1K-02

誤ラベルを含むデータ集合を用いた統計的機械学習

鈴木　美香 †　　安田　宗樹 ‡
† 山形大学大学院理工学研究科　　‡ 山形大学大学院理工学研究科

1　はじめに

深層学習 [1, 2] の発展によって，パターン認識システムの高い有用性が示されてきている．本稿では，条件付確率で記述される確率的パターン認識システムに注目する．確率的パターン認識システムは機械学習の方法（最尤法）によって最適化される．通常の機械学習の方法は，教師データ（ラベルデータ）が正しいことを前提としているため，誤ったラベルデータが混入する可能性を考慮していない．本稿では，ラベルデータに誤りが含まれている場合の，確率的パターン認識システムに対する統計的機械学習の方法を提案する．

まず，2 節で，ラベルデータの正誤情報（正誤データ）が既知の場合の学習方法（提案法 1）[5] について説明し，次に，3 節で，正誤データが未知の場合の学習方法（提案法 2）[5] について説明する．実用上は提案法 2 が重要である．提案法 1 に類似の手法が別の観点から提案されている [4].

しかし，提案法 2 はラベルデータの誤り率 ρ をハイパーパラメータとしてもっており，その値を人手で設定する必要があるという欠点をもつ．4 節で，ハイパーパラメータ ρ の推定法を提案し，5 節の数値実験により，提案法の有効性を示す．

2　提案法 1 正誤データを用いる最尤法

本節ではラベルデータの誤り部分が既知の場合の最尤法について考える．ラベルデータに誤りを含む場合の最尤法は誤りデータを捨て，正解データのみで学習する．しかし，捨てるデータも情報を持っているので非効率である．学習データを $\mathcal{D} = \{(\boldsymbol{x}^{(\mu)}, \boldsymbol{y}^{(\mu)}) | \mu = 1, 2, \ldots, N\}$ と定義する．$(\boldsymbol{x}^{(\mu)}, \boldsymbol{y}^{(\mu)})$ は μ 番目のデータである．データは入力データ $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ とその入力データに対する出力データ $\boldsymbol{y} = (y_1, y_2, \ldots, y_Y)^T$ の組の構成からなる．$y_k \in \{0, 1\}$,$\sum_{k=1}^{Y} y_k = 1$ である．この出力はラベルデータと呼ばれ 1-of-K 表現のベクトルである．データ $\mathcal{D}$ を用いた最尤法によりパラメータ $\boldsymbol{\theta}$ の最適な値を得ることを考える．尤度関数 $L(\boldsymbol{\theta}) = \prod_{\mu=1}^{N} p(\boldsymbol{y}^{(\mu)} \mid \boldsymbol{x}^{(\mu)}, \boldsymbol{\theta})$ とする．

正誤データが既知の場合の学習なので，ラベルデータの正誤情報を示す正誤データ $l^{(\mu)}$ を追加する．$l^{(\mu)} \in \{0, 1\}$ は出力データ $y^{(\mu)}$ が正解なら 0，間違いなら 1 となるデータとする．尤度関数 $L(\boldsymbol{\theta})$ を拡張し，次の新しい尤度関数を定義する．

$$L_0(\boldsymbol{\theta}) = \prod_{\mu=1}^{N} p_0(\boldsymbol{y}^{(\mu)} \mid \boldsymbol{x}^{(\mu)}, l^{(\mu)}, \boldsymbol{\theta}) \tag{1}$$

出力 $\boldsymbol{y}^{(\mu)}$ が正解データなら $l^{(\mu)} = 0$ となり $p(\boldsymbol{y}^{(\mu)} \mid x^{(\mu)}, \boldsymbol{\theta})$ の確率を高くする．出力 $\boldsymbol{y}^{(\mu)}$ が間違いデータなら $l^{(\mu)} = 1$ となり $\boldsymbol{y}^{(\mu)}$ 以外のデータが正解なので，他のデータの確率 $(1 - p(\boldsymbol{y}^{(\mu)} \mid \boldsymbol{x}^{(\mu)}, \boldsymbol{\theta}))$ を高くする．このとき条件付確率は次のように得られる

$$p_0(\boldsymbol{y} \mid \boldsymbol{x}, l, \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})^{1-l} \left\{ \frac{1 - p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})}{Y - 1} \right\}^{l} \tag{2}$$

