日本ロボット学会誌, *vol.33, no.2, pp.86–91, 2015.*

相互情報量を用いた機械学習とそのロボティクスへの応用

Machine Learning with Mutual Information and Its Application in Robotics

杉山 将[1,2]，入江 清[2,3]，友納 正裕[3]
[1]東京大学，[2]東京工業大学，[3]千葉工業大学
Masashi Sugiyama[1,2], Kiyoshi Irie[2,3], and Masahiro Tomono[3]
[1]The University of Tokyo, [2]Tokyo Institute of Technology,
and [3]Chiba Institute of Technology

Abstract

In this article, we review recent advances in machine learning with mutual information and its application in robotics. Software is available from `http://www.ms.k.u-tokyo.ac.jp/software.html`

Keywords

Machine learning, mutual information, divergence, map matching.

1　まえがき

インターネットやセンサー技術の発達と普及に伴い，文書，音声，画像，動画，電子商取引，電力，医療，生命，天文など様々な場面で膨大な量のデータが収集されるようになってきた．このようなビッグデータを活用して新たな価値を創造するためには，**機械学習**による知的なデータ解析が重要である．

機械学習には分類，回帰，特徴選択，異常検出など様々なデータ解析タスクが存在するが，古典的にはデータを生成する確率分布の推定により解析を行う**生成的アプローチ**が好まれてきた．なぜならば，データの生成確率を推定すればいくらでもデータを再生成できるため，生成的アプローチはあらゆるデータ解析に適用できるからである．しかし，データの生成過程に関する十分な事前知識がない場合は生成確率を精度良く推定することが困難であり，結果としてデータ解析の精度が低下するという問題があった．

そこで，データの生成確率の推定を経由せず，分類や回帰などのタスクを直接解こうというタスク特化型アプローチが近年盛んに研究されるようになった．サポートベクトルマシン [31] はその代表的な手法であり，データの生成確率を推定せずに直接パターン認識を行うことにより，認識精度を向上させた．しかし，解析すべきデータの量や次元数，および，データ解析タスクの複雑さは増大し続けており，機械学習の技術者やデータサイエンティストはアドホックなデータ解析に忙殺されるようになってしまった．

この問題を解決するために，我々は生成的アプローチとタスク特化型アプローチの折衷案を提案している [19]．具体的には，主要なデータ解析タスクの部分集合を考え，これらのタスク群に対して共通のデータ解析基盤技術を開発する．対象タスクを部分集合に限定することによって，すべてのタスクを対象とする生成的アプローチよりも性能の向上が期待される．また，主要なタスクをまとめて扱うことにより，各タスクを個別に扱うタスク特化型アプローチの高コスト問題を効果的に回避できる．

本稿では，そのようなデータ解析基盤技術の一例として，確率変数間の依存性の尺度である相互情報量を考える．第 2 節では，様々な相互情報量の定義と性質を示すとともに，確率分布の推定を介さないそれらの直接推定法を紹介する．次に，第 3 節では，相互情報量推定に基づく様々な機械学習アルゴリズムを概観する．具体的には，独立性検定 [18]，特徴選択 [26, 7]，特徴抽出 [25, 32, 12]，正準従属性分析 [8]，独立成分分析 [24]，オブジェクト適合 [33]，クラスタリング [17, 9, 11, 1]，因果推論 [34] のアルゴリズムを紹介する．そして第 4 節では，相互情報量のロボット自己位置推定への応用事例 [36] を紹介し，最後に第 5 節で今後の展望を述べる．

2 様々な相互情報量の定義と推定法

本節では，様々な種類の相互情報量の定義，性質，および，推定法を紹介する．

2.1 相互情報量

連続型確率変数 $\boldsymbol{x}$ と $\boldsymbol{y}$ の相互情報量 (mutual information, MI) は，

$$\mathrm{MI} = \iint p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x}) p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y} \tag{1}$$

で定義される．ただし，$p(\boldsymbol{x}, \boldsymbol{y})$，$p(\boldsymbol{x})$，$p(\boldsymbol{y})$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ の同時確率密度関数と周辺確率密度関数である．相互情報量は $p(\boldsymbol{x}, \boldsymbol{y})$ から $p(\boldsymbol{x})p(\boldsymbol{y})$ へのカルバック・ライブラー距離に対応しており，

$$\mathrm{MI} \geq 0 \quad \text{かつ} \quad \mathrm{MI} = 0 \Longleftrightarrow p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}) p(\boldsymbol{y})$$

を満たす．従って，相互情報量が大きいとき（小さいとき），$\boldsymbol{x}$ と $\boldsymbol{y}$ の従属性が高い（低い）とみなせる．

一般に確率密度関数 $p(\boldsymbol{x}, \boldsymbol{y})$，$p(\boldsymbol{x})$，$p(\boldsymbol{y})$ は未知であるため，相互情報量は直接計算できない．最も単純な相互情報量の推定法は，$p(\boldsymbol{x}, \boldsymbol{y})$，$p(\boldsymbol{x})$，$p(\boldsymbol{y})$ をそれぞれ標本 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$，$\{\boldsymbol{x}_i\}_{i=1}^n$，$\{\boldsymbol{y}_i\}_{i=1}^n$ から推定し，推定した確率密度関数を式 (1) に代入する方法であろう．しかしこの方法では，推定した確率密度関数の比を取ることによって推定誤差が拡大してしまうおそれがある．

