電子情報通信学会誌, *vol.97, no.5, pp.353–358, 2014.*

密度比推定によるビッグデータ解析

Big Data Analysis by Density Ratio Estimation

杉山　将
東京工業大学　計算工学専攻
sugi@cs.titech.ac.jp
http://sugiyama-www.cs.titech.ac.jp/˜sugi

概 要

統計的データ解析のあらゆるタスクは，データの背後に潜む確率分布を推定することにより解決できる．しかし，確率分布の推定は一般に困難であることが知られているため，これを回避しつつ所望のデータ解析を行うことが望ましい．本稿では，確率分布でなく確率密度の比を通して様々なデータ解析を行う密度比推定の枠組みを紹介する．この枠組には，非定常環境適応学習，異常値検出，二標本検定，独立性検定，特徴選択，独立成分分析，条件付き確率密度推定，確率的パターン認識など多くの重要なデータ解析パラダイムが含まれる．

キーワード
機械学習，確率分布，密度比，変化検知

1　はじめに

ビッグデータ時代には，データサイエンティストは多種多様なデータに対して，様々なデータ解析を迅速に行う必要がある．その際，データ解析タスクごとにアルゴリズムの開発・実装を行うのは効率が悪い．このような背景のもと，様々なデータ解析タスクを統一的に解決できる**密度比推定** [1] とよばれるアプローチが近年提案された．

密度比推定の枠組みには，非定常環境適応学習，異常値検出，二標本検定，独立性検定，特徴選択，独立成分分析，条件付き確率密度推定，確率的パターン認識など，多くの重要なデータ解析パラダイムが含まれる．そして，これらの多様なデータ解析タスクを，確率密度比関数の推定を通して統一的かつ精度良く解決する．

本稿では，まず第 2 節で様々な密度比推定の方法を概説し，第 3 節で密度比推定により解決できる様々な機械学習タスクの例を紹介する．そして，第 4 節では，密度比推定に基づく教師なし変化検知手法を詳しく説明し，最後に第 5 節で今後の展望を述べる．

2 密度比推定

確率密度 $p^*_{\mathrm{nu}}(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^{n_{\mathrm{nu}}}$ と，確率密度 $p^*_{\mathrm{de}}(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$ から，密度比

$$r^*(\boldsymbol{x}) = \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{p^*_{\mathrm{de}}(\boldsymbol{x})}$$

を推定する問題を考える．‘nu’ と ‘de’ は，分子 (numerator) と分母 (denominator) の頭文字である．

標本 $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^{n_{\mathrm{nu}}}$ から確率密度 $p^*_{\mathrm{nu}}(\boldsymbol{x})$ を，標本 $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$ から確率密度 $p^*_{\mathrm{de}}(\boldsymbol{x})$ をそれぞれ推定すれば，密度比 $r^*(\boldsymbol{x}) = p^*_{\mathrm{nu}}(\boldsymbol{x})/p^*_{\mathrm{de}}(\boldsymbol{x})$ を推定できる．しかし，推定した確率密度の比を取ることにより，確率密度の推定誤差が増幅されてしまう恐れがあるため，このような二段階の方式は推定が精度が良くない（図 1）．そこで本節では，各確率密度 $p^*_{\mathrm{nu}}(\boldsymbol{x})$ や $p^*_{\mathrm{de}}(\boldsymbol{x})$ の推定を経由せずに，密度比 $r^*(\boldsymbol{x}) = p^*_{\mathrm{nu}}(\boldsymbol{x})/p^*_{\mathrm{de}}(\boldsymbol{x})$ を直接推定できる手法を紹介する [1].

2.1 確率的分類法

確率的分類法では，$p^*_{\mathrm{nu}}(\boldsymbol{x})$ と $p^*_{\mathrm{de}}(\boldsymbol{x})$ から生成された標本に，ラベル $y =$ ‘nu’ と ‘de’ をそれぞれ割り当てる．このとき，$p^*_{\mathrm{nu}}(\boldsymbol{x})$ と $p^*_{\mathrm{de}}(\boldsymbol{x})$ を

$$p^*_{\mathrm{nu}}(\boldsymbol{x}) = p^*(\boldsymbol{x}|y = \text{‘nu’})$$
$$p^*_{\mathrm{de}}(\boldsymbol{x}) = p^*(\boldsymbol{x}|y = \text{‘de’})$$

と表せるため，ベイズの定理により，密度比を

$$r^*(\boldsymbol{x}) = \frac{p^*(y = \text{‘de’})}{p^*(y = \text{‘nu’})} \frac{p^*(y = \text{‘nu’}|\boldsymbol{x})}{p^*(y = \text{‘de’}|\boldsymbol{x})}$$