Statistical machine learning using data including error
†Mika Suzuki　‡Yasuda Muneki
†Graduate School of Science and Engineering ,Yamagata University
‡Graduate School of Science and Engineering ,Yamagata University

$Y - 1$ は規格化条件のための定数である．式 (1) の尤度関数を最大化するパラメータ $\boldsymbol{\theta}$ を求める．尤度関数を式 (2) のように拡張することで，最尤法を使う際に捨てていた誤りデータも有効に利用できる最尤法の枠組みが実現する．

3　提案法 2 正誤データを用いない最尤法

本節では，ラベルデータのどこに誤ラベルを含むか分からないようなより一般的な場合について考える．通常，誤ラベルがどこにあるか不明な場合が多い．2 節ではデータ構成は入力 $\boldsymbol{x}^{(\mu)}$, 出力 $\boldsymbol{y}^{(\mu)}$, 正誤データ $l^{(\mu)}$ の 3 つで 1 セットとなっているが, データの正誤情報を知らない状態での学習は正誤データ $l^{(\mu)}$ の情報がないことである．そこで出力データは確率 ρ で誤りになると仮定し，正誤データ l に対する次のベルヌイ分布 $B(l \mid \rho) = \rho^l (1 - \rho)^{1-l}$ を定義する．尤度関数 $L(\boldsymbol{\theta})$ を拡張し，次の新しい尤度関数を定義する．

$$L_1(\boldsymbol{\theta}) = \prod_{\mu=1}^{N} p_1(\boldsymbol{y}^{(\mu)} \mid \boldsymbol{x}^{(\mu)}, \boldsymbol{\theta}, \rho) \tag{3}$$

式 (3) は正誤データを含まないので，データの正誤情報を必要としない．式（2）の条件付確率 $p_0(\boldsymbol{y} \mid \boldsymbol{x}, l, \boldsymbol{\theta})$ から l を消すために式 (2) と $B(l)$ をかけて l に関して周辺化すると新しい条件付確率 $p_1(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ が次のように得られる．

$$p_1(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}, \rho) = \frac{Y(1 - \rho) - 1}{Y - 1} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) + \frac{\rho}{Y - 1} \tag{4}$$

式 (3) の尤度関数を最大化するパラメータ $\boldsymbol{\theta}$ を求める．

4　提案法 2 ハイパーパラメータ $\boldsymbol{\rho}$ の推定

本節では 3 節提案法 2 の誤り率 ρ であるハイパーパラメータ推定について考える．提案法 2 はデータの誤り率 ρ がハイパーパラメータであり，人手で設定する必要がある．設定値とデータの誤り率 ρ が一致しない場合は学習の精度が大幅に下がるためデータの最適な誤り率 ρ を推定する必要がある．

そこで一部の正誤データが判明しているとする．N 個のデータ中 M 個の正誤データが判明している場合を考える．正誤データが判明しているデータを M 個のデータからなる $\mathcal{D}_M = \{(\boldsymbol{x}^{(\mu)}, \boldsymbol{y}^{(\mu)}, \boldsymbol{l}^{(\mu)}) \mid \mu = 1, 2, \ldots, M\}$ とする．正誤データが判明してないデータを $N - M$ 個のデータからなる $\mathcal{D}_R = \{(\boldsymbol{x}^{(\mu)}, \boldsymbol{y}^{(\mu)}, \boldsymbol{l}^{(\mu)}) \mid \mu = M+1, M+2, \ldots, N\}$ とする．$\mathcal{D}_M$ はすべてのデータを信用できるので，正誤データあり学習とし考えることができる．$\mathcal{D}_R$ はラベルデータを信用できないので正誤データなし学習とし，提案法 2 を用いて学習することができる．次の新しい尤度関数を定義する．