そこで，それぞれの確率密度関数 $p(\boldsymbol{x}, \boldsymbol{y})$，$p(\boldsymbol{x})$，$p(\boldsymbol{y})$ を個別に推定せず，**確率密度比**

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \tag{2}$$

を直接推定する**最尤相互情報量推定法** (maximum likelihood MI, MLMI) が考案された [27]．最尤相互情報量推定法では，非負の基底関数ベクトル $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})$ を用いた密度比 $r(\boldsymbol{x}, \boldsymbol{y})$ のモデル

$$r_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \tag{3}$$

を，適当な制約条件のもとで対数尤度を最大化するように学習する．

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MLMI}} = \underset{\boldsymbol{\theta} \geq \boldsymbol{0}}{\mathrm{argmax}} \left[\sum_{i=1}^n \log \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right]$$
$$\text{subject to } \frac{1}{n^2} \sum_{i,i'=1}^n \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_{i'}) = 1$$

これは凸最適化問題であるため，射影勾配法などにより大域的な最適解 $\widehat{\boldsymbol{\theta}}_{\mathrm{MLMI}}$ を求められる．また，非負制約条件 $\boldsymbol{\theta} \geq \boldsymbol{0}$ により，$\widehat{\boldsymbol{\theta}}_{\mathrm{MLMI}}$ は疎ベクトルになるという特徴がある．$\widehat{\boldsymbol{\theta}}_{\mathrm{MLMI}}$ を用いて，MI は次式で近似できる．

$$\widehat{\mathrm{MI}} = \frac{1}{n} \sum_{i=1}^n \widehat{\boldsymbol{\theta}}_{\mathrm{MLMI}}^\top \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

カルバック・ライブラー距離に含まれる対数関数は原点付近で鋭い形状を持つため，確率密度の微小な変化を感度良く捉えられる．しかし一方では，標本からの相互情報量の推定が異常値の影響を受けやすいという問題もある．

2.2 二乗損失相互情報量

カルバック・ライブラー距離の異常値に対する過敏さは，対数関数を含まない**ピアソン距離**を用いれば改善できる．ピアソン距離を用いた相互情報量を**二乗損失相互情報量** (squared-loss MI, SMI) とよぶ [26]．

$$\mathrm{SMI} = \iint p(\boldsymbol{x})p(\boldsymbol{y}) \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1 \right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

カルバック・ライブラー距離とピアソン距離は，共に f**距離**とよばれるクラスに属し，似た性質を持つ．実際，二乗損失相互情報量に対しても

$$\mathrm{SMI} \geq 0 \quad \text{かつ} \quad \mathrm{SMI} = 0 \Longleftrightarrow p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$$

が成り立つため，二乗損失相互情報量も通常の相互情報量と同様に $\boldsymbol{x}$ と $\boldsymbol{y}$ の間の従属性尺度とみなせる．

二乗損失相互情報量は，モデル (3) を用いて密度比 (2) を直接推定する**最小二乗相互情報量推定法** (least squares MI, LSMI) によって精度良く近似できる．

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LSMI}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta}^\top \boldsymbol{G}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top \boldsymbol{h} + \lambda \boldsymbol{\theta}^\top \boldsymbol{\theta}\right]$$

ただし，$\lambda > 0$ は正則化パラメータであり，

$$\boldsymbol{G} = \frac{1}{n^2} \sum_{i,i'=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_{i'}) \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_{i'})^\top$$

$$\boldsymbol{h} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

である．最小二乗相互情報量推定法の解 $\widehat{\boldsymbol{\theta}}_{\mathrm{LSMI}}$ は，

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LSMI}} = (\boldsymbol{G} + \lambda \boldsymbol{I})^{-1} \boldsymbol{h}$$

と解析的に求められる．ただし，$\boldsymbol{I}$ は単位行列である．$\widehat{\boldsymbol{\theta}}_{\mathrm{LSMI}}$ を用いて，二乗損失相互情報量は次式で近似できる．

$$\widehat{\mathrm{SMI}} = \widehat{\boldsymbol{\theta}}_{\mathrm{LSMI}}^\top \boldsymbol{h} - 1$$

2.3 相対二乗損失相互情報量

確率密度関数の比 (2) は急峻な関数になる傾向があるため，二乗損失相互情報量を少数の標本から精度良く推定することは必ずしも容易でない．この問題は，密度比を滑らかにした**相対ピアソン距離**を用いれば改善できる．相対ピアソン距離を用いた相互情報量を**相対二乗損失相互情報量** (relative SMI, rSMI) とよぶ [35].

$$\mathrm{rSMI} = \iint q_\alpha(\boldsymbol{x}, \boldsymbol{y}) \left(\frac{p(\boldsymbol{x}, \boldsymbol{y})}{q_\alpha(\boldsymbol{x}, \boldsymbol{y})} - 1\right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

ただし，$0 \leq \alpha \leq 1$ に対して

$$q_\alpha(\boldsymbol{x}, \boldsymbol{y}) = \alpha p(\boldsymbol{x}, \boldsymbol{y}) + (1 - \alpha) p(\boldsymbol{x}) p(\boldsymbol{y})$$

とおいた．$\alpha = 0$ のとき相対二乗損失相互情報量はもとの二乗損失相互情報量に帰着され，$\alpha = 1$ のとき相対二乗損失相互情報量は常に 0 となる．α を大きく設定すると相対密度比関数は滑らかになるため，標本からの推定が容易になる．しかし，α が大き過ぎると従属性尺度としての感度が低下するため，実用上は α の値を適切に決定する必要がある．

