minority samples along the line of a minority sample and its manifold distance extracted neighbors. An under-sampling strategy (US) is firstly implemented to remove redundant majority samples when all of the k_1 neighbors belong to majority class, with the purpose of building a brief balanced dataset. Based on this, a new minority sample is generated in original space based on a minority sample and one of its neighbors measured by manifold distance. LLE(Roweis and Saul 2000) is employed to extract the neighbors of minority samples. The framework of MDOTE is listed in Algorithm 1.

Algorithm 1 MDOTE

Require: Training set(O, O^y); k_1, k_2 and k_3; $ndim$.

Ensure: The balanced dataset X_b, containing both the samples in U and X_{new}.

1: Removing majority samples whose k_1 neighbors are all belonging to majority class,

$$(U, U^y) = US(O, O^y, k_1) \quad (1)$$

2: Mapping the samples of U to LLE space and extracting neighbors,

$$indices = MNE(U, k_2, ndim, k_3) \quad (2)$$

3: Calculating the totally number of minority samples need to be generated,

$$g = u_l - u_s \quad (3)$$

4: **Do for** $i = 1, 2, ..., g$

5: Choosing one minority sample x_1 from U and another minority sample x_2 from its neighbors in $indices$.

6: Generating new minority samples between x_1 and x_2.

$$x_{new} = (x_2 - x_1) \times \lambda + x_1 \quad (4)$$

7: **End loop**

Here, U is the dataset after under-sampling. MNE represents the neighbor extraction method. $indices$ denotes

Table 1: UCI Datasets

	Dataset		Dataset
1	Abalone_18v9	13	Vehicle_VANvALL
2	CTG_PvN	14	Vehicle_SAABvALL
3	CTG_SvN	15	Vehicle_BUSvALL
4	Statland_4v12	16	Wine_3vALL
5	Libra_123vALL	17	Wine_2vALL
6	Libra_789vALL	18	BreastCancer_MvB
7	Yeast_ME1vNUC	19	Ionosphere_BvG
8	Yeast_ME2vCYT	20	PageBlocks_4v2
9	Yeast_ME2vNUC	21	PageBlocks_5v2
10	Yeast_ME3vCYT	22	Segment_4v123
11	Yeast_ME3vNUC	23	Segment_5v123
12	Ecoli_OMvCP		

archive saving k_3 neighbors for each sample in U used for over-sampling. u_l and u_s are the number of majority and minority samples after under-sampling. k_2 and $ndim$ are the number of neighbors used in LLE and the output dimension respectively.

Experiments

All experiments are carried out on 23 UCI datasets (Table 1) and the proposed method is compared with SMOTE, ADASYN, MWMOTE, ACOSampling by AUC value. The number of neighbors and the output dimension of LLE have a direct impact on manifold learning, therefore, are optimized by a simple grid search through the cross-validation evaluation process. The statistical classification performance of different algorithms at 10 independent running times is listed in Table 2, and the best one for each dataset is labelled by bold. Here, SMO, ADA, MWM and ACO represent SMOTE, ADASYN, MWMOTE and ACOSampling method respectively. 'R+', 'R-' and 'pval' are the results of Wilcoxon paired signed-rank test between MDOTE and other methods. 'R+' means the ranking of MDOTE is better than another algorithm and 'pval' means the p_{value} in hypothesis test. Lower 'pval' indicates that MDOTE has the better classification accuracy. As shown in Table 2, the proposed MDOTE outperforms other baselines for most tasks because the value of 'R+' is larger than 'R-' in all of the cases.

Conclusion

Manifold distance-based imbalance learning method is proposed to solve the class imbalanced problem with the overlapping and small disjunct data. The imbalanced dataset is transformed to a balanced one by generating minority samples around a original minority sample in terms of the neighbors found in manifold space. The experimental results on 23 UCI datasets show that the proposed method has the better classification accuracy. Combing the advanced optimizing techniques with MDOTE to improve the structure extracted by manifold learning is our future work.

Table 2: Comparison of AUC among different methods

	SMO	ADA	MWM	ACO	MDOTE
1	0.6590	0.6115	0.6250	0.6160	**0.6609**
2	0.9589	0.9649	**0.9650**	0.9561	0.9561
3	0.9492	0.9475	0.9313	0.9492	**0.9545**
4	0.9746	**0.9889**	0.9775	0.9579	0.9846
5	0.8758	0.9096	0.9069	0.9012	**0.9122**
6	0.9368	0.9493	0.9421	0.9306	**0.9524**
7	0.9673	0.9531	0.9714	0.9714	**0.9786**
8	0.9498	0.9498	**0.9782**	0.9564	0.9616
9	0.9141	0.9161	0.9373	0.9380	**0.9646**
10	0.9473	0.9371	0.9379	**0.9556**	0.9454
11	0.9349	0.9320	0.9190	**0.9364**	0.9114
12	0.9203	0.8870	0.8659	**0.9268**	0.8993
13	0.9437	0.9429	0.9472	0.9307	**0.9506**
14	0.8424	0.8077	0.8313	**0.8543**	0.8345
15	0.9515	0.9434	0.9410	**0.9564**	0.9508
16	0.9462	0.9538	0.9385	**0.9592**	0.9462
17	0.8898	0.8652	0.9401	0.8898	**0.9460**
18	0.9323	**0.9583**	0.9011	0.9435	0.9414
19	0.8594	0.8631	**0.8705**	0.8205	0.8632
20	0.9667	0.9667	0.9577	**0.9778**	0.9667
21	**0.9830**	0.9630	0.9662	0.9729	0.9662
22	0.9728	0.9684	**0.9731**	0.9692	0.9647
23	**0.9458**	0.9365	0.9437	0.9280	0.9380
R+	170.5	200.5	205.5	171.5	-
R-	105.5	75.5	70.5	104.5	-
pval	0.3051	0.0619	0.0424	0.3155	-

Acknowledgments

This work is jointly supported by National Natural Science Foundation of China under Grant 61573361, National Key Research and Development Program under Grant 2016YFC0801406.

References

Barua, S.; Islam, M. M.; Yao, X.; and Murase, K. 2014. Mwmote–majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on Knowledge and Data Engineering* 26(2):405–425.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16(1):321–357.

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks*, 1322–1328.

Lim, P.; Goh, C. K.; and Tan, K. C. 2017. Evolutionary cluster-based synthetic oversampling ensemble (eco-ensemble) for imbalance learning. *IEEE transactions on cybernetics* 47(9):2850–2861.

Roweis, S. T., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Yu, H.; Ni, J.; and Zhao, J. 2013. Acosampling: An ant colony optimization-based undersampling method for classifying imbalanced dna microarray data. *Neurocomputing* 101(2):309–318.