3L3-3

情報の独立性を強化したトピックモデル

A topic model whose information-independence is enhanced

神嶌 敏弘 *1　赤穂 昭太郎 *1　佐藤 一誠 *2
Toshihiro Kamishima　Shotaro Akaho　Issei Sato

*1産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

*2東京大学
The University of Tokyo

A topic model is a model for grouping documents based on their constituent words into so-called topics, which are kinds of soft clusters. We discuss a model whose topics so as to be independent of specified sensitive information. In other words, we try to obtain topics that do not contain the sensitive information.

1. はじめに

本稿では，指定したセンシティブ属性とよぶ変数と独立になるようなトピックを抽出するためのトピックモデルについて論じる．このトピックモデルには次のような応用が考えられる．与信などの判断を行う予測モデルが，性別や人種などのセンシティブ情報を排除しつつ予測を行うという公正配慮型分類問題がある．この問題について，センシティブな情報を含まない形式に元データを変換し，その後一般的な分類器で予測する事前処理型の手法が提案されている [Zemel 13]．元データをセンシティブ情報を含まないトピックに，このトピックモデルを用いて変換しておけば，同様の公正配慮型分類が可能となる．また，利用者が不要と判断した情報を利用せずにクラスタリングを行うモデルも提案されている [Gondek 04]．トピックモデルは一種のソフトクラスタリングでもあるため，この手法と同様の文書クラスタを獲得できる．本稿では，このような目的に利用できる独立性強化型トピックモデル (independence-enhanced topic model) について論じる．

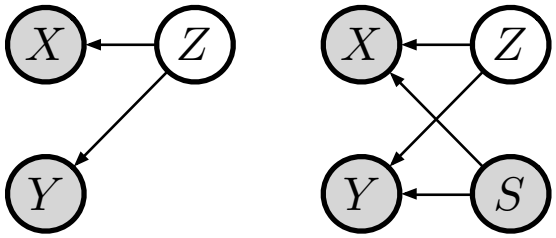
2. 手法

2.1 表記

変数などの表記を行う．まず，トピックモデルにとって基本的な変数について述べる．X は文書を表す確率変数であり，その実現値 $x \in \{1, \ldots, n\}$ は文書のインデックスである．Y は単語を表す確率変数であり，単語の実現値 y も $\{1, \ldots, m\}$ のインデックスで表す．Z はトピックを表す確率変数であり，$z \in \{1, \ldots, K\}$ の実現値をとる．このトピックとは，文書間での複数の単語の共起性によって創発される情報である潜在的意味のカテゴリのことである [佐藤 15]．そして，トピックモデルはこのトピックを文書集合から得るためのものである．

これらのトピックモデルにとって基本的な変数に加えて，本稿ではセンシティブ情報を表すセンシティブ属性 S を導入する．この変数に対し独立となるよう制約の下でトピック変数 Z の分布を定めることが，通常のトピックモデルと本稿の独立性強化型モデルとの違いである．ここでは，このセンシティブ変数 S は簡単のため二値変数としておき，その実現値は $s \in \{0, 1\}$ である．

トピックモデルに対する訓練データは N 個のタプルの集合 D である．各タプルは，通常のトピックモデルでは文書と単語の実現値の対 (x, y) であり，独立性強化型ではそれにセンシティブ変数の実現値を加えた三つ組 (x, y, s) である．この訓練データに対してあてはめることによって，トピックモデルのパラメータを推定する．



(a) pLSA　(b) STI-pLSA

図 1: 二つの pLSA モデルのグラフィカルモデル

2.2 確率的意味分析モデル

独立性強化型のモデルについて論じる前に，その元となる確率的潜在意味分析モデル (probabilistic latent semantic analysis model; pLSA)[Hofmann 99] について述べる．この pLSA モデルの生成モデルをグラフィカルモデルで表すと図 1(a) のようになる．グラフィカルモデルとは，生成モデルの観測変数を白丸で潜在変数を黒丸で表し，これらの変数間の依存関係を有向辺で図示したものである．そして，数式では次のようになる：

$$\Pr[X, Y] = \sum_Z \Pr[X|Z] \Pr[Y|Z] \Pr[Z] \qquad (1)$$

X, Y, および Z はいずれも多値離散確率変数であるため，$\Pr[X|Z]$，$\Pr[Y|Z]$，および $\Pr[Z]$ はいずれもカテゴリ分布に従うとする．確率質量 $\Pr[x|z]$ と $\Pr[y|z]$ はそれぞれ，トピック z に対し文書と単語の関連の強さを表すことになる．すなわち，これらの確率が大きいと，そのトピックに該当する文書であったり，そのトピックでよく使われる単語であったりする．

