3P1-8in2

ノイズあり属性統計量からの学習

Learning from Attribute Statistics with Noise

吉川 友也 *1
Yuya Yoshikawa

*1千葉工業大学 人工知能・ソフトウェア技術研究センター
Software Technology and Artificial Intelligence Research Laboratory, Chiba Institute of Technology

Estimating the attributes of SNS users is an important technique for marketing and advertisement delivery systems. The attributes are typically estimated in a supervised learning manner. However, labeling the users with their attributes manually is difficult and expensive. In this paper, we address the problem of learning a classifier that predicts users' attributes only from the attribute statistics of websites that the users have visited. To tackle this problem, we propose a probabilistic generative model for the attribute statistics, in which can capture the users' attributes as hidden variables. Moreover, since the attribute statistics of the SNS and the websites may be different, the proposed model is modeled so as to treat the difference as noise variables. In the experiment, we show that the proposed model outperforms the existing methods on synthetic datasets.

1. はじめに

個人の属性推定は，性別，年齢，職業，人種，居住地等をその人を表す特徴量から推定するタスクである．SNS 等のインターネットサービスでは，個人の属性情報を使い，マーケティングや広告配信システムを作成する．しかし，そのような属性情報が得られないケースが多いため，属性推定技術の開発が必要である．

機械学習による典型的な属性推定の方法は，正解となる属性情報を教師データとして作成し，そのデータから属性識別器を学習する方法である．しかし，この方法では，個人に属性ラベルを人手で付ける作業が発生するが，この作業は簡単ではない．例えば，SNS 上のユーザに対して属性ラベルを付けることを考える．性別は，ユーザの顔写真や発言内容等から比較的簡単にラベル付けできる．しかし，年齢は 30 代と 40 代の違いを正しくラベル付けすることは難しい．また，属性識別器の性能を上げるためには，より多くの属性ラベル付きデータが必要であるが，人手での作業のためコストがかかる．

図 1は本論文で想定するシチュエーションを示す．今，SNS 上の個人の属性を推定したいとし，個人を表す特徴量とどの企業アカウントにリンクを張っているか（フォローしているか）が既知であるとする．また，企業アカウントは Web サイトを持っており，その Web サイトに訪れるユーザの属性統計量（例えば，訪問ユーザの性別比が男性 60%，女性 40%）は，Quantcast*1 等によって提供されるとする．

このシチュエーションの下で，本稿では，個人の属性情報を教師データとして使わずに，個人の属性を推定する手法を提案する．観測値として，人手でラベル付けしたユーザ属性情報の代わりに，そのユーザが訪れた Web サイトとその Web サイトの属性統計量が与えられる．提案法では，個人の特徴量から属性が生成され，Web サイトの属性統計量はその Web サイトに訪問したことがある個人の属性の統計量として生成されると仮定する．ただし，Web サイトの属性統計量と，我々が知りたい個人の属性の統計量は異なることが想定される．すなわち，

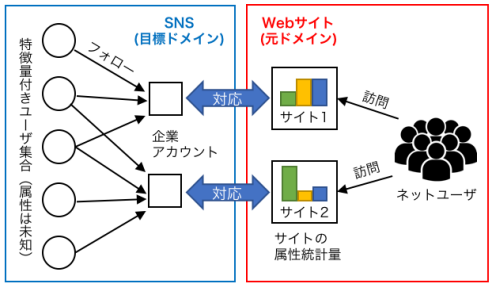

図 1: 想定するシチュエーション

Web サイトに訪れる集団と，属性を知りたい個人の集団は異なるため，属性統計量にズレが生じる．提案法では，このズレを考慮するようにモデル化を行う．実験では，提案法が図 1のシチュエーションにおいて有効であることを人工データを用いて示す．