相対二乗損失相互情報量は，最小二乗相互情報量推定法で用いた行列 $\boldsymbol{G}$ を

$$\boldsymbol{G}_\alpha = \frac{\alpha}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{y}_i)\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{y}_i)^\top + \frac{1-\alpha}{n^2}\sum_{i,i'=1}^{n}\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{y}_{i'})\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{y}_{i'})^\top$$

に置き換えるだけで，二乗損失相互情報量と全く同じアルゴリズムによって推定できる．

2.4 二次相互情報量

二次相互情報量 (quadratic MI, QMI) は，L_2 **距離**に基づく相互情報量である．

$$\mathrm{QMI} = \iint \Big(p(\boldsymbol{x},\boldsymbol{y}) - p(\boldsymbol{x})p(\boldsymbol{y})\Big)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

二次相互情報量には，二乗損失相互情報量のように確率密度関数の比 $\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}$ が含まれないため，雑音の影響を受けにくいという特徴がある．また，相対二乗損失相互情報量のように調整パラメータ α が含まれないため，実用上便利である．

二次相互情報量も，確率密度関数 $p(\boldsymbol{x},\boldsymbol{y})$，$p(\boldsymbol{x})$，$p(\boldsymbol{y})$ の推定を避けることにより精度良く近似できる．**密度差**

$$f(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x},\boldsymbol{y}) - p(\boldsymbol{x})p(\boldsymbol{y}) \tag{4}$$

を直接推定する**最小二乗二次相互情報量推定法** (least squares QMI, LSQMI)[11] では，密度差 $f(\boldsymbol{x},\boldsymbol{y})$ のモデル

$$f_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\beta}^\top\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{y}) \tag{5}$$

を，二乗誤差が最小になるように学習する．

$$\widehat{\boldsymbol{\beta}}_{\mathrm{LSQMI}} = \operatorname*{argmin}_{\boldsymbol{\beta}}\ \big[\boldsymbol{\beta}^\top\boldsymbol{U}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top\boldsymbol{v} + \lambda\boldsymbol{\beta}^\top\boldsymbol{\beta}\big]$$

ただし，$\lambda > 0$ は正則化パラメータであり，

$$\boldsymbol{U} = \iint \boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{\psi}(\boldsymbol{x},\boldsymbol{y})^\top \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}$$

$$\boldsymbol{v} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\boldsymbol{x}_i,\boldsymbol{y}_i) - \frac{1}{n^2}\sum_{i,i'=1}^{n}\boldsymbol{\psi}(\boldsymbol{x}_i,\boldsymbol{y}_{i'})$$

表 1: 様々な相互情報量

	推定量の計算	ロバスト性（感度）	調整パラメータ
相互情報量	数値的	低（高）	無
二乗損失相互情報量	解析的	中（中）	無
相対二乗損失相互情報量	解析的	調整可	有
二次相互情報量	解析的	高（低）	無

である．最小二乗二次相互情報量推定法の解 $\widehat{\boldsymbol{\theta}}_{\mathrm{LSQMI}}$ は，

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LSMI}} = (\boldsymbol{U} + \lambda \boldsymbol{I})^{-1}\boldsymbol{v}$$

と解析的に求められ，二次相互情報量は次式で近似できる．

$$\widehat{\mathrm{QMI}} = 2\widehat{\boldsymbol{\theta}}_{\mathrm{LSQMI}}^{\top}\boldsymbol{v} - \widehat{\boldsymbol{\theta}}_{\mathrm{LSQMI}}^{\top}\boldsymbol{U}\widehat{\boldsymbol{\theta}}_{\mathrm{LSQMI}}$$

2.5 まとめ

上記の相互情報量の性質を表 1 にまとめる．実用的には，推定量の計算の簡便さ，ロバスト性（感度）の高さ，調整パラメータの有無の観点から，適切な相互情報量の定義および推定法を選択することが望ましい．

3 相互情報量を用いた機械学習

本節では，相互情報量を用いた機械学習アルゴリズムを紹介する．

3.1 独立性検定

独立性検定の目的は，確率変数 $\boldsymbol{x}$ と $\boldsymbol{y}$ が統計的に独立かどうかを標本 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ から検定することである．$\boldsymbol{x}$ と $\boldsymbol{y}$ の独立性は，**並べ替え検定** [3] により検定できる [18].

具体的には，まず，もとの標本 $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ に対する相互情報量推定量 $\widehat{\mathrm{MI}}(\mathcal{D})$ を求める．次に，$\boldsymbol{x}_i$ と $\boldsymbol{y}_i$ の組合せをランダムに並び換えた標本 $\widetilde{\mathcal{D}}$ に対して相互情報量推定量 $\widehat{\mathrm{MI}}(\widetilde{\mathcal{D}})$ を求める．$\boldsymbol{x}_i$ と $\boldsymbol{y}_i$ の組合せをランダムに並び換えることにより $\boldsymbol{x}$ と $\boldsymbol{y}$ は統計的に独立になるため，$\widehat{\mathrm{MI}}(\widetilde{\mathcal{D}})$ はゼロに近い値を取る．このランダムな並び替えと相互情報量の推定を何度も繰り返し，相互情報量推定量 $\widehat{\mathrm{MI}}(\widetilde{\mathcal{D}})$ のヒストグラムを求めれば，もとの標本 $\mathcal{D}$ に対する相互情報量推定量 $\widehat{\mathrm{MI}}(\mathcal{D})$ の相対的な順位が計算できる．もし，$\widehat{\mathrm{MI}}(\mathcal{D})$ が $\widehat{\mathrm{MI}}(\widetilde{\mathcal{D}})$ のヒストグラムの上位 $\delta\%$ に入っていれば（通常は $\delta = 1$ か 5），$\boldsymbol{x}$ と $\boldsymbol{y}$ は従属だとみなす．