と表現できる．ここで，ラベルの事前確率 $p^*(y)$ の比を標本数の比で近似し，ラベルの事後確率 $p^*(y|\boldsymbol{x})$ を $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^{n_{\mathrm{nu}}}$ と $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$ に対する確率的分類器 $\widehat{p}(y|\boldsymbol{x})$（例えば，ロジスティック回帰や最小二乗確率的分類により求める）で近似すれば，密度比の近似 $\widehat{r}(\boldsymbol{x})$ を次式で求められる．

$$\widehat{r}(\boldsymbol{x}) = \frac{n_{\mathrm{de}}}{n_{\mathrm{nu}}} \frac{\widehat{p}(y = \text{‘nu’}|\boldsymbol{x})}{\widehat{p}(y = \text{‘de’}|\boldsymbol{x})}$$

2.2 積率適合法

積率適合法では，密度比のモデル $r(\boldsymbol{x})$ を用いて，$r(\boldsymbol{x})p^*_{\mathrm{de}}(\boldsymbol{x})$ の積率を $p^*_{\mathrm{nu}}(\boldsymbol{x})$ の積率に最小二乗適合させる．例えば一次の積率（すなわち期待値）を適合させる場合は，次式を解く．

$$\min_r \left\| \mathbb{E}_{p^*_{\mathrm{de}}}[\boldsymbol{x} r(\boldsymbol{x})] - \mathbb{E}_{p^*_{\mathrm{nu}}}[\boldsymbol{x}] \right\|^2$$

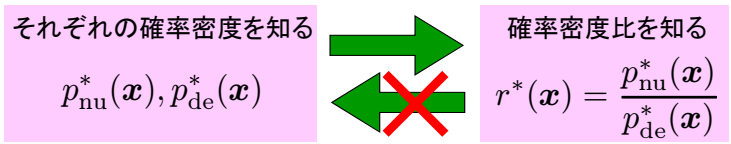


図 1: 密度比推定．分子と分母の密度 $p^*_{\mathrm{nu}}(\boldsymbol{x})$，$p^*_{\mathrm{de}}(\boldsymbol{x})$ がわかればそれらの比 $r^*(\boldsymbol{x})$ もわかるが，密度比 $r^*(\boldsymbol{x})$ がわかったとしてもそれぞれの密度はわからない．従って，分子と分母の密度を個別に推定するよりも，密度比を直接推定する方が易しいと考えられる．

ただし，$\|\cdot\|$ はユークリッドノルム，$\mathbb{E}$ は期待値を表す．真の密度比を正しく求めるためには全ての次数の積率を適合させる必要がある．ガウスカーネルのような普遍再生核関数 $K(\boldsymbol{x}, \boldsymbol{x}')$ を用いれば，これを効率良く実現できる．

$$\min_r \left\| \mathbb{E}_{p^*_{\mathrm{de}}}[K(\boldsymbol{x}, \cdot) r(\boldsymbol{x})] - \mathbb{E}_{p^*_{\mathrm{nu}}}[K(\boldsymbol{x}, \cdot)] \right\|^2_{\mathcal{H}}$$

ただし，$\|\cdot\|_{\mathcal{H}}$ は $K(\boldsymbol{x}, \boldsymbol{x}')$ が属するヒルベルト空間のノルムを表す．実際には，期待値を標本平均で近似した規準を最小化することにより解を求める．

2.3 密度適合法

密度適合法では，一般化カルバック距離のもとで $r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x})$ を $p^*_{\mathrm{nu}}(\boldsymbol{x})$ に適合させる．

$$\min_r \mathbb{E}_{p^*_{\mathrm{nu}}} \left[\log \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{r(\boldsymbol{x}) p^*_{\mathrm{de}}(\boldsymbol{x})} \right] + \mathbb{E}_{p^*_{\mathrm{de}}}[r(\boldsymbol{x})]$$

ただし，実際の推定には期待値を標本平均で近似した規準を用いる．$r(\boldsymbol{x})$ として，線形モデル，対数線形モデル，混合モデルを用いた手法が提案されている．

2.4 密度比適合法

密度比適合法では，密度比モデル $r(\boldsymbol{x})$ を真の密度比 $r^*(\boldsymbol{x})$ に最小二乗適合させる．

$$\min_r \mathbb{E}_{p^*_{\mathrm{de}}} \left[(r(\boldsymbol{x}) - r^*(\boldsymbol{x}))^2 \right]$$