2. 関連研究

これまで，インターネット上の個人の属性推定は多数行われてきた．推定する属性は，職業 [Li 14]，性別・年齢 [Rao 10]，民族 [Mislove 11]，居住地 [Cheng 10] 等がある．個人を表す特徴としては，SNS 上の発言 [Rao 10]，顔画像 [Kumar 09]，友人関係 [Dong 14] 等が用いられる．提案法は，特定の属性や特徴に特化しない汎用的なモデルであり，このモデルを拡張することにより，様々な属性や特徴に特化したモデルを作ることができる．また，これらの研究では，個人に対して属性ラベルが付けられ，それが教師データとして与えられると仮定する．提案法は，そのような作成にコストの掛かる教師データは使わずに，属性分類器を構築することができる．

提案法は，[Culotta 15] で提案された属性推定のアイディアから着想を得ている．彼らの論文では，図 1と同等のシチュエーションを想定し，個人の特徴から Web サイトの属性統計量を回帰するモデルに基づき，属性分類器を構築する．そのようなやり方では，企業アカウントをフォローする個人の属性統計量

連絡先: 吉川 友也，yoshikawa@stair.center
https://sites.google.com/site/yuyay4ml/

*1 https://www.quantcast.com/

と，その企業の Web サイトを訪問する個人の属性統計量が一致する前提で，属性分類器が構築される．提案法は，SNS と Web サイトのドメイン間の属性統計量の違いが考慮できるようにモデル化している点が異なる．

3. 提案法

3.1 モデル

提案法を一般的に説明するために，属性を予測したい個人がいる世界を「目標ドメイン」，サイトの属性統計量が観測できる世界を「元ドメイン」と呼ぶ．また，目標ドメインと元ドメインのサイトは一対一対応が取れているとする．

観測値として，目標ドメインにおける U 人の特徴量 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots, \boldsymbol{x}_U]^\top \in \mathbb{R}^{U\times D}$ と，元ドメインにおけるサイトの属性統計量 $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_N]^\top \in \mathbb{R}_+^{N\times M}$ が与えられる．ここで，D は特徴量の次元数，M は属性のクラス数を表す．また，各サイトの属性統計量の総和は必ず 1 となる．すなわち，$\sum_{m=1}^{M} s_{im} = 1$ である．E_i は，目標ドメインにおけるサイト i を訪問したことがあるユーザ集合を表す．簡単のため，$E = \{E_1, E_2, \cdots, E_N\}$ とする．

提案モデルでは，各個人 $u = 1, 2, \cdots, U$ の属性 $y_u \in \{1, 2, \cdots, M\}$ の確率は，その個人の特徴量 $\boldsymbol{x}_u$ と各属性の重みベクトル $\boldsymbol{w}_m \in \mathbb{R}^D$ の内積 $\boldsymbol{w}_m^\top \boldsymbol{x}_u$ を入力とするソフトマックス関数によって与えるとする．これは多クラスロジスティック回帰と同じアイディアであるが，提案法においては，y_u は潜在変数である．また，重みベクトル $\boldsymbol{w}_m$ は等分散ガウス分布 $\mathcal{N}(\boldsymbol{0}, \alpha_w^{-1}\boldsymbol{I}_D)$ から生成される．ここで，$\boldsymbol{I}_D$ は D 次元の単位行列である．

目標ドメインと元ドメインのサイトでは，ユーザの属性分布が異なると仮定する．この違いを表現するために，混同行列 $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_M]^\top$ を導入する．$\boldsymbol{c}_m = [c_{m1}, c_{m2}, \cdots, c_{mM}]$ は，目標ドメインにおいて個人の属性が m の時，元ドメインではどの属性に割り当てられるかの確率を表す．従って，個人 u の属性ラベルが y_u のとき，元ドメインのサイト i における属性 t_{ui} はカテゴリカル分布 $\mathrm{Cat}(\boldsymbol{c}_{y_u})$ から生成される．なお，各属性 m について，$\sum_{l=1} c_{ml} = 1$ となる必要があるため，$\boldsymbol{c}_m$ はディリクレ分布から生成することが適当である．ただし，多くの場合，目標ドメインと元ドメインの間で属性は変化しないと思われるため，混同行列 $\boldsymbol{C}$ の対角成分の値は比較的大きくなりやすいと考えられる．この事前知識を導入するため，各属性 m ごとにディリクレパラメータ $\boldsymbol{\beta}_m \in \mathbb{R}_+^M$ を超パラメータとして与える．ここで，

$$\beta_{ml} = \begin{cases} \alpha_{c_0} & (m \neq l) \\ \alpha_{c_1} & (m = l). \end{cases} \tag{1}$$