3.2 特徴選択・特徴抽出

回帰や分類など，入力 $\boldsymbol{x}$ から出力 $\boldsymbol{y}$ を予測する教師付き学習では，入力 $\boldsymbol{x}$ の次元が高いと出力 $\boldsymbol{y}$ の予測が困難である．このような場合，入力 $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ の各要素の中で出力 $\boldsymbol{y}$ の予測に有用なものだけを選べば，学習精度を向上させられる．これを**特徴選択**という．また，不要な要素を除去する代わりに，出力 $\boldsymbol{y}$ の予測に有用な $\boldsymbol{x}$ の低次元表現 $\boldsymbol{z}$ を求めることを**特徴抽出**という．

出力 $\boldsymbol{y}$ の予測に対する入力 $x^{(j)}$ の有用性を $\mathrm{MI}(x^{(j)}, \boldsymbol{y})$ で測ることにすれば，相互情報量推定量 $\widehat{\mathrm{MI}}$ を用いて特徴選択や特徴抽出が行える．例えば，出力 $\boldsymbol{y}$ の予測に最も有用な特徴 $x^{(\widehat{k})}$ は，

$$\widehat{k} = \operatorname*{argmax}_{j=1,\ldots,d} \widehat{\mathrm{MI}}(x^{(j)}, \boldsymbol{y})$$

で求められる [26]．また，$\boldsymbol{z} = \boldsymbol{T}\boldsymbol{x}$ と線形変換によって特徴抽出することにすれば，出力 $\boldsymbol{y}$ の予測に最も有用な表現 $\widehat{\boldsymbol{z}} = \widehat{\boldsymbol{T}}\boldsymbol{x}$ は，

$$\widehat{\boldsymbol{T}} = \operatorname*{argmax}_{\boldsymbol{T}:\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}} \widehat{\mathrm{MI}}(\boldsymbol{T}\boldsymbol{x}, \boldsymbol{y})$$

で求められる [25, 32]．ただし $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$ は $\boldsymbol{T}$ の正規化のための制約条件である．更に，ℓ_1 **正則化** [30] を用いれば，正方対角行列 $\boldsymbol{T} = \mathrm{diag}(t_1, \ldots, t_d)$ に対して，

$$(\widehat{t}_1, \ldots, \widehat{t}_d) = \operatorname*{argmax}_{t_1,\ldots,t_d \geq 0} \left[\widehat{\mathrm{MI}}(\boldsymbol{T}\boldsymbol{x}, \boldsymbol{y}) + \lambda \sum_{j=1}^{d} t_j \right]$$

によって特徴選択が行える [7]．ただし，$\lambda \geq 0$ は特徴数を調整する正則化パラメータである．

入力 $\boldsymbol{x}$ しか与えられない教師なし学習でも，$\boldsymbol{x}$ と $\boldsymbol{T}\boldsymbol{x}$ の相互情報量を用いれば特徴抽出（および特徴選択）が行える [12]．

$$\widehat{\boldsymbol{T}} = \operatorname*{argmax}_{\boldsymbol{T}:\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}} \widehat{\mathrm{MI}}(\boldsymbol{T}\boldsymbol{x}, \boldsymbol{x})$$

通常の教師なし特徴抽出法では，調整パラメータの値を勘と経験で決める必要があるため，得られた特徴抽出結果に主観性が残る．一方，相互情報量を用いた特徴抽出法では，相互情報量最大化の原理に基づいて系統的に調整パラメータの値を決定できるため，得られた結果の客観性が担保できる．

3.3 正準従属性分析

正準相関分析 [4] は，2 つの確率変数 $\boldsymbol{x}$ と $\boldsymbol{x}'$ に対する特徴抽出法であり，$\boldsymbol{T}\boldsymbol{x}$ と $\boldsymbol{T}'\boldsymbol{x}'$ の間の相関を最大にする行列 $\boldsymbol{T}$ と $\boldsymbol{T}'$ を求める．**正準従属性分析** [8] はその拡張であり，$\boldsymbol{T}\boldsymbol{x}$

と $\boldsymbol{T}'\boldsymbol{x}'$ の間の従属性を最大にする行列 $\boldsymbol{T}$ と $\boldsymbol{T}'$ を求める．正準従属性分析では，無相関でも従属性の高い特徴が抽出できる．

相互情報量を従属性の尺度として用いることにすれば，

$$(\widehat{\boldsymbol{T}}, \widehat{\boldsymbol{T}}') = \underset{\boldsymbol{T},\boldsymbol{T}':\boldsymbol{T}\boldsymbol{T}^\top=\boldsymbol{I},\boldsymbol{T}'\boldsymbol{T}'^\top=\boldsymbol{I}}{\operatorname{argmax}} \widehat{\mathrm{MI}}(\boldsymbol{T}\boldsymbol{x}, \boldsymbol{T}'\boldsymbol{x}')$$