カテゴリ分布のパラメータ $\{\Pr[z]\}$，$\{\Pr[x|z]\}$，および $\{\Pr[y|z]\}$ は最尤推定により求めるが，実験では 0 頻度問題に対処するため Laprace 平滑化を利用した．この最尤推定は EM アルゴリズムを利用することで解けることが広く知られている．

連絡先: http://www.kamishima.net/

2.3 厳密トピック独立潜在意味分析モデル

前節の pLSA を独立性強化型にしたモデルについて述べる．本稿ではトピック Z と与えられたセンシティブ変数 S との条件なし独立性 $S \perp\!\!\!\perp Z$ を満たすようなモデルについて考察する．トピックに対して厳密な独立性を考えるということで厳密トピック独立 pLSA モデル (strictly topic-independent pLSA model; STI-pLSA) と呼ぶことにする．

この STI-pLSA のグラフィカルモデルによる表示は図 1(b) のようになる．この図から分かるようにトピック Z とセンシティブ属性 S は条件 $S \perp\!\!\!\perp Z$ を満たす．このモデルの同時分布は次式で表される．

$$\Pr[X, Y, S] = \Pr[S] \sum_Z \Pr[X|Z, S] \Pr[Y|Z, S] \Pr[Z] \tag{2}$$

なお，pLSA と同様に，いずれの確率分布もカテゴリ分布に従う．

訓練データ D が与えられたとき，やはり EM アルゴリズムを用いた最尤推定によってパラメータを求めることができる．EM アルゴリズムは E ステップと M ステップを交互に収束するまで繰り返すアルゴリズムであり，各ステップでは次式によりパラメータの更新を行う．

E ステップ：

$$\Pr_{\text{new}}[z|x, y, s] \leftarrow \frac{\Pr[z] \Pr[x|z, s] \Pr[y|z, s]}{\sum_{z'} \Pr[z'] \Pr[x|z', s] \Pr[y|z', s]} \tag{3}$$

M ステップ：

$$\Pr_{\text{new}}[x|z, s] \leftarrow \frac{\sum_{y'} N(x, y', s) \Pr\left[z|x, y', s\right] + 1/n}{\sum_{x'y'} N(x', y', s) \Pr[z|x', y', s] + 1} \tag{4}$$

$$\Pr_{\text{new}}[y|z, s] \leftarrow \frac{\sum_{x'} N(x', y, s) \Pr\left[z|x', y, s\right] + 1/m}{\sum_{x'y'} N(x', y', s) \Pr[z|x', y', s] + 1} \tag{5}$$

$$\Pr_{\text{new}}[z] \leftarrow \frac{\sum_{x',y',s'} N(x', y', s') \Pr\left[z|x', y', s'\right] + 1/K}{N + 1} \tag{6}$$

$$\Pr_{\text{new}}[s] \leftarrow \frac{\sum_{x'y'} N(x', y', s) + 1/2}{N + 1} \tag{7}$$

ただし，$N(x, y, s)$ の表記は，訓練集合 D 中で，$X{=}x \wedge Y{=}y \wedge Z{=}z$ の条件を満たす事例数を表す．また，M ステップでは 0 頻度問題に対応するため Laplace 平滑化を導入している．実験では，最初の E ステップをパラメータが全て 1 の Dirichlet 分布に従う乱数で初期化後，50 回両ステップを反復させて収束させた．