$$L_2(\boldsymbol{\theta}, \rho) = \prod_{\mu=1}^{M} p_2(\boldsymbol{y}^{(\mu)}, l^{(\mu)} \mid \boldsymbol{x}^{(\mu)}, \boldsymbol{\theta}) + \prod_{\mu=M+1}^{N} p_1(\boldsymbol{y}^{(\mu)} \mid \boldsymbol{x}^{(\mu)}, \boldsymbol{\theta}, \rho) \tag{5}$$

$\mathcal{D}_M$ は尤度関数の第 1 項に対応し，$\mathcal{D}_R$ 第 2 項に対応する．$p_1(y \mid x, \theta, \rho) = \sum_{l=0,1} p_2(\boldsymbol{y}, l \mid \boldsymbol{x}, \boldsymbol{\theta}, \rho)$ より新しい条件付確率は次のよ

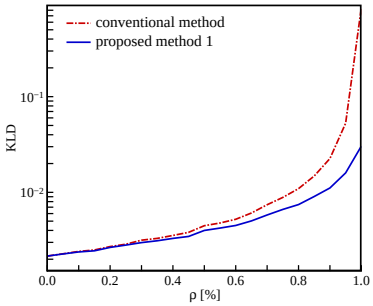

図 1: 提案法 1 の性能比較

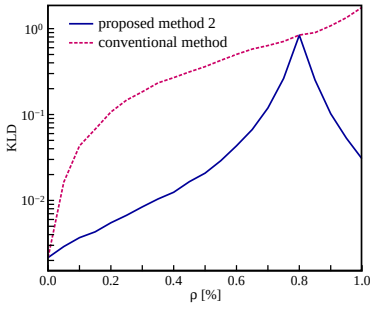

図 2: 提案法 2 の性能比較

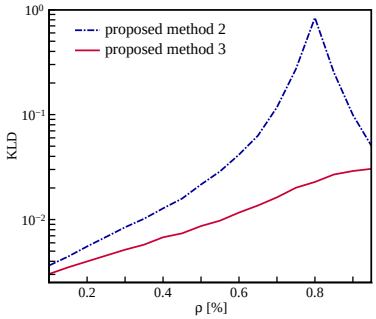

図 3: 提案法 3 の性能比較

うに得られる．

$$p_2(\boldsymbol{y}, l \mid \boldsymbol{x}, \boldsymbol{\theta}, \rho) = p_0(\boldsymbol{y} \mid \boldsymbol{x}, l, \boldsymbol{\theta}) B(l) \tag{6}$$

式 (5) の尤度関数を最大化するパラメータ $\boldsymbol{\theta}, \rho$ を求める．

5 数値実験

本節では人工データに対する数値実験を用いて，2〜4 節で提案した提案法 1〜3 の性能を調べる．数値実験では確率モデルとして多値ロジスティック回帰モデル (multi-valued logistic regression; MLR)[6] を用いて人工的に最適なデータ確率 $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^*)$ を設定する．$\boldsymbol{\theta}^*$ は $\mathcal{N}(\boldsymbol{\theta}^* \mid 0, 1)$ から生成する．次にデータ $\mathcal{D}$ を生成して提案法の学習を行う．データ $\mathcal{D}$ の生成過程は以下の通りである．最適なデータ確率 $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^*)$ に従ってランダムに決められた入力 $\boldsymbol{x}$ に対応した出力 $\boldsymbol{y}$ を生成する．この方法でまず N 個のデータを生成する．$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}^*)$ から生成された出力はすべて正解データであるので，確率 ρ（誤り率）でいくつかの出力を選択し，選ばれた出力の値をランダムに変更する．変更された出力の正誤ラベルは $l^{(\mu)} = 1$ に変更される．提案法 3 ではさらに正誤が判明しているデータ $\mathcal{D}_M$ と判明していないデータ $\mathcal{D}_R$ にラベルを付け, 区別する．プロットは 100 回の試行の平均である．データ数 $N = 10000$, 入力データ $n = 10$, 出力データ $Y = 5$ で数値実験を行う．