ただし，実際の推定には期待値を標本平均で近似した規準を用いる．$r(\boldsymbol{x})$ として線形モデルを用いれば，密度比適合法の解は解析的に求められる．

2.5 統一的枠組み

上記の最小二乗密度比適合法を一般化し，ブレグマン距離のもとで $r(\boldsymbol{x})$ を $r^*(\boldsymbol{x})$ に適合させる．

$$\min_r \mathbb{E}_{p^*_{\mathrm{de}}}[f(r^*(\boldsymbol{x})) - f(r(\boldsymbol{x})) \\ - f'(r(\boldsymbol{x}))(r^*(\boldsymbol{x}) - r(\boldsymbol{x}))]$$

ただし，$f(t)$ は微分可能な強凸関数であり，$f'(t)$ はその微分を表す．$f(t)$ を変えることにより，様々な密度比推定法が表現できる．

- ロジスティック回帰：$t\log t - (1+t)\log(1+t)$
- 再生核積率適合：$(t-1)^2/2$
- カルバック密度適合：$t\log t - t$
- 最小二乗密度比適合：$(t-1)^2/2$
- ロバスト密度比適合：$(t^{1+\alpha} - t)/\alpha,\ (\alpha > 0)$

2.6 次元削減付き密度比推定

$\boldsymbol{x}$ を線形射影により $\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{v}\end{bmatrix}$ と分解したときに，$\boldsymbol{v}$ 成分が $p^*_{\mathrm{nu}}(\boldsymbol{x})$ と $p^*_{\mathrm{de}}(\boldsymbol{x})$ で共通，すなわち，ある共通の $p^*(\boldsymbol{v}|\boldsymbol{u})$ を用いて $p^*_{\mathrm{nu}}(\boldsymbol{x})$ と $p^*_{\mathrm{de}}(\boldsymbol{x})$ が

$$p^*_{\mathrm{nu}}(\boldsymbol{x}) = p^*(\boldsymbol{v}|\boldsymbol{u})p^*_{\mathrm{nu}}(\boldsymbol{u}),\ p^*_{\mathrm{de}}(\boldsymbol{x}) = p^*(\boldsymbol{v}|\boldsymbol{u})p^*_{\mathrm{de}}(\boldsymbol{u})$$

と表現できるならば，密度比 $r^*(\boldsymbol{x})$ を $p^*_{\mathrm{nu}}(\boldsymbol{u})/p^*_{\mathrm{de}}(\boldsymbol{u})$ と簡略化できる．従って，$\boldsymbol{u}$ が属する部分空間（異分布部分空間とよぶ）を特定すれば，高次元の密度比推定問題を低次元の問題に還元できる．異分布部分空間の探索は，教師付き次元削減手法により $\{\boldsymbol{x}^{\mathrm{nu}}_i\}_{i=1}^{n_{\mathrm{nu}}}$ と $\{\boldsymbol{x}^{\mathrm{de}}_j\}_{j=1}^{n_{\mathrm{de}}}$ を最も良く分離する部分空間を求める，あるいは，$p^*_{\mathrm{nu}}(\boldsymbol{u})$ から $p^*_{\mathrm{de}}(\boldsymbol{u})$ へのピアソン距離

$$\mathbb{E}_{p^*_{\mathrm{de}}}[(p^*_{\mathrm{nu}}(\boldsymbol{u})/p^*_{\mathrm{de}}(\boldsymbol{u}) - 1)^2]$$

を最大にする部分空間を求めることにより行う．

3 密度比に基づく機械学習

本節では，密度比推定により解決できる機械学習タスクの例を示す．

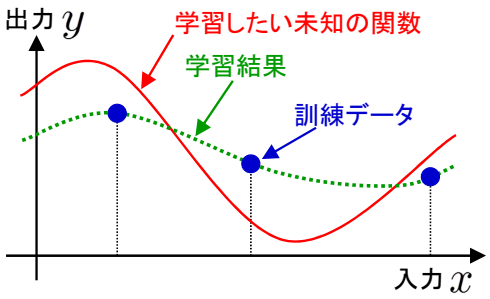


図 2: 教師付き学習．入力と出力が対になった訓練データから，その背後に潜んでいる入出力関係を推定する．

3.1 重点標本化

入力 $\boldsymbol{x}$ から出力 y への変換規則を学習する教師付き学習（図 2）において，訓練標本とテスト標本の入力分布が $p^*_{\rm tr}(\boldsymbol{x})$ から $p^*_{\rm te}(\boldsymbol{x})$ に変化するが，入出力関係 $p^*(y|\boldsymbol{x})$ は変化しない状況を共変量シフトとよぶ [2]．共変量シフト下では最尤推定などの学習法はバイアスを持つが，このバイアスは損失関数を重要度 $p^*_{\rm te}(\boldsymbol{x})/p^*_{\rm tr}(\boldsymbol{x})$ に従って重み付けすることにより打ち消される．すなわち，損失関数 $\ell(\boldsymbol{x})$ の $p^*_{\rm te}(\boldsymbol{x})$ に関する期待値は，損失関数の $p^*_{\rm tr}(\boldsymbol{x})$ に関する重要度重み付き期待値により計算できる（図 3）．