によって正準従属性分析が行える．

3.4 独立成分分析

独立成分分析 [6] の目的は，d 個の独立な信号 $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})^\top$ が未知の行列 $\boldsymbol{U}$ によって混合された信号

$$\boldsymbol{y} = (y^{(1)}, \ldots, y^{(d)})^\top = \boldsymbol{U}\boldsymbol{x}$$

を，もとの独立な信号に分解することである．独立成分分析では，

$$\boldsymbol{z} = (z^{(1)}, \ldots, z^{(d)})^\top = \boldsymbol{T}\boldsymbol{y}$$

と分解された各信号 $z^{(1)}, \ldots, z^{(d)}$ を最も独立にする分解行列 $\boldsymbol{T}$ を求める．

d 変数 $z^{(1)}, \ldots, z^{(d)}$ に対する相互情報量は，同時確率密度関数 $p(z^{(1)}, \ldots, z^{(d)})$ から周辺確率密度関数の積 $p(z^{(1)}) \times \cdots \times p(z^{(d)})$ への距離によって定義される．これを用いれば，相互情報量推定量の最小化によって独立成分分析が行える [24]．

$$\widehat{\boldsymbol{T}} = \underset{\boldsymbol{T}:\boldsymbol{T}\boldsymbol{T}^\top=\boldsymbol{I}}{\operatorname{argmin}} \widehat{\mathrm{MI}}(z^{(1)}, \ldots, z^{(d)})$$

3.5 オブジェクト適合

オブジェクト適合の目的は，対になっていない標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ と $\{\boldsymbol{y}_i\}_{i=1}^n$ を並び替えて，適合するように対 $\mathcal{D}_\pi = \{(\boldsymbol{x}_i, \boldsymbol{y}_{\pi(i)})\}_{i=1}^n$ にすることである [33]．ただし，π は $1, \ldots, n$ に対する並び替え関数である．

オブジェクト $\boldsymbol{x}$ と $\boldsymbol{y}$ の適合度を相互情報量で測ることにすれば，その最大化によってオブジェクト適合が行える．

$$\widehat{\pi} = \underset{\pi}{\operatorname{argmax}} \widehat{\mathrm{MI}}(\mathcal{D}_\pi)$$

3.6 クラスタリング

クラスタリングの目的は，入力標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ に対して，性質の似た（異なる）標本が同じ（別の）クラスタに属するようにクラスタラベル $\{y_i \mid y_i \in \{1,\ldots,c\}\}_{i=1}^n$ を与えることである．

クラスタリングを，多次元ベクトル標本 $\{\boldsymbol{x}_i\}_{i=1}^n$ を 1 次元のスカラー $\{y_i\}_{i=1}^n$ に情報圧縮する問題と捉えれば，$\{\boldsymbol{x}_i\}_{i=1}^n$ に対する相互情報量を最大にする $\{y_i\}_{i=1}^n$ を求める問題に帰着できる [17, 9, 11, 1]．

$$(\widehat{y}_1,\ldots,\widehat{y}_n) = \operatorname*{argmax}_{(y_1,\ldots,y_n)} \widehat{\mathrm{MI}}(\{(\boldsymbol{x}_i,y_i)\}_{i=1}^n)$$

通常のクラスタリングでは，調整パラメータの値を勘と経験で決める必要があるため，得られたクラスタリング結果に主観性が残る．一方，相互情報量を用いたクラスタリング法では，相互情報量最大化の原理に基づいて系統的に調整パラメータの値を決定できるため，得られた結果の客観性が担保できる．

3.7 因果推論

因果推論の目的は，相関のある標本対 $\{(x_i,y_i)\}_{i=1}^n$ から，x と y の因果関係（x が y の原因か，y が x の原因か）を知ることである [10]．非線形関数 f と非正規加法性雑音 e を用いて x から y への因果関係が $y = f(x) + e$ で与えられるとき，適当な条件のもとで，y から x への関係は加法性雑音を用いて表現できないことが知られている [5]．従って，入力と残差ができるだけ独立になるように，x から y への因果モデル $y = f(x) + e$ と y から x への因果モデル $x = f'(y) + e'$ を標本 $\{(x_i,y_i)\}_{i=1}^n$ から学習すれば，因果方向が同定できる．

入力と残差の独立性の判定に相互情報量を用いることにすれば，相互情報量推定量が最小になるように因果モデル f と f' を学習することによって因果推論が行える [34]．

$$\operatorname*{argmin}_f \widehat{\mathrm{MI}}(x, y - f(x)) \qquad \operatorname*{argmin}_{f'} \widehat{\mathrm{MI}}(y, x - f'(y))$$

4 相互情報量のロボット自己位置推定への応用

前節で紹介した特徴抽出，正準従属性分析，オブジェクト適合，クラスタリングの手法では，相互情報量の最大化を通して性質の異なるデータを適合している．本節では，この考え方を応用したロボット自己位置推定手法を紹介する．

自己位置推定とは，ロボットが地図上のどこにいるかをセンサデータから推定する問題であり，移動ロボットの重要な機能の一つである [38]．従来の自己位置推定法では，事前に収集したセンサデータを用いてあらかじめセンサ地図を構築しておき，実際の走行時の

