A Regularized Ensemble Framework of Deep Learning for Cancer Detection from Multi-class, Imbalanced Training Data

Xiaohui Yuan, Lijun Xie, Mohamed Abouelenien

PII: S0031-3203(17)30503-4
DOI: 10.1016/j.patcog.2017.12.017
Reference: PR 6402

To appear in: *Pattern Recognition*

Received date: 30 September 2017
Revised date: 14 November 2017
Accepted date: 17 December 2017

Please cite this article as: Xiaohui Yuan, Lijun Xie, Mohamed Abouelenien, A Regularized Ensemble Framework of Deep Learning for Cancer Detection from Multi-class, Imbalanced Training Data, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.12.017

A Regularized Ensemble Framework of Deep Learning for Cancer Detection from Multi-class, Imbalanced Training Data

Xiaohui Yuan[1,2], Lijun Xie[3,*], Mohamed Abouelenien[2]

Abstract

In medical diagnosis, e.g. bowel cancer detection, a large number of examples of normal cases exists with a much smaller number of positive cases. Such data imbalance usually complicates the learning process, especially for the classes with fewer representative examples, and results in miss detection. In this article, we introduce a regularized ensemble framework of deep learning to address the imbalanced, multi-class learning problems. Our method employs regularization that accommodates multi-class data sets and automatically determines the error bound. The regularization penalizes the classifier when it misclassifies examples that were correctly classified in the previous learning phase. Experiments are conducted using capsule endoscopy videos of bowel cancer symptoms and synthetic data sets with moderate to high imbalance ratios. The results demonstrate the superior performance of our method compared to several state-of-the-art algorithms for imbalanced, multi-class classification problems. More importantly, the sensitivity gain of the minority classes is accompanied by the improvement of the overall accuracy for all classes. With regularization, a diverse group of classifiers is created and the maximum accuracy improvement is at 24.7%. The reduction in computational cost is also noticeable and as the volume of training data increase, the gain of efficiency by our method becomes more significant.

Keywords: ensemble, deep learning, imbalanced data, cancer detection

1. Introduction

In many real-world applications, data annotation is expensive and the target of interest occurs much less frequently compared to the normal cases. In bowel cancer diagnosis, capsule endoscopy has been used as a screening method, which provides over 55,000 frames in a video. The majority of the frames in a video depict normal tissues and the cancer symptom is under-represented in the training frames. Such uneven distribution creates a difficult learning process for finding unbiased decision boundaries. In addition, multi-class problems complicate the learning process. Many methods have been developed to handle the learning from imbalanced data, which leverage One vs. One (OvO) or One vs. All (OvA) strategies. These conversions mostly result in degraded performance and elongated training time as the number of class increases.

The existing methods can be classified as data-level approaches or algorithmic-level approaches [1, 2, 3]. In data level approaches, undersampling or oversampling is applied to balance the minority and majority classes. The oversampling process differs in the way how synthetic examples are created. Some techniques randomly create synthetic examples while others create synthetic examples based on density distribution [4] or distance to the decision boundary [5]. Algorithmic-level approaches such as cost-sensitive learning methods assign higher costs to the minority class [6].

Hybrid methods were developed by integrating both sampling and algorithmic approaches to handle the imbalanced data sets such as boosting methods. AdaBoost [7] was originally introduced to sequentially train an ensemble of classifiers to achieve improved accuracy through an error minimization function. With the success of this method, AdaBoost was extended to multi-class classification.

The indirect conversion of AdaBoost transforms the multi-class into multiple binary classifications using the binarization methods such as AdaBoost.M2 [7] and AdaBoost.MH [8]. These methods require extended training time with many training iterations and the improvement of accuracy is limited to a large number of classes. The direct conversion applies the boosting method to the multi-class data sets by changing the loss function. AdaBoost.M1 [7] was introduced to train multi-class data sets in which the error bound is too strict compared to random guessing error for multi-class problems. Stage-wise Additive Modeling using Multi-class Exponential (SAMME) loss function [9] was developed to ease the error bound of AdaBoost.M1 by transforming it to that of random guessing of C classes, i.e., $\frac{C-1}{C}$. Mukherjee and Schapire [10] developed a theoretical approach to identify the optimal requirements of the trained weak classifiers and introduced a general framework for multi-class boosting. Saberian and Vasconcelos [11] introduced two multi-class boosting algorithms using multi-dimensional codewords and predictors based on coordinate and gradient descents. The

[*] Corresponding author

[1] College of Information Engineering, China University of Geosciences, Wuhan, China

[2] Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

[3] Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China