$\alpha_{c_1} > \alpha_{c_0}$ とすることにより，混同行列 $\boldsymbol{C}$ の対角成分は大きな値になりやすい．

$\boldsymbol{t}_i$ を個人 $u \in E_i$ の t_{ui} の統計量とする．すなわち，$\boldsymbol{t}_i = \frac{1}{|E_i|}\sum_{u\in E_i} v(t_{ui})$ とする．ここで，$v(t)$ は t 次元の値は 1 でそれ以外は 0 の M 次元ベクトルである．最終的に，元ドメインのサイト i の属性統計量 $\boldsymbol{s}_i$ は，$\boldsymbol{t}_i$ を平均とする等分散ガウス分布 $\boldsymbol{s}_i \sim \mathcal{N}(\boldsymbol{t}_i, \alpha_s^{-1}\boldsymbol{I}_M)$ から生成される．

まとめると，提案モデルの生成過程は以下のようになる．

1. 各クラス $m = 1, 2, \cdots, M$ に対して
   (a) 重み行列を生成: $\boldsymbol{w}_m \sim \mathcal{N}(\boldsymbol{0}, \alpha_w^{-1}\boldsymbol{I}_D)$
   (b) 混同行列を生成: $\boldsymbol{c}_m \sim \mathrm{Dir}(\boldsymbol{\beta}_m)$
2. 各ユーザ $u = 1, 2, \cdots, U$ に対して
   (a) 属性ラベルを生成: $y_u \sim \mathrm{Softmax}([\boldsymbol{w}_m^\top \boldsymbol{x}_u]_{m=1}^M)$
   (b) 各サイト $i = 1, 2, \cdots, N$ に対して
      i. サイト依存属性ラベルを生成: $t_{ui} \sim \mathrm{Cat}(\boldsymbol{c}_{y_u})$
3. 各サイト $i = 1, 2, \cdots, N$ に対して
   (a) 属性統計量を生成: $\boldsymbol{s}_i \sim \mathcal{N}(\boldsymbol{t}_i, \alpha_s^{-1}\boldsymbol{I}_M)$.
   ここで，$\boldsymbol{t}_i = \frac{1}{|E_i|}\sum_{u\in E_i} v(t_{ui})$.

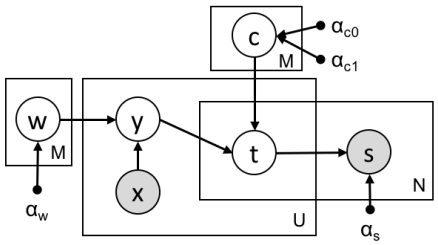

図 2: 提案法のグラフィカルモデル

図 2は，提案モデルのグラフィカルモデルを示す．白ノードは潜在変数，灰色ノードは観測値，小点は超パラメータを表す．超パラメータの集合は $\boldsymbol{\alpha} = \{\alpha_w, \alpha_{c_0}, \alpha_{c_1}, \alpha_s\}$ と表記する．

3.2 変分ベイズによる推論

図 2から分かるように，提案モデルは階層ベイズモデルの一種であるため，変分ベイズ等の近似推論法や MCMC を使うことで効率的な推論が可能になる．本稿では，変分ベイズによる推論法を説明する．

