半教師有り学習に基づくユーザ属性予測モデルに関する研究

1X18C050-0　竹内 瑞生
指導教員　後藤正幸

1. 研究背景・目的

従来より，ユーザの属性情報を利用して特定層をターゲットするセグメントマーケティングは広く様々なビジネス領域で活用されてきた．昨今においては，インターネット上のサービスに登録されたユーザアカウントに紐付けられた属性情報がマーケティングに活用されることも多い．しかし一般的に，そのようなサービスに登録をして利用する会員ユーザよりも，登録せずに利用する非会員ユーザの方が大多数を占めており，かつマーケティング目的からすれば非会員ユーザの属性情報の方が重要となる．そのため，属性情報がある少数の会員ユーザの閲覧・購買履歴データから，非会員ユーザに対して属性を予測し，属性情報が紐付けられたユーザの数を増やす方法が実務上，価値が高い．その際，属性情報の無いデータは大量に存在することが一般的であることから，少数の教師有りデータと多数の教師無しデータを併用して分類問題を解く半教師有り学習が有効であると考えられる．

一方，近年提案されている半教師有り学習の手法の中にLadder Network[1] がある．この手法は特徴量にノイズを付与し除去する操作を行うニューラルネットワークを用いた手法であり，画像データによる実験では高精度な結果を示している．また，特徴量が高次元になりやすい履歴データに対しても適用可能なため，ユーザ属性の予測モデルで利用するのに適した手法であると考える．しかし，従来の Ladder Network をそのままユーザ属性の予測モデルとする場合，予測するラベルに順序性がある情報が含まれている場合にも正誤のみに基づいて予測してしまう．例えば「30 代」という正解ラベルに対しては，「80 代」よりも「40 代」の方が予測として好ましく，このような順序性を考慮したモデルを構築するべきである．

そこで本研究では，Ladder Network をベースとし，ユーザの属性情報を適切に予測可能な仕組みを組み込んだモデルを提案する．具体的には，学習時にラベルの順序性を考慮可能な損失関数を導入する．最後に，実際の閲覧履歴データを用いてユーザの属性情報を予測し，その精度を従来手法と比較して，提案手法の有効性を示す．

2. Ladder Network

2.1. モデル概要

Ladder Network はニューラルネットワークを用いて分類を行う半教師有り学習の手法である． モデルの全体図を図 1 に示す．テストデータの予測ラベルを出力する Clean Encoder Path(以下，Cl E-path)，ノイズが付与された学習用の予測ラベルを出力する Corrupted Encoder Path(以下，Co E-path)，ノイズを除去する Decoder Path(以下，D-path) の 3 つの要素からモデルが構成される．

この手法では，ノイズの付与と除去を行うことにより，クラスの識別超平面を適切に決定することができる．

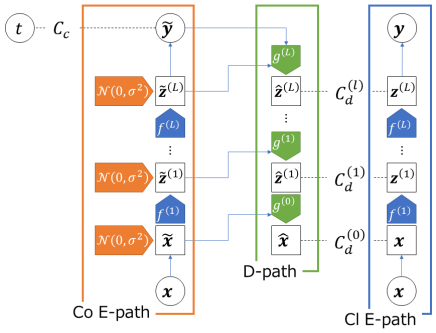

図 1: Ladder Network のモデルの全体図

2.2. 定式化

学習用データ数を M，そのうちの教師有りデータ数を N，n 番目のデータの特徴量を $\boldsymbol{x}_n$ として，教師有りデータの組を $\{\boldsymbol{x}_n, \boldsymbol{t}_n | 1 \leq n \leq N\}$，残りの教師無しデータを $\{\boldsymbol{x}_n | N + 1 \leq n \leq M\}$ とする．ただし，$\boldsymbol{t}_n$ は n 番目の教師有りデータの正解ラベルを表す．

このとき，テストデータの予測ラベル $\boldsymbol{y}$ を出力する Cl E-path における l 層目 $(1 \leq l \leq L)$ の中間表現 $\boldsymbol{z}^{(l)}$ は，直前の中間表現 $\boldsymbol{z}^{(l-1)}$ と学習パラメータである $\mathbf{W}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}^{(l)}$ を用いて非線形写像 $f^{(l)}$ により計算される．ただし，$\boldsymbol{z}^{(0)} = \boldsymbol{x}$ である．

$$\boldsymbol{z}^{(l)} = f^{(l)}(\boldsymbol{z}^{(l-1)}, \mathbf{W}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}^{(l)}) \tag{1}$$