$$\begin{aligned}\mathbb{E}_{p^*_{\rm te}}[\ell(\boldsymbol{x})] &= \int \ell(\boldsymbol{x}) p^*_{\rm te}(\boldsymbol{x}) {\rm d}\boldsymbol{x} \\ &= \int \ell(\boldsymbol{x}) \frac{p^*_{\rm te}(\boldsymbol{x})}{p^*_{\rm tr}(\boldsymbol{x})} p^*_{\rm tr}(\boldsymbol{x}) {\rm d}\boldsymbol{x} = \mathbb{E}_{p^*_{\rm tr}}\left[\ell(\boldsymbol{x}) \frac{p^*_{\rm te}(\boldsymbol{x})}{p^*_{\rm tr}(\boldsymbol{x})}\right]\end{aligned}$$

交差確認などのモデル選択法も共変量シフト下では不偏性を失うが，同様に重要度重み付けを行うことにより不偏性が回復できる．

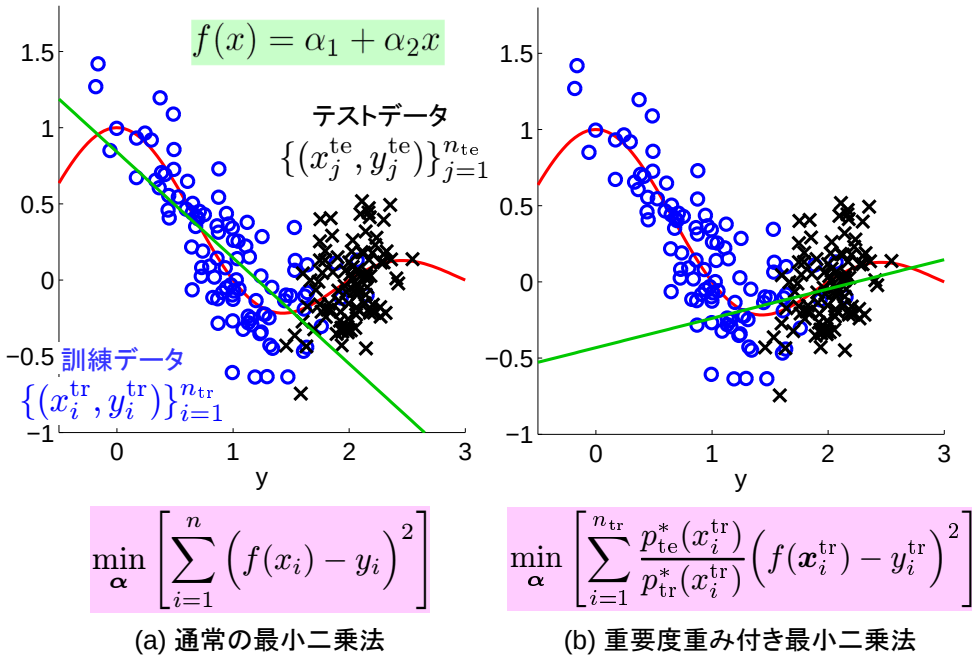
3.2 確率分布比較

正常標本集合に基づいて，評価標本集合に含まれる異常値を検出する問題を考える．これら二つの標本集合に対する密度比を考えれば，正常値に対する密度比の値は 1 に近く，異常値に対する密度比の値は 1 から大きく離れる．従って，密度比の値を評価基準とすることにより異常値を検出できる（図 4）．

また，密度比 $r^*(\boldsymbol{x}) = p^*_{\rm nu}(\boldsymbol{x})/p^*_{\rm de}(\boldsymbol{x})$ の推定量 $\widehat{r}(\boldsymbol{x})$ を用いることにより，分布 $p^*_{\rm nu}(\boldsymbol{x}), p^*_{\rm de}(\boldsymbol{x})$ 間の距離を精度良く推定できる [3].

