The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

深層学習抽出特徴量から生成した擬似特徴量を用いた不均衡データ多クラス画像分類

Pseudo-feature generation from feature map in deep learning for imbalanced data multi-class image classification

紺野 友彦　藤井 秀明　岩爪 道昭
Tomohiko Konno　Hideaki Fujii　Michiaki Iwazume

国立研究開発法人　情報通信研究機構　知能科学融合研究開発推進センター
AI Science R&D Promotion Center in National Institute of Information and Communications Technology

If some classes of the data have an only small number of samples, the accuracies of the classes become too low. It is well known as an imbalanced data problem. We often encounter imbalanced data in reality. In a sense, all the wild data are imbalanced.

In this paper, we make pseudo-feature from feature map in lower layers of deep neural networks, and we augment the data of minor classes to improve the imbalanced-data problem. We compare our proposed method with existing ones in imbalanced data multi-class image classification problems.

1. はじめに

データの中に少量しかないクラスがあると、そのクラスの識別率が極端に低下してしまう場合がある。そのようなデータは不均衡データと呼ばれており、現実のデータ分析では頻繁に遭遇する。不均衡データ問題が起きる理由は直感的には以下の通りである。例えば、がん検診で負例が 99%、正例が 1%であったとしよう。学習機は全データに対して負例であると言うだけで 99%の精度で予測できてしまい、そのような学習をしてしまう傾向がある。不均衡データ問題の解決方法としては例えば以下のような方法が知られている 。

1. アップサンプリング: マイナークラス（少量しかデータの無いクラス）のデータを重複してサンプリングして他のクラスのデータ量に合わせる方法
2. ダウンサンプリング:　マイナークラスのデータ量に合わせて他のクラスのサンプル数も少なく取る方法
3. コスト関数の重みの変更: クラスのサンプル数に合わせてコスト関数の重みを変更する方法
4. 人工的なマイナークラスデータの生成 (参考 SMOTE(Synthetic Minority Over-sampling Technique)[Chawla 02]）
5. Bagging: bootstrap sampling を行い、アンサンブルで決定する方法。

連絡先: 紺野 友彦 tomohiko@nict.go.jp

実際には特定の方法のみを用いるのではなく、ダウンサンプリング+Bagging など複数の方法を組み合わせて不均衡データ問題に取り組むことが多い。([Wallace 11] など参照)

ダウンサンプリングを行うとマイナークラスの識別率が高くなる。しかしながら、マイナークラスに合わせて他クラスの利用データも減らすことになるため学習データの総量は少なくなってしまう。結果として識別問題全体の予想精度が低下してしまう。

我々は以下に示すように、深層学習の途中層で得られた特徴量を基に人工的な擬似特徴量を生成し、マイナークラスのデータ数を増強する方法を提案する。この方法による擬似特徴量生成は不均衡データ問題のみに限られる手法ではないと期待される。

2. 提案手法

提案手法の概要を図 1 に示した。手法は次の手続きからなる。

Step1: (図 1 左) 全データを用いて Neural Network 1 で学習を行う。

Step2: (図 1 中) 学習済み Neural Network 1 の途中層から訓練データの特徴量分布を取り出す。

Step3: （図 1 中）取り出した特徴量分布をガウス分布と仮定し、それに従う擬似特徴量を生成する。(2.1 で解説)