また，ノイズが付与された学習用の予測ラベル $\tilde{\boldsymbol{y}}$ を出力する Co E-path における l 層目 $(1 \leq l \leq L)$ の中間表現 $\tilde{\boldsymbol{z}}^{(l)}$ は，直前の中間表現 $\tilde{\boldsymbol{z}}^{(l-1)}$ と学習パラメータである $\mathbf{W}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}^{(l)}$ に加えて，正規分布に従うノイズ $\boldsymbol{\varepsilon}^{(l)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2)$ を用いて $f^{(l)}$ により計算される．

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{z}}^{(0)} = \boldsymbol{x} + \boldsymbol{\varepsilon}^{(0)} \tag{2}$$

$$\tilde{\boldsymbol{z}}^{(l)} = f^{(l)}(\tilde{\boldsymbol{z}}^{(l-1)}, \mathbf{W}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\gamma}^{(l)}) + \boldsymbol{\varepsilon}^{(l)} \tag{3}$$

D-path における l 層目 $(0 \leq l \leq L)$ の中間表現 $\hat{\boldsymbol{z}}^{(l)}$，特徴量 $\hat{\boldsymbol{x}} = \hat{\boldsymbol{z}}^{(0)}$ は直前の中間表現 $\hat{\boldsymbol{z}}^{(l+1)}$ と学習パラメータ $\mathbf{V}^{(l+1)}, \boldsymbol{a}^{(l)}$ に加えて，同じ層の Co E-path の中間表現 $\tilde{\boldsymbol{z}}^{(l)}$ を用いて非線形写像 $g^{(l)}$ により計算される．ただし，$\hat{\boldsymbol{z}}^{(L+1)} = \tilde{\boldsymbol{y}}$ である．

$$\hat{\boldsymbol{z}}^{(l)} = g^{(l)}(\hat{\boldsymbol{z}}^{(l+1)}, \tilde{\boldsymbol{z}}^{(l)}, \mathbf{V}^{(l+1)}, \boldsymbol{a}^{(l)}) \tag{4}$$

これらの中間表現をもとにして，教師有りデータに対するコスト関数 C_c と教師無しデータに対するコスト関数 C_d から成る以下のコスト関数 C を最小化することでモデルの学習を行う．

$$C = C_c + C_d \tag{5}$$

$$C_c = -\frac{1}{N}\sum_{n=1}^{N} \log P(\tilde{\boldsymbol{y}}_n = \boldsymbol{t}_n | \boldsymbol{x}_n) \tag{6}$$

$$C_d = \sum_{l=0}^{L} \frac{\lambda_l}{N m_l} \sum_{n=1}^{M} \| \boldsymbol{z}_n^{(l)} - \hat{\boldsymbol{z}}_{BN,n}^{(l)} \|^2 \tag{7}$$

ただし, $\hat{\boldsymbol{z}}_{BN,n}^{(l)}$ はバッチごとに標準化した D-path の中間表現, λ_l はハイパーパラメータ, m_l は l 層目の中間表現の次元数である.

3. 提案手法

3.1. 提案への着想

本研究では評価実験として性別と年代のクロスラベル予測を行う. Ladder Network では, 教師有りデータに対するコスト関数 C_c に交差エントロピーを適用している. しかし, 交差エントロピーは年齢等の順序性を無視したコストを計算してしまう.

そこで本研究では, 教師有りデータに対するコスト関数 C_c を性別予測についてのコスト関数 C_s と年代予測についてのコスト関数 C_a に分割し, それぞれに適切な関数を設定して学習する方法を提案する.

3.2. 変数の定義と定式化

教師有りデータに対するコスト関数 C_c を, 性別と年代それぞれのコスト関数 C_s, C_a に分割し, ハイパーパラメータ λ_c を用いて以下の式で定義する.

$$C_c = C_s + \lambda_c C_a \tag{8}$$

また, 長さ 2 のベクトルとして性別についての n 番目 $(1 \leq n \leq N)$ の正解ラベル, 予測ラベルを $\boldsymbol{t}_{s,n}, \tilde{\boldsymbol{y}}_{s,n}$, 長さ K のベクトルとして年代についての n 番目 $(1 \leq n \leq N)$ の正解ラベル, 予測ラベルを $\boldsymbol{t}_{a,n}, \tilde{\boldsymbol{y}}_{a,n}$ とする.