- カルバック距離：$\frac{1}{n_{\rm nu}} \sum_{i=1}^{n_{\rm nu}} \log \widehat{r}(\boldsymbol{x}^{\rm nu}_i)$
- ピアソン距離：$\frac{2}{n_{\rm nu}} \sum_{i=1}^{n_{\rm nu}} \widehat{r}(\boldsymbol{x}^{\rm nu}_i)$
  $-\frac{1}{n_{\rm de}} \sum_{j=1}^{n_{\rm de}} \widehat{r}(\boldsymbol{x}^{\rm de}_j)^2 - 1$

これらの距離推定量を用いれば，並べ替え検定により二つの分布の同一性を検定できる．



(a) 通常の最小二乗法　　(b) 重要度重み付き最小二乗法

図 3: 重要度重み付き最小二乗法による共変量シフト適応．(a) 通常の最小二乗法ではモデルを訓練データに適合させるため，テストデータが訓練データと異なる分布に従う場合はテストデータをうまく予想できない．(b) テストデータに近い訓練データに強い重みをつけることにより，テスト出力に適合させる．

3.3 相互情報量推定

確率密度 $p^*_{\mathrm{x,y}}(\boldsymbol{x}, \boldsymbol{y})$ を持つ分布に独立に従う n 個の標本 $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ から，$\boldsymbol{x}$ と $\boldsymbol{y}$ の相互情報量

$$\mathbb{E}_{p^*_{\mathrm{x,y}}} \left[\log \frac{p^*_{\mathrm{x,y}}(\boldsymbol{x}, \boldsymbol{y})}{p^*_{\mathrm{x}}(\boldsymbol{x}) p^*_{\mathrm{y}}(\boldsymbol{y})} \right]$$

を推定する問題を考える．ただし，$p^*_{\mathrm{x}}(\boldsymbol{x})$ と $p^*_{\mathrm{y}}(\boldsymbol{y})$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ の周辺密度である．$\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}_{k=1}^n$ を分子の確率分布からの標本とみなし，$\{(\boldsymbol{x}_k, \boldsymbol{y}_{k'})\}_{k,k'=1}^n$ を分母の確率分布からの標本とみなせば，密度比推定により相互情報量が推定できる．同様に，二乗損失版の相互情報量

$$\mathbb{E}_{p^*_{\mathrm{x}}(\boldsymbol{x})} \mathbb{E}_{p^*_{\mathrm{y}}(\boldsymbol{y})} \left[\left(\frac{p^*_{\mathrm{x,y}}(\boldsymbol{x}, \boldsymbol{y})}{p^*_{\mathrm{x}}(\boldsymbol{x}) p^*_{\mathrm{y}}(\boldsymbol{y})} - 1 \right)^2 \right]$$

も推定できる．相互情報量は確率変数間の独立性を表す指標であり，その推定量は，独立性検定，特徴選択，特徴抽出，クラスタリング，独立成分分析，オブジェクト適合，因果推定など，様々な機械学習タスクに応用できる [4]（図 5）．

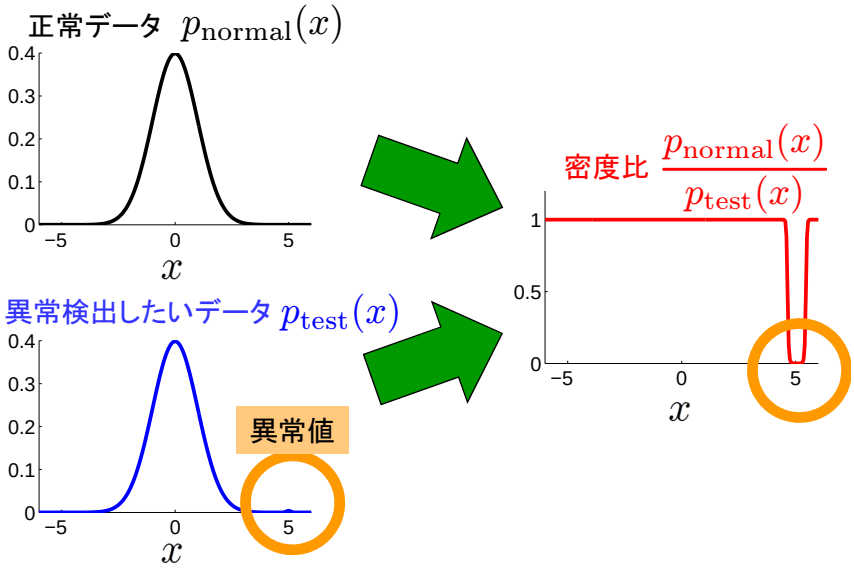


図 4: 密度比に基づく異常値検出．$p(\boldsymbol{x})$ の異常値を直接検出するのは困難だが（左下），正常データの密度（左上）との比を取ることにより，異常値が強調され検出が容易になる（右）．

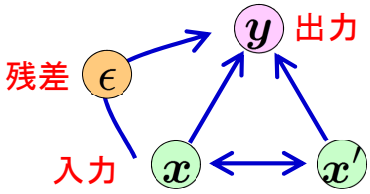


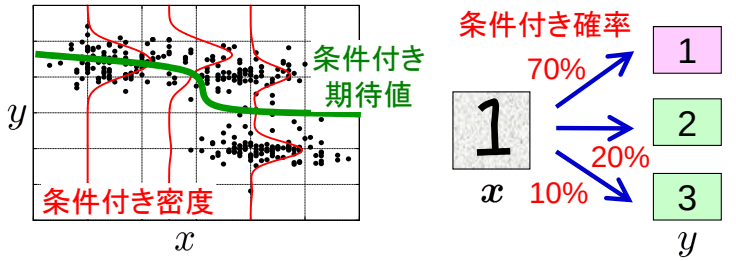
図 5: 相互情報量に基づく独立性判定．入出力間の独立性判定により，特徴選択，特徴抽出，クラスタリングなどが行え，入力間の独立性判定により，独立成分分析やオブジェクト適合などが行える．また，入力と残差の間の独立性判定により，因果推論が行える．

3.4 条件付き確率推定

確率密度 $p^*_{\mathrm{x,y}}(\boldsymbol{x},\boldsymbol{y})$ を持つ分布に独立に従う n 個の標本 $\{(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1}^n$ から，条件付き確率

$$p^*_{\mathrm{y|x}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{p^*_{\mathrm{x,y}}(\boldsymbol{x},\boldsymbol{y})}{p^*_{\mathrm{x}}(\boldsymbol{x})}$$

を推定する問題を考える．$\{(\boldsymbol{x}_k,\boldsymbol{y}_k)\}_{k=1}^n$ を分子の確率分布からの標本とみなし，$\{\boldsymbol{x}_k\}_{k=1}^n$ を分母の確率分布からの標本とみなせば，密度比推定により条件付き確率が推定できる（図 6）．$\boldsymbol{y}$ が連続変数の場合，これは条件付き密度推定に対応し，$\boldsymbol{y}$ がカテゴリ変数の場合は確率的パターン認識となる（図 6）．



(a) 条件付き密度推定　　(b) 確率的パターン認識

図 6: 条件付き確率推定．(a) 出力変数 $\boldsymbol{y}$ が連続値を取るとき，条件付き密度の推定に対応する．これは，条件付き期待値を推定する回帰分析の一般化になっており，出力の条件付き分布が多峰性や非対称性を持つときに有用である．(b) 出力変数 y がカテゴリ値を取るとき，確率的パターン認識とよばれ，カテゴリの予測だけでなく予測の信頼度も同時に得られる．

図 7: 二次のマルコフネットワーク．$k=1,\ldots,d$ が頂点で，$\|\boldsymbol{\theta}_{k,k'}\|>0$ のときに頂点 k,k' 間に枝があるグラフとして，確率分布が表現される．マルコフネットワークの変化は，グラフの枝の有無の変化に対応する．

4　密度比推定に基づく教師なし変化検知

本節では，確率分布比較の一例である変化検知について詳しく説明する．

教師なし変化検知の目的は，確率密度 $p^{\mathrm{A}}(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}_i^{\mathrm{A}}\}_{i=1}^{n_{\mathrm{A}}}$ と確率密度 $p^{\mathrm{B}}(\boldsymbol{x})$ を持つ確率分布に独立に従う標本 $\{\boldsymbol{x}_j^{\mathrm{B}}\}_{j=1}^{n_{\mathrm{B}}}$ からが与えられたとき，$p^{\mathrm{A}}(\boldsymbol{x})$ と $p^{\mathrm{B}}(\boldsymbol{x})$ が等しいかどうかを判定することである．

第 3.2 節で述べた方法で $p^{\mathrm{A}}(\boldsymbol{x})$ と $p^{\mathrm{B}}(\boldsymbol{x})$ の距離を推定すれば，$p^{\mathrm{A}}(\boldsymbol{x})$ と $p^{\mathrm{B}}(\boldsymbol{x})$ が等しいかどうかを判定できる．ここではさらに，$p^{\mathrm{A}}(\boldsymbol{x})$ と $p^{\mathrm{B}}(\boldsymbol{x})$ が二次のマルコフネットワーク

$$q(\boldsymbol{x};\boldsymbol{\theta}) \propto \exp\left(\sum_{k\geq k'} \boldsymbol{\theta}_{k,k'}^{\top} \boldsymbol{f}(x^{(k)}, x^{(k')})\right)$$

