CS446: Machine Learning, Fall 2017

Lecture :Addressing Imbalance in Training Data using SMOTE

Lecturer: *Scribe: Anirudh Madhusudan*

Imbalanced data refers to an issue that arises in the classification problem, when some classes are not equally represented. An good example to illustrate this would be the fraud-detection case, where non-fraudulent transactions represent the majority class and the fraudulent transactions represent the minority class. The number of fraudulent cases are sparse and consequently due to lack of sufficient training data from this minority class, the learner doesn't train well to detect these fraudulent cases (which are critical in our learner).

An **accuracy paradox** arises as a result of imbalanced dataset because we get a high accuracy measure, but the accuracy is only reflecting the underlying class distribution. i.e for example in the fraud-detection case, if the learner only predicted non-fraudulent case, the accuracy will still be high because majority of the instances are non-fraudulent.

The need to address this imbalance is critical to improve the learner ability to distinguish classes better. This literature review will study methods to create synthetic sample for instances in the minority class. In the last decade Synthetic Minority Over-Sampling Technique (SMOTE) has become popular to address this imbalanced dataset problem. SMOTE selects two or more similar instances and creates synthetic data by interpolating levels within these minority instances. SMOTE also over-samples the minority class and under-samples the majority class to establish some balance in the dataset.

In my literature survey, I plan to do an in-depth analysis on the SMOTE algorithm and check its validity for different kinds of classification problems. Classifier performance before and after using SMOTE shall be studied.

References

N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer (2002) "SMOTE: Synthetic Minority Over-sampling Technique", Volume 16, pages 321-357

Han H., Wang WY., Mao BH. (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang DS., Zhang XP., Huang GB. (eds) Advances in Intelligent Computing. ICIC 2005. Lecture Notes in Computer Science, vol 3644. Springer, Berlin, Heidelberg

Chawla N.V. (2009) Data Mining for Imbalanced Datasets: An Overview. In: Maimon O.,

Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA