Vol.2018-DBS-167 No.24
Vol.2018-IFAT-132 No.24
2018/9/13

オートエンコーダを使用したサンプリング手法による不均衡データの再現度向上

阪本宏輔[†1]　新美礼彦[†2]

概要：近年，機械学習の精度が向上し，様々なデータに対して機械学習が使用されている．しかし，不均衡データに対して機械学習でクラス判別を行う場合，少数クラスの再現度が低くなるという問題がある．このことから，本稿では不均衡データに機械学習を行った際の再現度向上を目的としている．そこで，本稿ではオートエンコーダを使用し，オーバーサンプリングを行う手法の提案を行った．提案手法では，単に同じデータを複製するわけではなく，少数クラスのデータの特徴を学習することで，その特徴を持った新しいデータを作成することができる．これにより，少数クラスのデータが存在する領域を増やし，少数クラスの再現度向上が期待できる．提案手法の評価実験から，決定木，SVM，Deep Learning，それぞれの機械学習手法において，再現度の向上が確認できた．

1. はじめに

近年，機械学習の精度が向上し，様々なデータに対して機械学習が使用されている．しかし，実データにはクラス間で大きくデータ数が異なる不均衡データが数多く存在する[1][2]．不均衡データは，機械学習の性能を下げてしまうため，様々な手法が提案されている[3]．そのため，本稿では不均衡データを機械学習で学習した際に，少数クラスの再現度が低下してしまう問題の解決を目的とする．

ここで，不均衡データにおける機械学習の 1 例を紹介する．表 1 は，人工不均衡データを決定木でクラス判別を行った結果である．表 1 から，少数クラスの再現度は 0.05 と小さい値だが，多数クラスの再現度は 0.99 と高い値になっていることがわかる．このような結果になってしまう原因として考えられる理由は 2 つある．

表 1 不均衡データの精度，再現度

	precision	recall
majority_class	0.90	0.99
minority_class	0.66	0.05

まず 1 つは，少数クラスのデータ数が少ないため，観測データと同じ特徴を持つテストデータが存在しないため，より多くの特徴を持つ多数クラスに分類されてしまうことだ．本稿では，オートエンコーダでオーバーサンプリングを行うことでこの問題の解決を行う．GAN [4]や VAE [5]など，ニューラルネットワークを使用する生成モデルはいくつか提案されている．ニューラルネットワークは本来，生成モデルではないため，これらの手法は学習器の構造や誤差関数を工夫することで，母集団の確率分布に出力を近づけるように学習している．しかし，GAN では学習が Discriminator と Generator の学習速度を調整しなければいけないため，学習が難しく，VAE は正規分布を仮定しているため，ほかの分布には対応できないといった様々な欠点がある．そのため，提案手法ではオートエンコーダを生成モデルではなく，オーバーサンプリングの手法として使用する．オートエンコーダは入力をある確率分布 p(z) から発生するデータ z，教師を確率分布 p(x) から発生する観測データ x とする．学習器は二乗誤差を使用すると z を x に変換するようにパラメータを更新する．つまり，z を x に近似するような学習器を作成する．このことから，学習後のモデルに，確率分布 p(z) から発生したデータを入力すると p(x) に従うデータを生成できる．p(x) に従うデータを生成できるのであれば，観測データにない特徴を持ったデータを生成できる可能性があるため，オートエンコーダでこの問題の解決を行う．

もう 1 つは，各クラスのデータ差により機械学習の学習器が学習できていないことだ．本稿では，アンダーサンプリングを使用してこの問題の解決を行う．アンダーサンプリングにより，各クラスのデータ数を揃えることで学習器の性能が上がることが期待できる．

本稿では以上の理由から，オートエンコーダとアンダーサンプリングを使用した手法の提案を行い，少数クラスの再現度向上を行う．

以下では，2 章で関連研究の紹介，3 章ではオートエンコーダによるデータ生成の確認，4 章で提案手法の実験を行い，評価を行う．最後に 5 章にて本稿をまとめる．

2. 関連研究

本章では，不均衡データに使用されるサンプリング手法やニューラルネットワークを使用した生成モデルの研究を紹介する．

2.1 SMOTE: Synthetic Minority Over-Sampling Technique

SMOTE [6]は不均衡データによく使用されるオーバーサンプリングの手法である．SMOTE は，データとそのデータ付近のデータ間には，データがあると仮定し，データ間に

†1 KOSUKE SAKAMOTO, Future University Hakodate
†2 AYAHIKO NIIMI, Future University Hakodate

Vol.2018-DBS-167 No.24
Vol.2018-IFAT-132 No.24
2018/9/13

データを生成する手法である．そうすることで，データが無い領域にデータを生成することができるため，学習器の性能向上が期待できる．ROC カーブを使用して，性能比較を行った結果，性能の向上が確認された．

SMOTE は，少数クラスの再現度向上が期待できる手法である．なぜなら，データが無い領域にデータを生成することができるからだ．しかし，SMOTE ではデータ間にあるデータしか増やすことができない．本稿では，データ間にあるデータは，データと離れているが同じクラスであるデータより，同じ特徴を持っている可能性が高いため，SMOTE を使用しなかった．オートエンコーダが母集団のデータを生成できると仮定するとデータ間以外のデータも生成できるため，観測データにない特徴を持つデータを生成できる可能性は高いと判断した．

2.2 Generative Adversarial Nets

GAN (Generative Adversarial Nets)[4]は Discriminator と Generator という 2 つのニューラルネットワークを競わせることで学習を行う生成モデルである． Discriminator は，データが Generator により生成されたデータか，オリジナルデータかを判別するように学習する． Generator は，Generator が生成したデータを Discriminator がオリジナルデータと間違えるように学習する． GAN のメリットは母集団の分布を仮定していないことだ．これにより，様々な分布のデータに対して，データを生成することができる．しかし，Discriminator と Generator の学習は調整が難しい．Discriminator の学習が進まないと Generator は学習できず，Discriminator の学習が進みすぎても Generator の学習は進まない．

不均衡データの少数クラスは，データ数が少ないため，分布を仮定しない GAN のようなデータ生成手法が適している．しかし，学習が安定しないことから，提案手法で GAN を使用しなかった．

2.3 Auto-Encoding Variational Bayes

VAE (Auto-Encoding Variational Bayes)[5]は，オートエンコーダに平均と分散のニューロンを置き，潜在変数 z を正規分布から発生させるように工夫することで，誤差関数で分布の近似をする生成モデルである．理論上は分布を近似しているため，観測データにないようなデータの生成を行うことができる．しかし，正規分布を仮定してしまっているため，学習できるデータは正規分布のデータに限られる．

本稿の目的は，不均衡データの再現度向上であるため，生成するデータは少数クラスのデータである．少数クラスのデータは，多数クラスのデータに比べ，データ数が少なく，正規分布に従っている可能性が低いため，提案手法で VAE を使用しなかった．

3. オートエンコーダによるデータ生成

本章では，オートエンコーダによるデータ生成が可能であるか検証を行う．

3.1 仮定

オートエンコーダは，教師データに観測データを使用するニューラルネットワークである．次元削減では，エンコーダの層のみ使用されるため，デコーダの層は使用されない．しかし，学習器で学んでいることは，入力データ z と観測データ x の二乗誤差の最小化である．つまり，学習器は入力データ z を観測データ x に変換できるようにパラメータ θ を更新する．ここで入力データを確率分布 p(z) から発生するデータ z とすると，出力は確率分布 p(x) から発生するデータ x になると考えられる．この過程が正しいならば，オートエンコーダは観測データにない特徴を持ったデータを生成できる可能性があるため，不均衡データの少数クラスの再現度低下問題を解決できる．

3.2 確認実験

本節では，オートエンコーダに関する確認実験を行う．

3.2.1 目的

オートエンコーダが観測データにない特徴を持ったデータを生成できることの確認を行う．

3.2.2 方法

観測データにない特徴を持ったデータを知るためには，観測データにある特徴を持ったデータが分かればよい．そのために，観測データのみで学習した学習器を使用する．この学習器で判別されるデータは観測データにある特徴を持ったデータである．観測データのみで学習した学習器が観測データにある特徴を持ったデータを判別でき，オートエンコーダが観測データにない特徴を持ったデータを生成できるという仮定が正しいならば，観測データにオートエンコーダで生成したデータを加えることで観測データにない特徴を持ったデータも判別できるようになると考えられる．

このことから，観測データのみで学習した学習器と観測データにオートエンコーダで生成したデータを加えたデータで学習した学習器の比較を行う．比較対象は少数クラスのデータの TP の数を比較する．

学習手法によって差が出る可能性を考慮して，決定木，SVM，Deep Learning の学習器を作成し，比較する．また，学習器の性能を最大限に生かすため，アンダーサンプリングを使用して，各クラスのデータ数を揃える．

Vol.2018-DBS-167 No.24
Vol.2018-IFAT-132 No.24
2018/9/13

3.2.3 データセット

データセットは Adult データセットを使用する．このデータセットは，年齢や結婚歴などから年収が 50K を超えているかどうかを判別するデータセットである[7].

データ加工については，欠損値を含むデータ除去とカテゴリーデータのダミー変数化を行った．また，観測データにない特徴を持ったデータを増やすために，観測データの少数クラスのデータを 7.5 割減らしている．加工後のデータ情報を表 2 に示す.

表 2 Adult データセット(加工後)

パラメータ	値
Majority data (training)	18109
Minority data (training)	1505
Majority data (test)	4545
Minority data (test)	1488
Attribute	96

3.2.4 結果

実験の結果を図 1 に示す．学習結果にばらつきがあったため，5 回の結果の平均をとった．また，生成データを混ぜる割合で結果が変わったため，割合を変えて実験を行った．0 割が観測データのみで学習した結果になるため，0 割と他の割合との比較を行う.

すべての手法において，観測データのみで学習した学習器よりも観測データに生成データを加えたデータで学習した学習器の方が TP の数が多くなることが分かった．しかし，生成データの割合を増やしすぎると TP が下がってしまうことも分かった.

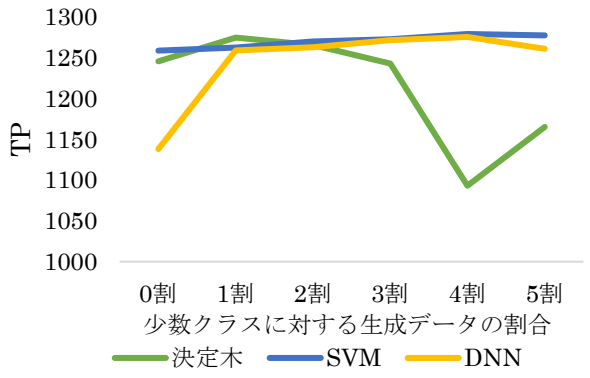

図 1 確認実験の結果

3.3 考察

実験の結果から，観測データのみで学習した学習器よりも観測データに生成データを加えたデータで学習した学習器の方が TP の数が多くなることが分かった．この結果より，生成データによって増えたデータ分の差が影響しているため，生成データは観測データにない特徴を持ったデータを生成できることが確認できた.

また，生成データを増やしすぎると TP の数が下がってしまうことも分かった．本来，オートエンコーダは観測値を教師データにしているため，増やす割合が増えても，TP の数は下がらないはずである．しかし，オートエンコーダには誤差があるため，完全に観測値の特徴をとらえきれない．そのため，生成データを増やしすぎると，誤差が大きくなっていき, TP の数が下がってしまっていることが考えられる.

この実験により，オートエンコーダで観測データにない特徴を持ったデータを生成できることとオートエンコーダによる生成データは誤差を含んでいるため，生成データを増やしすぎないことが重要であることが分かった.

4. 提案手法

本章では，オートエンコーダによるオーバーサンプリングを用いた手法の提案を行う.

4.1 アプローチ

オートエンコーダによる生成データが誤差を含んでいるため，生成できるデータ数に制限がある．また，オートエンコーダが生成するデータは，観測データにある特徴を持ったデータと観測データにない特徴を持ったデータを含んでいる．これらのことから，生成データを観測データにある特徴を持ったデータと観測データにない特徴を持ったデータに分け，観測データにない特徴を持ったデータのみを訓練データに加えることで生成データの数を減らすことができるのではないかと考えた．観測データにある特徴を持ったデータは，観測データで判別することができるため，オーバーサンプリングをする必要がない．そのため，観測データに観測データにない特徴を持ったデータを加えることで，少数クラスの再現度を向上することができる.

提案手法は 2 つの学習器を使用して, データ生成を行う．1 つはデータ生成を行う Generator である．Generator は，オートエンコーダを用いて少数クラスのデータを学習し，データを生成する．もう一つは観測データにある特徴を持ったデータと観測データにない特徴を持ったデータを判別する Discriminator である．Discriminator は，観測データのみで学習した学習器で，入力されたデータが多数クラスか，少数クラスかを判別する．生成データは少数クラスのデータを教師データにしているため，少数クラスの特徴を持ったデータである．そのため，観測データのみで学習した学習器に多数クラスと判別されたなら，それは観測データの少数クラスが持っていなかった特徴を持っていると考えられる．このことから，Discriminator に，生成データを入力

Vol.2018-DBS-167 No.24
Vol.2018-IFAT-132 No.24
2018/9/13

し, 多数クラスなら観測データにない特徴を持ったデータ, 少数クラスなら観測データにある特徴を持ったデータだと判断することができる．このように生成した観測データにない特徴を持ったデータと観測データの少数クラスのデータを合わせたデータをサンプリングした少数クラスとし, サンプリングした少数クラスのデータと同じデータ数になるように観測データの多数クラスをアンダーサンプリングすることでデータをサンプリングする．また, Discriminator は，サンプリングしたデータで学習する手法と同じ手法を使用する．そうすることで，学習手法によって学習できる特徴が違う場合に対応することができる．

この提案手法により, 各クラスのデータ数に差ができず, 観測データにないデータを訓練データにすることができるため，少数クラスの再現度が向上する．

4.2 アルゴリズム

Algorithm 1 により提案手法のアルゴリズムを示す．

Algorithm 1: Over sampling algorithm using Auto encoder	
1	train generator (input: noise, teacher: minority data)
2	generate data with generator (input: noise)
3	train discriminator (input: observed data, teacher: minority or majority)
4	select data having new feature with discriminator (input: generated data)
5	sampling data = observed data + data having new feature
6	reduce majority data with under sampling (sampling data)

Algorithm 1 に沿って，提案手法のアルゴリズムを説明する． 1 行目は，Generator の訓練を行う．入力はある分布に基づき，乱数生成で生成したノイズ，教師は観測値の少数クラスのデータを使用する．また，学習手法はオートエンコーダを使用する．2 行目は，Generator でデータを生成する．入力は訓練に使用した分布から乱数生成したノイズを使用する．3 行目は，Discriminator の訓練を行う．入力は観測データ，教師はデータが少数クラスか多数クラスかを使用する．また，学習手法はサンプリングしたデータで学習を行う手法と同じ手法を使用する．4 行目は，Discriminator で観測データにない特徴を持ったデータを判別する．入力は，Generator で生成したデータを使用する．Discriminator で多数クラスと判別されたデータを観測データにない特徴を持ったデータとする．5 行目は，観測データに観測データにない特徴を持ったデータを加える．6 行目は，5 行目で作成したデータをアンダーサンプリングすることで多数クラスと少数クラスのデータ数を揃える．

提案手法では以上のようにして，サンプリングを行う．

4.3 評価実験

本節では，提案手法の評価実験を行う．

4.3.1 目的

観測されたデータのみで学習した学習器より，提案手法を使用して生成したデータで学習した学習器の方が再現度向上しているかの確認を行う．

4.3.2 方法

観測データのみで学習した学習器と提案手法を使用して生成したデータで学習した学習器の少数クラスの TP を比較する．

学習手法によって差が出る可能性を考慮して，決定木, SVM，Deep Learning の学習器を作成し，比較する．また，観測したデータのみで学習した学習器には，学習器の性能を最大限に生かすため, アンダーサンプリングを使用して, 各クラスのデータ数を揃える．再現度ではなく，少数クラスの TP を比較する理由は，観測データのみで判別できる少数クラスのデータが多いため，少数クラスの再現度の変化が分かりにくいからである．

4.3.3 データセット

3 章の 3.2 節 3.2.3 項と同じデータセットを使用した．

4.3.4 結果

実験の結果を図 2 に示す．学習結果にばらつきがあったため，5 回の結果の平均をとった．観測が観測データのみで学習した学習器，提案が提案手法で生成したデータで学習した学習器である．

