F-009

糖尿病重症化予防のためのトピックモデルを用いた時系列ヘルスケアデータからのリスク評価

Risk evaluation of severe diabetes by topic model using time series healthcare data

永田 雅俊† 松本 一則† 橋本 真幸†
Masatoshi Nagata Kazunori Matsumoto Masayuki Hashimoto

1. はじめに

近年の診療報酬明細(レセプト)電子化に伴い、多くの健保組合で健診とレセプトデータを解析できる環境になっており、これらを用いた疾病の発症予測や医療費全体の削減が求められている。中でも糖尿病は、重症化して糖尿病腎症を経て人工透析となった場合に高額な医療費につながるため、糖尿病に罹患している患者の適切な指導や治療が重要である。

傷病発症の予測には、これまでロジスティック回帰や Cox 比例ハザードモデルによる手法が多く行われてきたが、近年、機械学習の一つである潜在的ディレクレ配分法(LDA)[1]を利用した先行研究[2][3]では、LDA に健診だけでなく問診やレセプトも用いることで精度が向上することが示唆されていた。しかしながら、発症者が少ない傷病に罹患する人を予測するような場合、正解データが少ないため、機械学習による発症の予測モデル作成が困難になるという問題があった。また、レセプトに現れる傷病名や医薬品名には、別名称でも機能的に同等なものや、予測対象としている傷病とはほとんど関係ない単語も含まれている。一般に特徴選択の観点からは、特徴量の中から有用なものを選び出し、次元数を抑えることで精度の向上が見込まれる[4]ため、ヘルスケアデータにおいても発症に関連するレセプトの特徴量セットを用いることで、予測精度を向上できると期待される。そこで本稿では、医療分野の実際のヘルスケアデータに対し特徴選択を適用することで、LDA による発症予測を行ったときの特性を評価した。

2. 手法

2.1 LDA

LDA は確率的トピックモデルとして様々な分野で応用されている。トピックモデルにおいては、文書は潜在的トピックの集合とみなし、そのトピックは単語の分布の集まりとみなされる。LDA のグラフィカルモデルを図 1 に示す。α, β はハイパーパラメータで、確率分布 θ に従いトピック Z が選択され、単語はトピックと確率分布 ϕ に従って生成される。実験ではトピック数を 10 から 100 までとし、確率分布 θ, ϕ は Collapsed 変分ベイズから推定した。

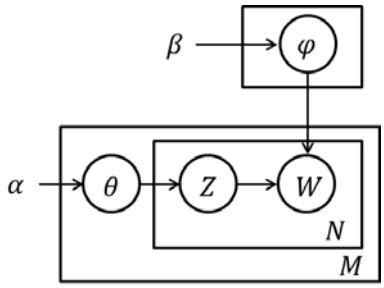

図 1 LDA のグラフィカルモデル

予測対象とする疾病は、人数の異なるデータセットとして糖尿病と糖尿病性腎症に関するデータセットを用いた。LDA に用いる単語(Bag-of-Words) は健診、問診、レセプトに使用されている項目を集め特徴量とした。健診の個々の測定値は、正規化した値の分布から標準偏差を基準に低中高に分類した。

2.2 特徴選択

特徴選択にあたっては、単語 w の有無と発症者の人数から表 1 に示すようなクロス集計表を作成し、文献[5]を参考に独立モデル(IM)と従属モデル(DM)の AIC（赤池情報量規準）をそれぞれ下記式で算出した。実験には、*AIC_IM(w)* と *AIC_DM(w)*の差が 2 以上のものを発症に関連するとみなして採用した。

また、単語間の組み合わせによる特徴量生成を行い、同様に独立/従属 AIC による発症の関連性を調べた。AIC の差が 2 以上となった単語の組み合わせを持つ場合に、新たに単語として追加した。表 2、表 3 に糖尿病性腎症予測に用いたデータセットの単語について、AIC を計算した一部を示す。

$$AIC\_IM(w) = -2 \times MLL\_IM + 2 \times 2$$
$$MLL\_DM = Ne(w)\log Np(w) + N(w)\log N(w) + Nn(w)\log Nn(w) + N(\neg s)\log N(\neg s) - 2N\log N$$
$$AIC\_DM(w) = -2 \times MLL\_DM + 2 \times 3$$
$$MLL\_DM = N_{11}(w)\log N_{11}(w) + N_{12}(w)\log N_{12}(w) + N_{21}(w)\log N_{21}(w) + N_{22}(w)\log N_{22}(w) - N\log N$$

表 1 発症者と未発症者の単語 w をもつ人数

	発症者	未発症者	合計
単語 w あり	$N_{11}(w)$	$N_{12}(w)$	Ne
単語 w なし	$N_{21}(w)$	$N_{22}(w)$	Nn
合計	N(w)	N(¬w)	N

† （株）KDDI 研究所 , KDDI R&D Laboratories, Inc.

Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 2　単語と発症の関連性（1 項目）

単語名称	AIC_IM-AIC_DM
高_HbA1C	229
高_空腹時血糖	195
中_HbA1C	125
中_空腹時血糖	122
経口血糖降下薬	84

表 3　単語と発症の関連性（2 項目）

単語名称 1	単語名称 2	AIC_IM–AIC_DM
高_HbA1C	高_空腹時血糖	211
高_HbA1C	中_γ.GTP	158
高_HbA1C	中_身長	153
高_HbA1C	中_年齢	141
高_空腹時血糖	中_γ.GTP	137

3. 実験，結果

LDA の結果得られた θ からリスク者を多く分類できるクラスタ数をモデルに採用した。訓練データとテストデータは 3-fold cross validation により分割して実験を行った。

実験では、健診・問診・レセプトのデータ項目すべてを使う場合（特徴選択なし）、発症に関する独立と従属 AIC の差が 2 以上のものを使う場合（特徴選択 1）、さらにそれらの単語の組み合わせも使用する場合（特徴選択 2）に分けて比較した。

3.1 実験条件

2011 年から 2015 年で 4 年以上の健診・レセプトデータがある匿名化されたデータを使用した。対象者は 30 歳から 64 歳までの男女とした。発症判定期間は 3 年間とし、発症判定は糖尿病予測の場合は健診結果を、糖尿病性腎症の場合はレセプトで診断された傷病名を用いた。初年度の時点で健診が糖尿病基準に達している人、レセプトに当該傷病名がある人、および関連する服薬のある人は除外した。糖尿病発症に用いるデータセットでは、さらに初年度で特定保健指導基準にないことを条件とし、全体人数 3394 人（うち 22 人がその後 3 年間発症者）のデータセットを用いた。また、糖尿病性腎症では全体人数 8238 人（うち 94 人がその後 3 年間の発症者）のデータセットを用いた。

3.2 結果

糖尿病発症予測に用いるデータセットにおいて、発症に関与することが想定された単語に絞り込んだために、824 種類から 24 種類となった。単語の組み合わせも利用した場合では、使用する単語の種類は 1165 種類となった。糖尿病性腎症予測に用いるデータセットでは、特徴選択によりもともとの単語 1402 種類から 79 種類となり、単語の組み合わせも利用した場合には 4988 種類となった。

特徴選択の有無による予測性能の比較を表 4、表 5 に示す。糖尿病予測に用いたデータセットにおいて、特徴選択 1 では特徴選択なしの場合とほとんど同じだが、特徴選択 2 では Precision が上がり F 値がやや向上した。一方、糖尿病性腎症予測の場合では、Recall が上がったが F 値には大きな違いは認められなかった。これは糖尿病性腎症に用いたデータでは、糖尿病発症のデータに比べて全体的に AIC が高く、用いられている単語からの予測が困難であるためと考えられる。

表 4　特徴選択の有無による糖尿病の予測性能の比較

手法	Recall	Precision	F 値
特徴選択なし	0.59	0.45	0.51
特徴選択 1	0.55	0.48	0.51
特徴選択 2	0.50	0.63	0.56

表 5　特徴選択の有無による糖尿病性腎症の予測性能の比較

手法	Recall	Precision	F 値
特徴選択なし	0.32	0.23	0.26
特徴選択 1	0.55	0.18	0.27
特徴選択 2	0.60	0.17	0.27

4. おわりに

本稿では発症者の少ない傷病を対象として、ヘルスケアデータに特徴選択を導入することで、LDA による予測性能の向上がみられるか検証した。実験の結果から、健診やレセプトに現れる単語の組み合わせを新たに単語とみなして LDA に用いることで予測性能を向上できる見込みを得た。一方、糖尿病腎症予測に関するデータセットでは大きな性能向上はみられず、予測力の低い特徴量の場合には効果は薄いと考えられる。

医療分野においては、症例が少ないために機械学習にはデータ数が不足することも多い。単純に発症と関連の少ない特徴量を削減して、関連の高い特徴量に注目するだけでは予測モデルの向上は難しいが、組み合わせを含めて予測力の高い特徴量を増やすことで、ヘルスケアデータにおいても予測モデルの向上ができると考えられる。

参考文献

[1] Blei DM, et al. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research,* 3:993-1022, (2003)

[2] Ogawa K, et al. “Method of Screening the Health of Persons with High Risk for Potential Lifestyle-related Diseases using LDA.” *HEALTHINF*, 502-507, (2015)

[3] 畠山 豊, 宮野伊知郎, 片岡 浩巳, 中島 典昭, 渡部 輝明, 奥原 義保, “問診データに対する潜在トピックモデルに基づく健診データ解析” 医療情報学, 33(5):267-277, (2013)

[4] Liu H. and Motoda H. “Feature Selection for Knowledge Discovery and Data Mining” Kluwer Academic Publishers Norwell, (1998)

[5] 鈴木義一郎，情報量基準による統計解析入門，（株）講談社サイエンティフィク（編），pp. 80−96，（株）講談社，東京，1995

Copyright © 2016 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.