によりモデル化できると仮定する．ただし，$\boldsymbol{f}(x,x')$ は特徴ベクトルである．

このモデルのパラメータ $\boldsymbol{\theta}$ の変化から，多次元ベクトル $\boldsymbol{x}=(x^{(1)},\ldots,x^{(d)})^{\top}$ の要素 $x^{(k)}, x^{(k')}$ 間の相互作用の変化を捉えられる（図 7）．$\boldsymbol{f}(x,x')=xx'$ とおけば，二次のマルコフネットワークは正規モデルと一致し，$x^{(k)}, x^{(k')}$ 間の共分散をモデル化することに対

応する．$\boldsymbol{f}(x,x')=[x^t,x^{t-1}x',\ldots,x,x',1]^\top$ のような高次特徴を考えると，$x^{(k)},x^{(k')}$ 間の高次の共分散を考慮したモデル化が行える．

モデル $q(\boldsymbol{x};\boldsymbol{\theta})$ を標本 $\{\boldsymbol{x}_i^{\mathrm{A}}\}_{i=1}^{n_{\mathrm{A}}}$ と $\{\boldsymbol{x}_j^{\mathrm{B}}\}_{j=1}^{n_{\mathrm{B}}}$ にそれぞれ適合させ，推定したパラメータ $\widehat{\boldsymbol{\theta}}^{\mathrm{A}},\widehat{\boldsymbol{\theta}}^{\mathrm{B}}$ の差 $\widehat{\boldsymbol{\theta}}^{\mathrm{A}}-\widehat{\boldsymbol{\theta}}^{\mathrm{B}}$ を求めることにより，変数間の相互作用の変化を捉えられる．例えば，スパース最尤推定を用いると，

$$\max_{\boldsymbol{\theta}^{\mathrm{A}}}\sum_{i=1}^{n_{\mathrm{A}}}\log q(\boldsymbol{x}_i^{\mathrm{A}};\boldsymbol{\theta}^{\mathrm{A}})-\lambda_{\mathrm{A}}\sum_{k\geq k'}\|\boldsymbol{\theta}_{k,k'}^{\mathrm{A}}\|,\quad \lambda_{\mathrm{A}}\geq 0$$
$$\max_{\boldsymbol{\theta}^{\mathrm{B}}}\sum_{j=1}^{n_{\mathrm{B}}}\log q(\boldsymbol{x}_j^{\mathrm{B}};\boldsymbol{\theta}^{\mathrm{B}})-\lambda_{\mathrm{B}}\sum_{k\geq k'}\|\boldsymbol{\theta}_{k,k'}^{\mathrm{B}}\|,\quad \lambda^{\mathrm{B}}\geq 0$$

しかし，我々が知りたいのはパラメータが変化したかどうかだけであり，個々のパラメータの値 $\widehat{\boldsymbol{\theta}}^{\mathrm{A}},\widehat{\boldsymbol{\theta}}^{\mathrm{B}}$ を求めることが目的ではない．

そこで，二つのパラメータの差 $\boldsymbol{\theta}^{\mathrm{A}}-\boldsymbol{\theta}^{\mathrm{B}}$ を直接スパース化する学習法が提案された．

$$\max_{\boldsymbol{\theta}^{\mathrm{A}},\boldsymbol{\theta}^{\mathrm{B}}}\sum_{i=1}^{n_{\mathrm{A}}}\log q(\boldsymbol{x}_i^{\mathrm{A}};\boldsymbol{\theta}^{\mathrm{A}})+\sum_{j=1}^{n_{\mathrm{A}}}\log q(\boldsymbol{x}_j^{\mathrm{B}};\boldsymbol{\theta}^{\mathrm{B}})$$
$$-\gamma\sum_{k\geq k'}\|\boldsymbol{\theta}_{k,k'}^{\mathrm{A}}-\boldsymbol{\theta}_{k,k'}^{\mathrm{B}}\|,\quad \gamma\geq 0$$

しかし，この方法でもまだ二つのパラメータ $\boldsymbol{\theta}^{\mathrm{A}},\boldsymbol{\theta}^{\mathrm{B}}$ を明示的に推定している．また，マルコフネットワークの正規化項 $\int\exp\left(\sum_{k\geq k'}\boldsymbol{\theta}_{k,k'}^\top\boldsymbol{f}(x^{(k)},x^{(k')})\right)\mathrm{d}\boldsymbol{\theta}$ は，正規モデルやノンパラ正規モデル [5] など，限られたモデルに対してしか効率良く計算することができない．