3. 実験

前節の pLSA と STI-pLSA を Reuters-21578 コーパスに適用する実験を行った．コーパス全体で 10 回以上出現している単語を選んだ後，不用語や数字などの除去の作業を行った．その結果，文書数 n は 10786 個，単語数 m は 8210 個となった．$K = 2$ で pLSA を適用し，コーパス中の各文書 x を $\Pr[z|x]$ の値が最大となるトピック z にに分類しセンシティブ情報とする．すなわち，トピック 0 に分類された文書に含まれる単語が観測されたときのセンシティブ属性の値は全て 0 となる．このデータに K=2 で STI-pLSA を適用すると，pLSA で抽出されたトピックとは独立なトピックが得られることが期待される．

モデル (2) では，センシティブ属性 S とトピック Z は独立であり，真にこのモデルからデータが生成されていれば $S \perp\!\!\!\perp Z$ の条件は満たされるだろう．しかし，最尤推定によって得られたモデルは，モデル集合の中で最も訓練データに近い分布を表すだけで，訓練データそのものを表すわけではない．そこで訓練データに対する経験的な S と Z の同時分布を計算し，実際に独立になるかどうかを検証した．

この経験的な分布を求めてみる．トピックはデータを与えたときのトピックの分布として表されるので次式で計算できる：

$$\Pr[z|x, y, s] = \frac{\Pr[x, y, z, s]}{\sum_{z'} \Pr[x, y, z', s]} \tag{8}$$

$$\Pr[x, y, z, s] = \Pr[s] \Pr[z] \Pr[x|z, s] \Pr[y|z, s] \tag{9}$$

(X, Y, S) 上の真の分布を訓練データの平均による経験分布で近似することで (X, Y, Z, S) 上の経験同時分布を得ることができる．

$$\Pr_{\text{emp}}[x, y, z, s] = \frac{1}{T} \sum_{x',y',s'} N(x', y', s') \Pr[z|x, y, s] \tag{10}$$

$$T = \sum_{x',y',s'} N(x', y', s') \tag{11}$$

これを X と Y について周辺化すれば，S と Z の経験的な同時分布を得る．

$$\Pr_{\text{emp}}[z, s] = \sum_{x',y'} \Pr_{\text{emp}}[x, y, z, s] \tag{12}$$

この同時分布について χ 二乗検定を行うと p 値はほぼ 0 となり $S \perp\!\!\!\perp Z$ の独立性は予想に反し棄却されてしまった．以前，公正配慮型の分類問題を扱う問題において，データ分布とモデルによる分布の差であるモデルバイアスが，結果として得られるモデルの独立性を損なう問題について論じた [Kamishima 13]．今回，独立性が強化できなかったのはこのモデルバイアスの影響であると考える．特に今回はトピック数が 2 と非常に小さな次元にまで次元削減をおこなったため，モデルバイアスの影響は大きいことも影響しているであろう．

4. まとめ

本論文では，独立性を強化するトピックモデルについて考察した．モデル上では独立でも，実際のデータについてはモデルバイアスの影響が大きく独立性を強化したトピックが得られなかった．今後は，[Kamishima 13] で論じたような経験的な分布に対して数値最適化を用いて独立性を強化する方法などを考慮する必要があると考える．

謝辞：本研究は JSPS 科研費 16700157，21500154，および 24500194 の助成を受けた．

参考文献

[Gondek 04] Gondek, D. and Hofmann, T.: Non-Redundant Data Clustering, in *Proc. of the 4th IEEE Int'l Conf. on Data Mining*, pp. 75–82 (2004)

[Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Analysis, in *Uncertainty in Artificial Intelligence 15*, pp. 289–296 (1999)

[Kamishima 13] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: The Independence of the Fairness-aware Classifiers, in *Proc. of the 4th IEEE Int'l Workshop on Privacy Aspects of Data Mining*, pp. 849–858 (2013)

[佐藤 15] 佐藤 一誠：トピックモデルによる統計的潜在意味分析, 自然言語処理, 第 8 巻, コロナ社 (2015)

[Zemel 13] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C.: Learning Fair Representations, in *Proc. of the 30th Int'l Conf. on Machine Learning* (2013)