性別についてのコスト関数は, 順序性の考慮が必要ないため, 以下の式により定義される交差エントロピーを用いる.

$$C_s = -\sum_{n=1}^{N} \log P(\tilde{\boldsymbol{y}}_{s,n} = \boldsymbol{t}_{s,n} | \boldsymbol{x}_n) \tag{9}$$

一方, 年代は順序のある属性であるため, そのコスト関数として, $\boldsymbol{t}_{a,n}, \tilde{\boldsymbol{y}}_{a,n}$ の各要素 $t_{a,n}^{(k)}, \tilde{y}_{a,n}^{(k)} (1 \leq k \leq K)$ の二乗誤差と, 正解ラベルとのマンハッタン距離 $d_1^{(k)}$ をもとにした以下の式で定義される関数を用いる.

$$C_a = \sum_{n=1}^{N}\sum_{k=1}^{K} d_1^{(k)} (t_{a,n}^{(k)} - \tilde{y}_{a,n}^{(k)})^2 \tag{10}$$

このコスト関数を用いることで, 不正解の中でもより正解に近いラベルを予測するようになると期待できる.

4. 評価実験

4.1. データ概要

評価実験には, 株式会社ヴァリューズより提供された Web サイトの閲覧履歴データを用いる. データラベルには性別と年代のクロスラベルを用いる. ユーザ数は 45,481 件, Web サイトの種類数は 12,194 件であった.

4.2. 実験条件

45,481 件のデータのうち, 36,000 件を学習用データ, 残りをテストデータとした. 学習用データは 20%にあたる 7,200 件を教師有りデータ, 残りを教師無しデータとした. データラベルに関して, 性別は男性・女性の 2 種, 年代は 10 代–90 代の 9 種とし, クロスラベルの種類数はその組み合わせをとった計 18 種とした.

比較手法として, 従来の Ladder Network と Light GBM[2] を用いる. 提案手法と Ladder Network はバッチサイズ B を $B = 32$ とした. Light GBM は教師有りデータとした 7,200 件のみを学習に用いて予測を行った.

予測結果の評価指標として, Accuracy と F 値のマクロ平均に加えて, 性別と年代のそれぞれの予測性能を評価するために以下の式で定義される評価指標を用いた. それぞれ正解ラベルと予測ラベルとの距離の総和を計算しており, 年代については予測を外した際の順序の近さがこの指標により確認できる.

$$\mathrm{met}_s = \frac{B}{N}\sum_{n=1}^{N} |\operatorname*{argmax}_k t_{s,n}^{(k)} - \operatorname*{argmax}_k \tilde{y}_{s,n}^{(k)}| \tag{11}$$

$$\mathrm{met}_a = \frac{B}{N}\sum_{n=1}^{N} |\operatorname*{argmax}_k t_{a,n}^{(k)} - \operatorname*{argmax}_k \tilde{y}_{a,n}^{(k)}| \tag{12}$$

4.3. 実験結果と考察

ここで, 表 1 に各手法における各評価値の値を示す.

表 1: 実験結果

	Accuracy	F 値のマクロ平均	met_s	met_a
提案手法	0.3558	**0.1757**	6.011	22.47
Ladder Network	0.3582	0.1593	6.099	22.39
Light GBM	0.3537	0.1420	5.863	23.44

表 1 より, 3 手法で Accuracy の値はほぼ同等であるが, F 値のマクロ平均では提案手法が最も優れている. これは従来手法がデータ数の多いラベルに予測が集中した中, 提案手法がデータ数の少ないラベルに対しても予測が適切に行えたことを示しており, 提案手法が有効であると言える. しかし, met_s においては Light GBM が最も良い結果となった. これは性別のラベルの種類数が 2 つと少なかったことにより, 予測がデータ数の多いラベルに集中した場合でも, 予測が外れた際の誤差が小さな値に抑えられたためと考えられる.

5. まとめと今後の課題

本研究では, 半教師有り学習の手法を用いてユーザの履歴データから属性情報を予測するモデルを提案し, 実データを用いた評価実験によりその有効性を示した. 今後の課題として, 精度向上のためのモデルの見直しのほか, 履歴データやデータラベルを変更した評価実験の実施が挙げられる.

参考文献

[1] Rasmus, Antti, et al. "Semi-Supervised Learning with Ladder Networks", *Advances in Neural Information Processing Systems*, vol.28, 2015

[2] Ke, Guolin, et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", *Advances in Neural Information Processing Systems*, vol.30, 2017