まず，$\boldsymbol{T}, \boldsymbol{Y}$ を周辺化したパラメータ $\boldsymbol{W}, \boldsymbol{C}$ の対数周辺事後確率は，

$$\begin{aligned} &\log p(\boldsymbol{W}, \boldsymbol{C}|\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\alpha}) \\ &\quad\propto \log \sum_{\boldsymbol{T},\boldsymbol{Y}} p(\boldsymbol{S}|\boldsymbol{T}, \alpha_s) p(\boldsymbol{T}|\boldsymbol{Y}, \boldsymbol{C}) p(\boldsymbol{Y}|\boldsymbol{W}, \boldsymbol{X}) \\ &\qquad + \log p(\boldsymbol{W}|\alpha_w) + \log p(\boldsymbol{C}|\alpha_{c_0}, \alpha_{c_1}) \end{aligned} \tag{2}$$

と書ける．提案モデルの生成過程より，

$$p(\boldsymbol{S}|\boldsymbol{T}, \alpha_s) = \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{s}_i|\boldsymbol{t}_i, \alpha_s^{-1}\boldsymbol{I}_M), \tag{3}$$

$$p(\boldsymbol{T}|\boldsymbol{Y}, \boldsymbol{C}) = \prod_{i=1}^{N} \prod_{u\in E_i} c_{y_u, t_{ui}}, \tag{4}$$

$$p(\boldsymbol{Y}|\boldsymbol{W}, \boldsymbol{X}) = \prod_{u=1}^{U} \frac{\exp(\boldsymbol{w}_{y_u}^\top \boldsymbol{x}_u)}{\sum_{m=1} \exp(\boldsymbol{w}_m^\top \boldsymbol{x}_u)}, \tag{5}$$

$$p(\boldsymbol{W}|\alpha_w) = \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{w}_m|\boldsymbol{0}, \alpha_w^{-1}\boldsymbol{I}_D), \tag{6}$$

$$p(\boldsymbol{C}|\alpha_{c_0}, \alpha_{c_1}) = \prod_{m=1}^{M} \frac{\Gamma\left(\sum_{l=1}^{M} \beta_{ml}\right)}{\prod_{l=1}^{M} \Gamma(\beta_{ml})} \prod_{l=1}^{M} c_{ml}^{\beta_{ml}-1}, \tag{7}$$

となる．式 (2) が最大となる $\boldsymbol{W}, \boldsymbol{C}$ を求めることが目的であるが，$\boldsymbol{T}, \boldsymbol{Y}$ のとり得る値の全ての組み合わせを考慮し計算しなければならず，計算量的に困難である．従って，イェンセンの不等式により，式 (2) に対する変分下限 $\mathcal{L}(\Theta)$ を以下のように導出し，これを目的関数として最適化する．

$$\begin{aligned}
&\log p(\boldsymbol{W}, \boldsymbol{C}|\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\alpha}) \\
&\geq \sum_{\boldsymbol{T},\boldsymbol{Y}} q(\boldsymbol{T}, \boldsymbol{Y}) \log \frac{p(\boldsymbol{S}|\boldsymbol{T}, \alpha_s) p(\boldsymbol{T}|\boldsymbol{Y}, \boldsymbol{C}) p(\boldsymbol{Y}|\boldsymbol{W}, \boldsymbol{X})}{q(\boldsymbol{T}, \boldsymbol{Y})} \\
&\quad + \log p(\boldsymbol{W}|\alpha_w) + \log p(\boldsymbol{C}|\alpha_{c_0}, \alpha_{c_1}) \\
&= \mathcal{L}(\Theta). \qquad (8)
\end{aligned}$$

変分分布 $q(\boldsymbol{T}, \boldsymbol{Y})$ は，$q(\boldsymbol{T}, \boldsymbol{Y}) = q(\boldsymbol{Y}|\boldsymbol{\zeta}) q(\boldsymbol{T}|\boldsymbol{\eta})$ の因子分解を仮定する．ここで，

$$q(\boldsymbol{Y}|\boldsymbol{\zeta}) = \prod_{u=1}^{U} \zeta_{u y_u}, \quad q(\boldsymbol{T}|\boldsymbol{\eta}) = \prod_{i=1}^{N} \prod_{u=1}^{U} \eta_{i u t_{ui}}, \qquad (9)$$