これらの問題点は，パラメータ $\boldsymbol{\theta}^{\mathrm{A}},\boldsymbol{\theta}^{\mathrm{B}}$ を明示的に推定せず，それらの差 $\boldsymbol{\alpha}=\boldsymbol{\theta}^{\mathrm{A}}-\boldsymbol{\theta}^{\mathrm{B}}$ を直接推定することにより解決できる [6]．パラメータの差 $\boldsymbol{\alpha}$ の推定は，対数線形密度比モデルの学習に対応する．

$$r(\boldsymbol{x};\boldsymbol{\alpha})=\frac{q(\boldsymbol{x};\boldsymbol{\theta}^{\mathrm{A}})}{q(\boldsymbol{x};\boldsymbol{\theta}^{\mathrm{B}})}\propto\exp\left(\sum_{k\geq k'}\boldsymbol{\alpha}_{k,k'}^\top\boldsymbol{f}(x^{(k)},x^{(k')})\right)$$

対数線形密度比モデルは，第 2.3 節で紹介した一般化カルバック距離のもとでの密度適合法により，効率良く学習できる．

$$\min_{\boldsymbol{\alpha}}\int p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{p'(\boldsymbol{x})r(\boldsymbol{x};\boldsymbol{\alpha})}\mathrm{d}\boldsymbol{x}+\gamma\sum_{k\geq k'}\|\boldsymbol{\alpha}'_{k,k'}\|$$

5　まとめ

様々な機械学習タスクを統一的に解決できる密度比推定の枠組みを紹介した．密度比推定の精度や計算効率を向上させれば，密度比推定に基づく全ての機械学習アルゴリズムの

性能を改善できるため，密度比推定手法の更なる発展が望まれる．近年，密度比よりも安定した量である相対密度比 [7]

$$r^*_\beta(\boldsymbol{x}) = \frac{p^*_{\mathrm{nu}}(\boldsymbol{x})}{\beta p^*_{\mathrm{nu}}(\boldsymbol{x}) + (1-\beta)p^*_{\mathrm{de}}(\boldsymbol{x})}, \quad 0 \leq \beta < 1$$

や密度差 [8]

$$d(\boldsymbol{x}) = p^*_{\mathrm{nu}}(\boldsymbol{x}) - p^*_{\mathrm{de}}(\boldsymbol{x})$$

の直接推定法も研究されている．また，連続出力の隠れマルコフモデルの学習 [9] や教師なしラベル付け [10] など，密度比推定や密度差推定により解決できる新たな機械学習タスクを開拓することも今後の重要な課題である．

参考文献

[1] M. Sugiyama, T. Suzuki, and T. Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, Cambridge, UK, 2012.

[2] M. Sugiyama and M. Kawanabe, Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation, MIT Press, Cambridge, Massachusetts, USA, 2012.

[3] M. Sugiyama, S. Liu, M.C. duPlessis, M. Yamanaka, M. Yamada, T. Suzuki, and T. Kanamori, “Direct divergence approximation between probability distributions and its applications in machine learning.,” Journal of Computing Science and Engineering, vol.7, no.2, pp.99–111, 2013.

[4] M. Sugiyama, “Machine learning with squared-loss mutual information,” Entropy, vol.15, no.1, pp.80–112, 2013.

[5] H. Liu, J. Lafferty, and L. Wasserman, “The nonparanormal: Semiparametric estimation of high dimensional undirected graphs,” The Journal of Machine Learning Research, vol.10, pp.2295–2328, 2009.

[6] S. Liu, J. Quinn, M.U. Gutmann, and M. Sugiyama, “Direct learning of sparse changes in Markov networks by density ratio estimation,” Machine Learning and Knowledge Discovery, Part II, eds. by H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, vol.8189, pp.596–611, Lecture Notes in Computer Science, Springer, Berlin, Sep. 23–27 2013.

[7] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama, “Relative density-ratio estimation for robust distribution comparison,” Neural Computation, vol.25, no.5, pp.1324–1370, 2013.

[8] M. Sugiyama, T. Suzuki, T. Kanamori, M.C. duPlessis, S. Liu, and I. Takeuchi, "Density-difference estimation," Neural Computation, vol.25, no.10, pp.2734–2775, 2013.

[9] J.A. Quinn and M. Sugiyama, "Density ratio hidden Markov models," Technical Report 1302.3700, arXiv, 2013.

[10] M.C. duPlessis and M. Sugiyama, "Clustering unclustered data: Unsupervised binary labeling of two datasets having diffrent class balances," Technical Report 1305.0103, arXiv, 2013.