2019-1 Deep Learning Homework #2

Dae-Ki Kang

March 24, 2019

(Deadline : April 7)

Train GAN to perform over-sampling in order to solve data imbalance problems. Seventeen datasets (for binary classification) are attached with the homework. Perform the following experiments for each of the datasets.

1. Use 'Conditional GAN'.
   (a) Divide a dataset into 'train' and 'test'. The test set should be balanced and has about 50% instances of minority class. For example, if, in the original dataset, class + has 100 instances and class - has 500 instances (100,500), then divide the set so that the training set has (50,450), and the test set has (50,50).

   (If a dataset is not that heavily imbalanced, perform the experiments in the same way because it is still important for comparison with other imbalanced datasets)

   (b) Train a neural network with the training set and classify the test set. (You can use SVM if you want.) This is **the result without sampling** which is a baseline.

   (c) Perform over-sampling using GAN from both classes of a balanced subset of the training set. For example, if the training set has (50,450), use a (50,50) subset of the training set to GAN to generate new 400 + instances.

   (d) Now, merge the training set (50, 450) and the new instances (400,0). Train a neural network with the merged (450,450). Classify the test set (50,50). (You can use SVM if you want.) This result is **the result with GAN sampling of both classes**.

   (e) Instead of using both classes, perform over-sampling using GAN with the minority class only. For example, in the step 1c, if the training set has (50,450), use only the 50 + instances (without 450 - instances) of the training set to GAN to generate new 400 + instances. Follow the step 1d. This result is **the result with GAN sampling of one class**.

   (f) Instead of GAN, perform over-sampling using SMOTE. Follow the steps 1c and 1d. This result is **the results with SMOTE sampling**.

   (g) Compare and analyze the experimental results.

2. Use 'Wasserstein GAN' for the same experiments as the question 1.

3. Do your own investigation in the published literature to find GAN which is perfect for this data imbalance problem. And perform the same experiment with above.

4. [**Optional Extra Credit**] Find your own way to perform effective over-sampling for this kind of data imbalance problem. Perform the experiments to justify your answer.

Write a detailed report for all the experiments above and send the report to `dkkang@gmail.com`. The report has to be as detailed as possible.