すべての手法において，観測データのみで学習した学習器よりも提案手法で生成したデータで学習した学習器の方が TP の数が多くなることが分かった．

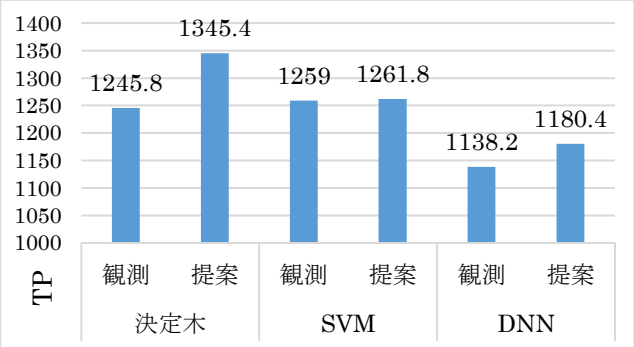

図 2 比較実験の結果

4.4 評価

実験の結果から，観測データのみで学習した学習器よりも提案手法で生成したデータで学習した学習器の方が TP の数が多くなることが分かった．この結果より，生成した

データにより，観測データだけでは判別できなかった少数クラスのデータを判別できていると確認できた．このことから，Discriminator により観測データにない特徴を持ったデータを判別できていることが分かる．

また，決定木に提案手法を試した結果は 1345.4 と非常に高い値を示している．この値は 3 章の 3.2 節 3.2.4 項の結果である図 1 のどの値よりも高いため，決定木は Discriminator により観測データにない特徴を持ったデータの判別がしやすい手法である．しかし，SVM や Deep Learning は図 1 と比べ，低い数値になっているため，Discriminator により観測データにない特徴を持ったデータの判別がしにくい手法であることが分かった．

これらの結果から，提案手法は観測データにない特徴を持ったデータを観測データに加えることで，少数クラスの再現度を向上することができることが分かった．また，決定木との相性が良く，提案手法と組み合わせることで，SVM や Deep Learning より少数クラスの再現度を向上することができる．

5. おわりに

本稿では，不均衡データの少数クラスの再現度低下問題に取り組んだ．この問題の原因は，少数クラスのデータ数が少ないため，観測データと同じ特徴を持つテストデータが存在しないことと各クラスのデータ差により機械学習の学習器が学習できていないことである．そのため，観測データにない特徴を持ったデータと観測データを訓練データに加え，アンダーサンプリングすることで，この問題を解決した．しかし，オートエンコーダで作成したデータには誤差があるため，多くのデータを生成してしまうと学習に悪影響を与えてしまうことがわかった．この問題を解決するために，オートエンコーダで生成したデータを観測データにない特徴を持ったデータのみにする手法を提案した．提案手法は，決定木，SVM，Deep Learning のすべての手法で，アンダーサンプリングよりも高い少数クラスの TP の数を示した．これらのことから，提案手法は不均衡データに対する機械学習に効果的であることがわかる．

今後は，提案手法の性能向上を目的とした研究を行う．提案手法は，決定木と組み合わせると非常に高い性能を示すが，SVM や Deep Learning では Discriminator がうまく機能しなかった．この問題を解決することで，様々な手法に応用できる手法になるだろう．また，提案手法はデータが無い領域にデータを生成する手法に Generator を置き換えることができる．既存の生成モデルや SMOTE などに手法を置き換えることで，性能の向上が行える可能性がある．

参考文献

[1] Chawla, N. V. et al. : Editorial: Special Issue on Learning from Imbalanced Data Sets, SIGKDD Explorations 6 (2004).
[2] Niimi, A. : Majority Rule Approach to Deep Learning for Real Credit Card Transaction Data, World Congress on Internet Security, pp.35-39 (2017).
[3] Wallace, B. C. et al. : Class imbalanced, Redux, In 11th IEEE International Conference on DataMining, pp.754-763 (2011).
[4] Goodfellow, I. J. et al. : Generative Adversarial Nets, In NIPS, pp. 2672–2680 (2014).
[5] Kingma, D. P. and Welling, M. : Auto-Encoding Variational Bayes, In The 2nd International Conference on Learning Representations (2013).
[6] Chawla, N. V. et al. : SMOTE: Synthetic Minority Over-Sampling Technique, Journal of Artificial Intelligence Research 16, pp.321-357 (2002).
[7] “UCI Machine Learning Repository Adult Data Set”. https://archive.ics.uci.edu/ml/datasets/adult, (参照 2018-08-08).