である．また，$\Theta = \{\boldsymbol{W}, \boldsymbol{C}, \boldsymbol{\zeta}, \boldsymbol{\eta}\}$ は最適化するパラメータ集合を表す．

次に，$\boldsymbol{W}, \boldsymbol{C}, \boldsymbol{\zeta}, \boldsymbol{\eta}$ の更新則を導出する．

$\boldsymbol{W}$ の更新：$\boldsymbol{W}$ は閉形式で更新できないため，以下のように勾配を計算し，準ニュートン法等の最適化法に使い，更新する．

$$\frac{\partial \mathcal{L}(\Theta)}{\partial \boldsymbol{w}_m} = \sum_{u=1}^{U} \left(\zeta_{um} - \frac{\exp(\boldsymbol{w}_m^\top \boldsymbol{x}_u)}{\sum_{l=1}^{M} \exp(\boldsymbol{w}_l^\top \boldsymbol{x}_u)} \sum_{l=1}^{M} \zeta_{ul} \right) \boldsymbol{x}_u - \alpha_w \boldsymbol{w}_m. \qquad (10)$$

$\boldsymbol{C}$ の更新：$\sum_{l=1}^{M} c_{ml} = 1$ の制約があるので，ラグランジュの未定乗数法によって，以下の更新則を導出する．

$$c_{ml} = \frac{\sum_{i=1}^{N} \sum_{u=1}^{U} \zeta_{um} \eta_{ium} + \beta_{ml} - 1}{\sum_{m'=1}^{M} \sum_{i=1}^{N} \sum_{u=1}^{U} \zeta_{um'} \eta_{ium'} + \sum_{l'=1}^{M} \beta_{ml'} - M}. \qquad (11)$$

$\boldsymbol{\zeta}$ の更新：$\sum_{m=1}^{M} \zeta_{um} = 1$ の制約があるので，ラグランジュの未定乗数法によって，以下の更新則を導出する．

$$\zeta_{um} \propto \exp\left\{ \sum_{i=1}^{N} \sum_{l=1}^{M} \eta_{iul} \log c_{ml} + a_{um} - \log \sum_{m'=1}^{M} \exp(a_{um'}) \right\}. \qquad (12)$$

ここで，$a_{um} = \boldsymbol{x}_u^\top \boldsymbol{w}_m$ である．式 (12) の値を計算した後，$\sum_{m=1}^{M} \zeta_{um} = 1$ となるように，値を正規化する．

$\boldsymbol{\eta}$ の更新：$\sum_{m=1}^{M} \eta_{ium} = 1$ の制約があるので，ラグランジュ未定乗数法で更新する．ただし，閉形式で更新はできないため，η_{ium} と λ_{iu} の勾配を以下のように計算し，準ニュートン法等により交互に更新する．

$$\begin{aligned}
\frac{\partial \mathcal{L}(\Theta)}{\partial \eta_{ium}} &= -\frac{\alpha_s}{|E_i|} \left(s_{im} - \mathbb{E}[\boldsymbol{t}_i]_m \right) + \sum_{l=1}^{M} \zeta_{ul} \log c_{lm} \\
&\quad - \log \eta_{ium} - 1 + \lambda_{iu}, \qquad (13) \\
\frac{\partial \mathcal{L}(\Theta)}{\partial \lambda_{iu}} &= \sum_{m=1}^{M} \eta_{ium} - 1. \qquad (14)
\end{aligned}$$

ここで，$\mathbb{E}[\boldsymbol{t}_i]_m = \frac{1}{|E_i|} \sum_{u \in E_i} \eta_{ium}$ である．

上記のようにパラメータを逐次的に更新し，式 (8) の値が収束したら学習を終了する．超パラメータ $\boldsymbol{\alpha}$ は，クロスバリデーションによって適切な値を求める．