センサデータと適合させることによって現在地を推定する [29]．しかしこのような手法では，事前にセンサ地図を構築するコストが高く，ロボットを初めての場所に持っていったときに直ちに運用できない．

これらの問題点は，Google Map や OpenStreetMap のような既存の 2 次元市街地図とセンサデータを直接適合させることにより原理的には克服できる．しかし市街地図とセンサデータは全く性質が異なるため，二乗誤差や相関などの標準的な尺度のもとではうまく適合できない．そこで，二乗損失相互情報量を用いた自己位置推定法 [36] が提案された．以下では，単眼カメラを用いた自由度 2 の自己位置推定（道路上の進行方向に対する回転角度 θ [deg] と左右の平行移動量 τ [m]）の例を紹介する．

まず，単眼カメラによって得られた画像（図 3(a)）の各画素から，色（RGB）とエッジ（強度と向き）の 5 次元の特徴ベクトル $\boldsymbol{x}$ を抽出する．一方，市街地図中の領域や境界線にはあらかじめ適当な ID 番号を付与しておき，各画素の ID 番号を y とする（図 1）．そして，ロボットの位置仮説 $\boldsymbol{w} = (\theta, \tau)$ のもとで，市街地図をカメラ画像平面に投影する．こうして，各画素の特徴ベクトル $\boldsymbol{x}_i$ と地図の ID 番号 $y_i^{(\boldsymbol{w})}$ を対応付けることができ，標本対 $\mathcal{D}_{\boldsymbol{w}} = \{(\boldsymbol{x}_i, y_i^{(\boldsymbol{w})})\}_{i=1}^n$ が得られる（図 2）．この標本対を用いて相互情報量を推定し，それを最大にするロボットの位置 $\widehat{\boldsymbol{w}}$ を求めることによって，自己位置推定を行う．

$$\widehat{\boldsymbol{w}} = \operatorname*{argmax}_{\boldsymbol{w}} \widehat{\mathrm{MI}}(\mathcal{D}_{\boldsymbol{w}})$$

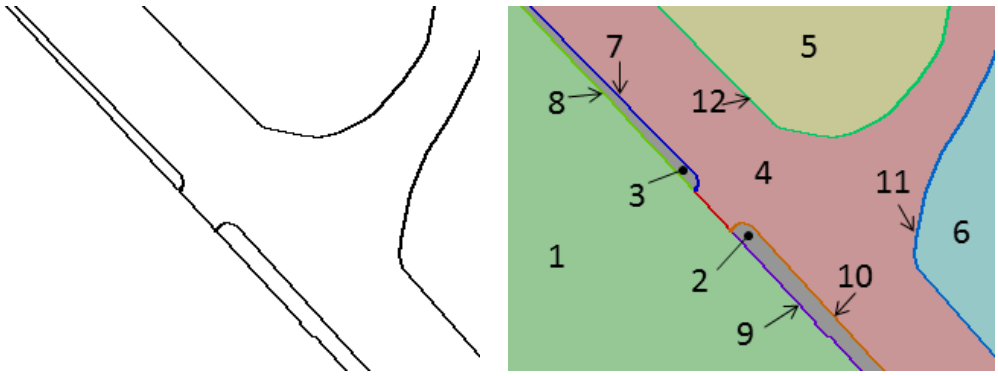
図 3 に実験結果の例を示す．この実験では，図 3(b) に示した真の自己位置 $\boldsymbol{w}^* = (\theta^*, \tau^*)$ の $\theta^* \pm 20$ [deg] および $\tau^* \pm 2$ [m] の範囲で総当り探索を行い，自己位置推定を行った．図 3(c) より，二乗損失相互情報量推定量は真の値 $\boldsymbol{w}^*$ 付近で最大値を取ることがわかる．この自己位置推定結果に基づいて，道路境界をカメラ画像に投影した結果を図 3(d) に示す．これより，二乗損失相互情報量推定量の最大化によって道路境界が推定できていることがわかる．

5　まとめと今後の展望

本稿では，様々な距離尺度に基づく相互情報量を導入し，確率分布を推定すること無く相互情報量を直接近似する手法を紹介した．このように確率分布の推定を経由せず，推定すべき量を直接求めるアプローチは**バプニックの原理** [31] とよばれており，近年の様々な機械学習アルゴリズムの設計原理となっている．

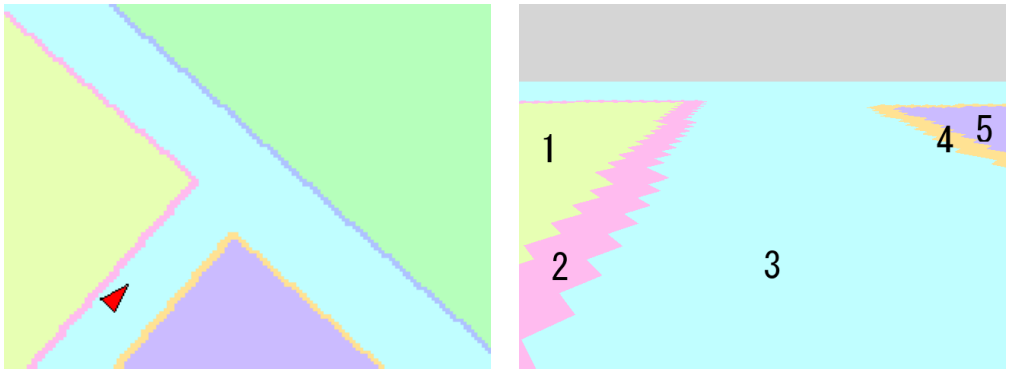
そして，相互情報量推定を用いた独立性検定，特徴選択，特徴抽出，正準従属性分析，独立成分分析，オブジェクト適合，クラスタリング，因果推論のアルゴリズムを概観した．相互情報量は，性質の異なるデータの適合に特に有効であり，単眼カメラを用いたロボット自己位置推定への応用例も紹介した．