図 1 に 2 節提案法 1 の誤ラベルデータを考慮する学習と従来法の誤ラベルデータを捨てる学習との比較を示す．図 2 に 3 節提案法 2 のデータ集合に誤ラベルデータを含むことを前提とする学習と従来法のすべてのデータを正解とする学習の比較を示す．図 3 に 4 節提案法 3 の一部の正誤が判明しているラベルデータによるハイパーパラメータ ρ を推定した学習と 3 節提案法 2 との比較を示す．それぞれの横軸は誤り率 ρ，縦軸はカルバック・ライブラー情報量 (Kullback-Leibler divergence;KLD) である．KLD は人工的に設定した最適なデータ確率と最尤法により求まった学習モデルとの差であり，小さいほど精度が良い．

図 1 では提案法 1 が従来法より高精度であることが分かる．従来法は誤りデータを捨てて学習するため学習できる情報が少なく精度が悪い．提案法 1 では捨てるデータも情報を持っているので学習に取り入れたことで精度が向上した．

図 2 では提案法 2 が従来法に対して全体的に高精度であることが分かる．誤り率 $\rho = 0.8$ の特異点付近で精度が下がる原因は誤り率 ρ が出力の数 Y と $\rho = 1 - \frac{1}{Y}$ の関係になる時である．この場合はすべての出力データの確率が同等でり，ρ を式 (3) の尤度関数に代入するとパラメータの項が 0 になるため学習することができない．通常の最尤法では誤り部分を知らないということはすべてのデータを信用して正解データとして学習する方法しかない．提案法 2 ではすべてのデータが何らかの確率で間違っているとする学習であり，正誤データが未知の場合は高性能である．

図 3 では誤り率を推定する学習である提案法 3 の方が精度が高いことが分かる．提案法 3 はハイパーパラメータを推定し，各データの誤り率 ρ に最適化させるため，提案法 2 のような特定誤り率に影響されず，データの誤り率 ρ に関係なく全体的に高性能である．

6 まとめ

本稿では誤ラベルを含むデータ集合をデータの正誤情報により二つの提案手法を用いた新しい統計的機械学習を提案した．またより実用性を考え，3 節の正誤データなし学習を拡張し，誤ラベルの誤り率 ρ を人手で設定するハイパーパラメータの推定法を導出したことで，従来法の性能の向上を実現することができた．正誤データがあれば 2 節の提案法 1，なければ 3 節の提案法 2 を使い分けることで高精度の学習を得られる．また一部の正誤データが判明している信用できるデータがあれば 4 節の提案法 3 よりデータの誤り率のハイパーパラメータを推定するすることで，提案法 2 よりも更に高精度の学習が得られる．提案法の枠組みはすべての確率的識別モデルに対して適用可能である．今後は MNIST などの画像データを用いて画像認識への適用を進める．

謝辞

本研究の一部は，JSPS 科研費（15K00330，15H03699）及び，JST CREST（JPMJCR1402）の補助を受けて行われたものである．

参考文献

[1] 上嶌敏弘（編）：深層学習，近代科学社，2015.
[2] 岡谷貴之: 機械学習プロフェッショナルシリーズ　深層学習，講談社，2015.
[3] 杉山将: 統計的機械学習 生成モデルに基づくパターン認識，オーム社，2011.
[4] T. Ishida, G, Niu, W. Hu and M. Sugiyama: Learning from Complementary Labels, In proceedings in NIPS2017, 2017.
[5] 小堀美香，安田宗樹: 誤り教師データを含むデータセットを用いた統計的機械学習に関する研究，情報処理学会第 79 回全国大会，2017.
[6] C. M. Bishop: Pattern Recognition and Machine Learning, Springer, 2006.