表 1: 人工データによる属性推定精度

	$\alpha_{c_1} = 1$	$\alpha_{c_1} = 10$	$\alpha_{c_1} = 100$
提案法	**0.43**	**0.52**	**0.45**
MTEN [Culotta 15]	0.32	0.51	0.31
Ridge [Culotta 15]	0.24	0.48	0.24

4. 実験

図 1のシチュエーションにおいて，提案法が有効であることを確認するために，人工データを使った実験を行った．

今回は $M = 4$ 種類の属性の分類問題を解く設定とする．人工データは，3.1節のモデルの生成過程で説明した手順に従う．ただし，

1. 個人の特徴量を平均 0，分散 1 の $D = 200$ 次元ガウス分布から生成する．
2. サイトの数を $N = 1,000$，個人の数を $U = 100$ とする．各サイトはランダムに選ばれた個人 30 人からフォローされる．
3. 式 (1) に従って $\boldsymbol{\beta}$ を設定する．ここで，$\alpha_{c_0} = 1$，$\alpha_{c_1} \in \{1, 10, 100\}$ とする．

比較手法として，[Culotta 15] で提案された属性統計量を回帰するモデルに基づいて個人の属性を当てる方法を用いる．この方法では，まず，個人の特徴量からその個人と関係のあるサイトの属性統計量を回帰するモデルを学習する．ここで，属性統計量は一般に多変量になるため，[Culotta 15] は各変量間の関係を捉えられる Multi-Task Elastic Net (MTEN) と，変量間で独立な回帰モデルを構築する Ridge 回帰のケースを考えた．予測の際は，この回帰モデルによって属性統計量を予測し，その値が最も大きい属性をその個人の属性として出力する．

表 1は，α_{c_1} を変えて人工データを生成した際の，個人の属性推定の精度を示す．α_{c_1} は小さい時，元ドメインと目標ドメインの属性統計量の差が大きくなり，元ドメインの属性統計量の値を回帰するようにモデル化した既存手法では，ランダムに予測した場合とほとんど変わらない精度となった．一方，提案法は，その差異を混同行列 $\boldsymbol{C}$ で捉えることができるため，比較的良い精度で属性を推定可能であった．また，α_{c_1} が大きくなった場合でも，提案法の予測精度が最も良かった．

5. おわりに

本稿では，個人の属性情報を教師データとして使わずに，個人の属性を推定する手法を提案した．具体的には，提案法は，別のドメインから比較的簡単に得られる属性統計量に基づいて，個人の属性情報を推定する．その際，提案法は，我々が属性を知りたい個人の存在するドメインと属性統計量のドメインの間に発生する属性統計量の違いを考慮する．

今後の研究では，実データでの実験を行い，提案法の有効性を検証していく．

参考文献

[Cheng 10] Cheng, Z., Caverlee, J., and Lee, K.: You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users, in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 759–768 (2010)

[Culotta 15] Culotta, A., Ravi, N. K., and Cutler, J.: Predicting the Demographics of Twitter Users from Website Traffic Data, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 72–78 (2015)

[Dong 14] Dong, Y., Yang, Y., Tang, J., Yang, Y., and Chawla, N. V.: Inferring User Demographics and Social Strategies in Mobile Social Networks, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 15–24 (2014)

[Kumar 09] Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K.: Attribute and Simile Classifiers for Face Verification, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 365–372 (2009)

[Li 14] Li, J., Ritter, A., and Hovy, E.: Weakly Supervised User Profile Extraction from Twitter, *ACL*, pp. 165–174 (2014)

[Mislove 11] Mislove, A., Lehmann, S., Ahn, Y.-y., Onnela, J.-p., and Rosenquist, J. N.: Understanding the Demographics of Twitter Users, in *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 554–557 (2011)

[Rao 10] Rao, D., Yarowsky, D., Shreevats, A., and Gupta, M.: Classifying Latent User Attributes in Twitter, in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, p. 37 (2010)