Step4: （図 1 中）生成された擬似特徴量と実訓練データ（の抽出特徴量）を合わせて Neural Network 2 で学習を行う。

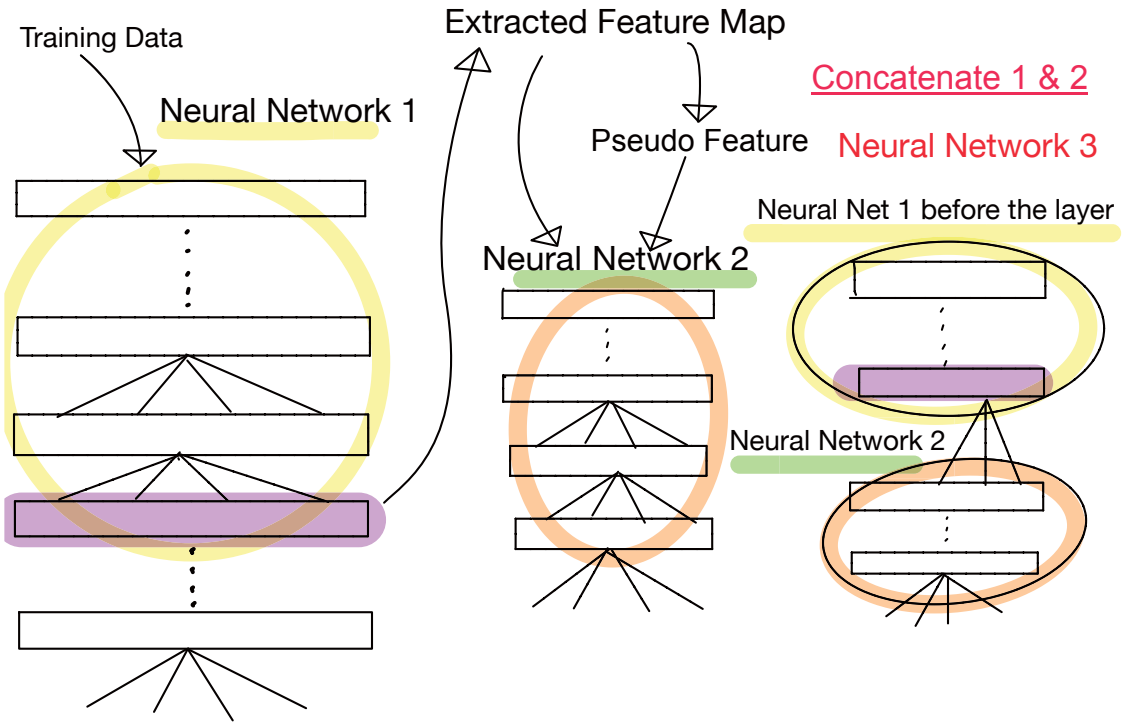

図 1: 提案手法の概要

Step5: (図 1 右) 学習済み Neural Network 1 の上部（特徴量を抽出した層以上の部分）と学習済み Neural Network 2 を結合して Neural Network 3 を作る。

Step6: (図 1 右) Neural Network 3 を用いて推論を行う。

今回は不均衡データの分析であるので、擬似特徴量で増強するのはマイナークラスである。

2.1 擬似特徴量の生成方法

擬似特徴量の作り方を説明する。クラス毎にそれぞれ別の擬似特徴量を生成する。特徴量分布がガウス分布に従うと仮定して 2 つの方法で擬似特徴量を生成した。特徴量分布の次元を D とする。1 つは特徴量が D 次元それぞれで独立なガウス分布に従うと仮定して生成した擬似特徴量分布である。つまり、$\mu_{i,c}, \sigma^2_{i,c}$ をそれぞれクラス c の特徴量分布の i 次元目の平均と分散として、クラス c の擬似特徴量の i 次元目の値 $x_{i,c}$ は以下の確率で次元ごとに独立なガウス分布に従うと仮定して得たものである。

$$\Pr(x_{i,c}) = \mathcal{N}(\mu_{i,c}, \sigma^2_{i,c}) \quad \text{for } i = 1, 2, \cdots, D \quad (1)$$

2 つ目は特徴量が多変量ガウス分布に従うと仮定して得られた擬似特徴量である。つまり $\vec{\mu_c}, \Sigma_c$ をそれぞれクラス c の D 次元特徴量分布の平均ベクトルと分散共分散行列とし、クラス c の擬似特徴量の分布 $\vec{x_c}$ は以下の多変量ガウス分布に従うと仮定して得られたものである。

$$\Pr(\vec{x_c}) = \mathcal{N}_D(\vec{\mu_c}, \Sigma_c) \quad (2)$$

図 2 で我々の提案した擬似特徴量が識別決定境界を押し出す直感的な様子をスケッチする。(図 2 左) 不均衡データの状況ではデータ数が過小なクラスは本来の境界（破線）よりも内側にサンプルが収まってしまう。識別決定境界（黒実線）を押せれば良い。（図 2 右) 我々の擬似特徴量生成は実データが作る境界（赤実線）を超える事が出来る。SMOTE ではこの境界を超えることはできない。この効果によって識別決定境界をより押し出すことが出来る。