このように，相互情報量推定は様々な機械学習技術の共通基盤となる重要な技術であり，ビッグデータ時代において，その重要性は益々高まっていくものと考えられる．近年，



(a) 市街地図　　(b) 内部表現

図 1: 市街地図の前処理：適当な領域や境界線に固有の ID 番号を割り当てる．



(a) 地図とロボット位置　　(b) ID 番号を画像に投影

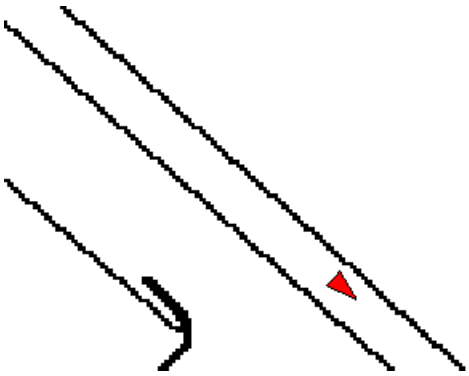
図 2: センサ情報 $\boldsymbol{x}$ と地図の ID 番号 $y^{(\boldsymbol{w})}$ との対応付け．画像の色やテクスチャをセンサ特徴 $\boldsymbol{x}$ とし，各画素に市街地図の ID 番号を (b) のように画像に割り当てる．

二乗損失相互情報量推定の改良法 [21] や，二次相互情報量の微分の直接推定法 [28] が提案されており，相互情報量の推定技術の更なる発展が期待される．

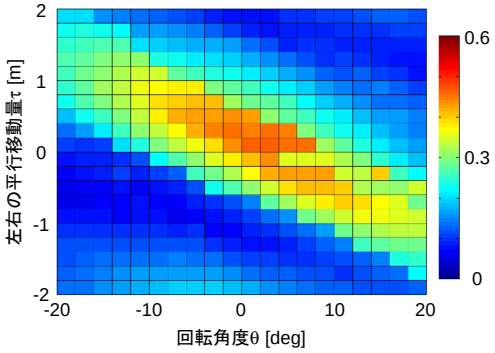
相互情報量を用いた機械学習技術は，信号や画像などのマルチメディアデータ解析 [32, 33] をはじめ，生命情報 [26, 34] や精密工学 [22] など幅広い分野に応用されている．ロボティクス分野においても，第 4 節で紹介した移動ロボットの自己位置推定法を発展させ，相互情報量を**パーティクルフィルタ** [2] に適用した位置追跡法が提案されている [37]．相互情報量の応用領域の更なる開拓が期待される．

第 1 節でも述べたように，機械学習によるデータ解析技術が応用分野で成功を収めるためには，推定精度や計算効率だけでなく，**汎用性**が重要である．例えば，相互情報量推定

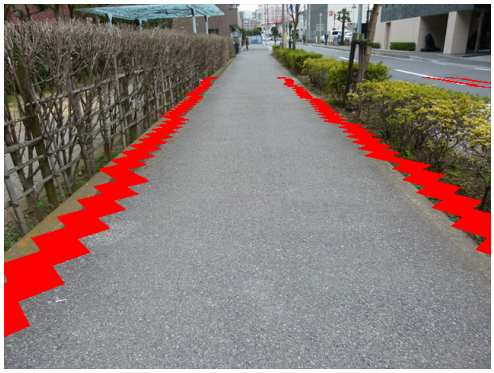
(a) 単眼カメラによる入力画像



(b) Google Map より得た市街地図（三角形は真の自己位置）



(c) 回転角度 $\theta^* \pm 20$ [deg] および左右の平行移動量 $\tau^* \pm 2$ [m] に対する二乗損失相互情報量推定量の値（原点が真値）



(d) 二乗損失相互情報量推定量の最大化によって推定した自己位置に基づいて，道路境界をカメラ画像に投影した結果

図 3: 単眼カメラからの自己位置推定の例.

で用いた**密度比推定** [19] や**密度差推定** [20]，および，**確率分布間の距離推定** [16] は幅広い応用をもつ汎用的な基盤技術であり，ロボティクス分野においても様々な応用が可能であると考えられる．また，近年提案された**密度微分推定** [23] も新たな汎用的基盤技術として注目されており，今後の更なる発展が期待される．

最後に，本稿では触れなかったが，機械学習分野における**強化学習** [15] とよばれる技術は，ロボットの**運動制御**に有用である [13]．ロボティクス分野での更なる活用が期待される．

謝辞

杉山将は，科学研究費補助金 25700022 の支援を受けた．

参考文献

[1] D. Calandriello, G. Niu, and M. Sugiyama. Semi-supervised information-maximization clustering. *Neural Networks*, 57:103–111, 2014.

[2] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, New York, NY, USA, 2001.

[3] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, NY, USA, 1993.

[4] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3–4):321–377, 1936.

[5] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21*, pages 689–696, 2009. MIT Press.

