A Distance-based Over-sampling Method for Learning from Imbalanced Data Sets

Jorge de la Calleja
Computer Science Department
I.N.A.O.E.
Tonantzintla, Puebla, 72840, Mexico

Olac Fuentes
Computer Science Department
University of Texas at El Paso
El Paso, Texas, 79968, U.S.A.

Abstract

Many real-world domains present the problem of imbalanced data sets, where examples of one classes significantly outnumber examples of other classes. This makes learning difficult, as learning algorithms based on optimizing accuracy over all training examples will tend to classify all examples as belonging to the majority class. We introduce a method to deal with this problem by means of creating a balanced data set, which allows to improve the performance of classifiers. Our method over-samples the minority class, using a randomized weighted distance scheme to generate synthetic examples in the neighborhood of each minority example.

Introduction

The class imbalance problem occurs when there are many more examples of some classes than others. By convention, the class label of the minority examples is positive, and the class label of the majority instances is negative. Generally, classifiers perform poorly on imbalanced data sets because they generalize from sample data and output the simplest hypothesis that best fits the data (Akbani, Kwek, & Japkowicz 2004). However, with imbalanced data sets, the simplest hypothesis is generally the one that classifies almost all examples as negative. Therefore, we will have biased classifiers that obtain high predictive accuracy over the negative class, but poor predictive accuracy over the positive class of interest.

The problem of imbalanced data sets has been addressed using two main approaches (Akbani, Kwek, & Japkowicz 2004). The first approach consists of assigning different costs to the training examples, weighting more heavily those in the minority class (Pazzani *et al.* 1994; Domingos 1999). The second approach is to re-sample the original dataset, either by over-sampling the minority class and/or under-sampling the majority class (Japkowicz 1997; Kubat & Matwin 1997; Chawla *et al.* 2002). The approach we propose in this paper is based on over-sampling the minority class.

In order to deal with imbalanced data sets we developed a method that generates synthetic examples with the purpose of creating a balanced data set that allows to improve the performance of classifiers. Our method creates new examples by over-sampling the minority class. We find the closest examples to the minority class instances using the weighted distance. When we find the closest instances to our query sample, we only consider a data set with positive examples. Then we use Locally Weighted Linear Regression to perform the classification task.

The Method

Our method performs similarly to SMOTE (Chawla *et al.* 2002), i.e. the minority class is over-sampled by taking each minority class sample and adding synthetic examples in the original data. Also, we operate in the "feature space" rather than the "data space". However, instead of selecting a nearest neighbor at random among the k nearest neighbors (as SMOTE does), we average these neighbors to obtain the mean example. In addition, we only consider the positive data set to find the closest instances using the weighted distance.

To create the new synthetic positive examples in our proposed method we do the following: separate positive and negative examples from the original data set. Find the n closest examples to each positive example using the weighted distance. For doing this, we only consider the positive data set. Then, average these n closest instances to obtain the mean example. Take the difference between the minority example and the mean instance. After that, multiply this difference by a random number between 0 and 1, to select a random point. Finally, add the new synthetic positive instance to the original data set. To perform the classification task we use Locally weighted linear regression. Details of this method can be founded in (de la Calleja & Fuentes 2004).

Experimental Results

In order to asses the effectiveness of the proposed method, we test it on ten different data sets from the UCI Machine Learning Repository. Since most of these data sets have more than two classes, we select the class which has the fewest examples to be the minority class, i.e. the positive class, while the other examples were grouped to create the majority (negative) class.

Table 1: The table below shows the performance of our proposed method using different amount for over-sampling.

	100%		200%		400%		1000%	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
balance	.050	.141	.053	.152	.123	.180	**.196**	**.230**
car	.448	**.756**	.600	.747	.798	.686	**.851**	.637
chess	.988	**.993**	.990	.985	**.992**	.981	.987	.975
glass	**.887**	.867	.876	**.876**	.855	.831	.878	.841
ionosphere	.563	.687	.577	.689	.574	.743	**.592**	**.775**
nursery	.802	**.987**	.925	.959	.982	.846	**1.000**	.645
thyroid	**.910**	.861	.886	**.880**	.892	.871	.872	.876
tic-tac-toe	.691	**.996**	.731	.965	.691	.819	**.750**	.692
wine	.821	**.718**	.753	.615	.825	.686	**.827**	.661
yeast	.322	**.384**	.369	.351	.391	.314	**.441**	.288

In all the experiments reported here we used 10-fold cross-validation. Also, we vary the amount for over-sampling from 100% to 1000%. In addition, we use the five closest examples to create the mean example. The results we show later correspond to the average of five runs.

Since accuracy is not a good metric for imbalanced data sets we evaluate our method using three metrics used in information retrieval: *precision*, *recall* and *f-measure* (Han, Wang, & Mao 2005).

In Table 1 we show the performance of our proposed method for different degrees of over-sampling. We can observe that the best results using the recall metric where obtained when data is over-sampled by 1000%. On the other hand, when data is over-sampled by 100%, we obtained six of the best results using the precision metric. In Table 2 we show the F-measure for the data sets.

Conclusions

We have introduced a method for dealing with imbalanced data sets. Our experimental results show that our proposed method helps to improve the classification of the minority class. We can also say that the optimal amount of over-sampling depends on the data set we are analyzing, thus future research needs to be aimed at characterizing the potential benefits of over-sampling methods and developing heuristics to determine, given a dataset, the degrees of over-sampling that are likely to yield best results. Present and future work also includes testing the method in real-world applications, for example, classifying astronomical objects or biological structures, where the imbalanced class problem is very common.

Acknowledgement The first author wants to thank CONACyT for supporting this work under grant 166596.

References

Akbani, R.; Kwek, S.; and Japkowicz, N. 2004. Applying support vector machines to imbalanced datasets. In *Proceedings of ECML*, 39–50.

Chawla, N.; Bowyer, K.; Hall, L.; and Kegelmeyer, P. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence* 16:321–357.

de la Calleja, J., and Fuentes, O. 2004. Machine learning and image analysis for morphological galaxy classification. *Monthly Notices of the Royal Astronomical Society* 349:87–93.

Domingos, P. 1999. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*, 155–164.

Han, H.; Wang, W.; and Mao, B. 2005. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Proceedings of ICIC*, 878–887.

Japkowicz, N. 1997. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning*, 111–117.

Kubat, M., and Matwin, S. 1997. Addressing the curse of imbalanced training sets: One sided selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 179–186.

Pazzani, M.; Merz, C.; Murphy, P.; Ali, K.; Hume, T.; and Brunk, C. 1994. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, 217–225.

Table 2: The table below show the performance of our proposed method using F-measure.

	100%	200%	400%	1000%
balance	.073	.078	.146	**.211**
car	.562	.665	**.737**	.728
chess	**.990**	.987	.986	.981
glass	**.876**	.876	.842	.859
ionosphere	.618	.628	.647	**.671**
nursery	.884	**.941**	.908	.784
thyroid	**.884**	.883	.881	.874
tic-tac-toe	.815	**.831**	.749	.719
wine	**.766**	.677	.749	.734
yeast	.350	**.359**	.348	.348