3. 実験

結果は表 1 と表 2 にまとめた。現時点で、我々は入手の容易な Cifar10 [Krizhevsky 09] データセットを用いて人工的に不均衡データの状況を作り実験を行なった。Cifar10 は 10 クラスの画像を持ち訓練データ 5 万、テストデータ 1 万から構成される。クラス毎にそれぞれ 5000 枚、1000 枚の画像がある。値は試行の平均値である。

表 1: 全体の識別率 (Accuracy)

#Minor	Baseline	Down-Sampling	Proposed: Multivariate Normal	Proposed: Independent Normal
4	0.711	0.518	0.699	0.697
5	0.662	0.513	0.660	0.655
6	0.608	0.525	0.603	0.600
7	0.585	0.521	0.589	0.583
8	0.558	0.517	0.549	0.556
9	0.511	0.522	0.501	0.532

表 2: 識別率が最低のクラスの識別率 (Min of class-accuracy)

#Minor	Baseline	Down-Sampling	Proposed: Multivariate Normal	Proposed: Independent Normal
4	0.344	0.356	0.479	0.444
5	0.188	0.327	0.351	0.314
6	0.227	0.349	0.323	0.296
7	0.207	0.368	0.291	0.260
8	0.250	0.333	0.298	0.280
9	0.219	0.350	0.268	0.296

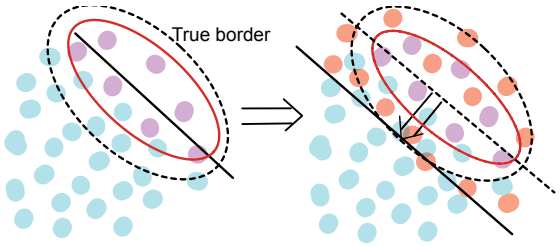

図 2: 直感的なスケッチ　青: メジャークラス　赤: マイナークラス　オレンジ: 擬似的に生成されたデータ

表中の#Minor とはデータの少ないクラスの数である。例えば、#Minor が 4 であるとは学習時には 10 種類の画像クラスのうち 4 クラスが 500 枚しかなく、残り 6 クラスは 5000 枚ある状況である。テスト時には不均衡データではなく 1 万枚のテストデータをそのまま用いた。表中の Baseline とは何もせずに全てのデータを用いて Neural Network 1 を使って学習したものである。Down-Sampling とは Neural Network 1 でダウンサンプリングで学習したものである。Proposed: Multivariate Normal とは特徴量分布から多次元ガウス分布を用いて擬似特徴量を生成した提案手法の結果である。Proposed: Independent Normal とは特徴量分布のそれぞれの次元で独立なガウス分布を用いて擬似特徴量を生成した提案手法の結果である。

3.1　実験結果のまとめ

結果は表 3 にまとめた。マイナークラス数が 4 以上の場合、我々の提案手法は全体の識別率 (Accuracy) は全データを利用した場合 (Baseline) と同じ程度に良く、識別率が最低なクラスの識別率 (Min of class-accuracy) もダウンサンプリングと同じ程度に良い識別率を得ることができた。つまり識別率 (Accucary) も識別率が最低なクラスの識別率 (Min of class accuracy) も共に良い [*1]。

表 3: 結果のまとめ (Summary)

	全体の識別率 (Accuracy)	最低識別率 (Min of Acc.)
Baseline	◯	×
Down-Sampling	×	◯
Proposed	◯	≈ ◯

4.　おわりに

深層学習の途中層から抽出した特徴量の分布から擬似的な特徴量分布を人工的に生成し、それを用いてマイナークラスのデータを増強した。それによって実験では全体の精度も最低精度も共に良い分類を行うことができた。我々が人工的に生成したデータは特徴量分布の境界を超えるものである。

*1　現時点の実験ではマイナークラス数が 3 以下の場合は我々の提案手法はそれほど良い結果を収めなかった。例えばマイナークラスが 1 クラスで残りが 9 クラスがメジャーの場合、これは転移学習のように残り 9 クラスで十分に学習してしまい、マイナーな 1 クラスのサンプル数が少なくてもマイナークラスの識別能力が高くなってしまうからでは無いかと推測している。

参考文献

[Chawla 02] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, Vol. 16, pp. 321–357 (2002)

[Krizhevsky 09] Krizhevsky, A. and Hinton, G.: Learning multiple layers of features from tiny images (2009)

[Wallace 11] Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A.: Class imbalance, redux, in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pp. 754–763IEEE (2011)