[6] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, New York, NY, USA, 2001.

[7] W. Jitkrittum, H. Hachiya, and M. Sugiyama. Feature selection via ℓ_1-penalized squared-loss mutual information. *IEICE Transactions on Information and Systems*, E96-D(7):1513–1524, 2013.

[8] M. Karasuyama and Sugiyama. Canonical dependency analysis based on squared-loss mutual information. *Neural Networks*, 34:46–55, 2012.

[9] M. Kimura and M. Sugiyama. Dependence-maximization clustering with least-squares mutual information. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 15(7):800–805, 2011.

[10] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2000.

[11] J. Sainui and M. Sugiyama. Direct approximation of quadratic mutual information and its application to dependence-maximization clustering. *IEICE Transactions on Information and Systems*, E96-D(10):2282–2285, 2013.

[12] J. Sainui and M. Sugiyama. Unsupervised dimension reduction via least-squares quadratic mutual information. *IEICE Transactions on Information and Systems*, E76-D(10):2806–2809, 2014.

[13] N. Sugimoto, V. Tangkaratt, T. Wensveen, T. Zhao, M. Sugiyama, and J. Morimoto. Efficient reuse of previous experiences in humanoid motor learning. In *Proceedings of IEEE-RAS International Conference on Humanoid Robots (HUMANOIDS2014)*, Madrid, Spain, Nov. 18-20, 2014.

[14] M. Sugiyama. Machine learning with squared-loss mutual information. *Entropy*, 15(1):80–112, 2013.

[15] M. Sugiyama. *Statistical Reinforcement Learning: Modern Machine Learning Approaches*. CRC Press, to appear.

[16] M. Sugiyama, S. Liu, M. C. du Plessis, 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*, 7(2):99–111, 2013.

[17] M. Sugiyama, G. Niu, M. Yamada, M. Kimura, and H. Hachiya. Information-maximization clustering based on squared-loss mutual information. *Neural Computation*, 26(1):84–131, 2014.

[18] M. Sugiyama and T. Suzuki. Least-squares independence test. *IEICE Transactions on Information and Systems*, E94-D(6):1333–1336, 2011.

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

[20] M. Sugiyama, T. Suzuki, and T. Kanamori, M. C. du Plessis, S. Liu, and I. Takeuchi. Density-difference estimation. *Neural Computation*, 25(10):2734–2775, 2013.

[21] T. Sakai and M. Sugiyama. Computationally efficient estimation of squared-loss mutual information with multiplicative kernel models. *IEICE Transactions on Information and Systems*, E97-D(4):968–971, 2014.

[22] T. Sakai, M. Sugiyama, K. Kitagawa, and K. Suzuki. Registration of infrared transmission images using squared-loss mutual information. *Precision Engineering*, 39:187–193, 2015.

[23] H. Sasaki, Y.-K. Noh, and M. Sugiyama. Direct density-derivative estimation and its application in KL-divergence approximation. *arXiv*, 1406.7638, 2014.

[24] T. Suzuki and M. Sugiyama. Least-squares independent component analysis. *Neural Computation*, 23(1):284–301, 2011.

[25] T. Suzuki and M. Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural Computation*, 3(25):725–758, 2013.

[26] T. Suzuki, M. Sugiyama, T. Kanamori, and J. Sese. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, 10(1):S52 (12 pages), 2009.

[27] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. In *Proceedings of ECML-PKDD2008 Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery (FSDM2008)*, pages 5–20, Antwerp, Belgium, Sep. 15 2008.

[28] V. Tangkaratt, H. Sasaki, and M. Sugiyama. Direct estimation of the derivative of quadratic mutual information with application in sufficient dimension reduction. submitted.

[29] S. Thrun, D. Fox, and and F. Dellaert W. Burgard. Robust Monte Carlo localization for mobile robots. *Artificial Intelligence*, 128(1–2):99–141, 2000.

[30] R. Tibshirani. Regression shrinkage and subset selection with the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[31] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

[32] M. Yamada, G. Niu, J. Takagi, and M. Sugiyama. Computationally efficient sufficient dimension reduction via squared-loss mutual information. In *Proceedings of the Third Asian Conference on Machine Learning (ACML2011)*, pages 247–262, Taoyuan, Taiwan, Nov. 13-15 2011.

[33] M. Yamada and M. Sugiyama. Cross-domain object matching with model selection. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS2011)*, pages 807–815, Fort Lauderdale, Florida, USA, Apr. 11-13 2011.

[34] M. Yamada, M. Sugiyama, and J. Sese. Least-squares independence regression for non-linear causal inference under non-Gaussian noise. *Machine Learning*, 96(3):249–267, 2014.

[35] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5):1324–1370, 2013.

[36] 入江 清, 杉山 将, 友納 正裕. 初めて訪れる場所における 2 次元市街地図を用いた自己位置推定. **第 *32* 回日本ロボット学会学術講演会講演論文集**, RSJ2014AC2I2-02, 福岡, 2014 年 9 月 4〜6 日.

[37] 入江 清, 杉山 将, 友納 正裕. 初めて訪れる場所における二次元市街地図を用いた自己位置推定. **第 *20* 回ロボティクスシンポジア講演論文集**, 投稿中.

[38] 友納 正裕. 移動ロボットのための確率的な自己位置推定と地図構築. **日本ロボット学会誌**, 29(5):423–426, 2011.