傾向スコアを用いた除籍資料予測の試み

Intro. Method Result Discus.

橋本 郷史
東邦大学医学メディアセンター 大橋病院図書室

2022年7月16日 MIS 37@ONLINE

これまでの経緯1

【きっかけになった業務上の問題意識】

除籍資料の選定には多くの労力がかかるのでそれを軽減したい。

- 除籍対象の選出は勘や経験頼りではなく，その根拠となるデータ的な要素があるはず。
- その要素を蔵書データから抽出できれば，除籍対象をある程度機械的に予測できるのでは？

(学術的な研究ではなく，業務上の問題を解決するための研究。)

これまでの経緯2

- 2021年12月のJMLAの学術集会で，**蔵書データに含まれる資料の要素情報から，各要素と除籍のされやすさの数値的な関係がわかることを報告した**。

除籍に関係する要素

リスク増：改訂版，複本，基礎医学，薬学，
リスク減：眼科・耳鼻科 ……など

※どのような要素が除籍とどのように関係するかは機関ごとに異なる。ここでの結果はあくまで本学での結果。

今回の目的と結果

Intro.

Result

- 今回の研究では，蔵書データのそれらの要素(変数)を用いて実際に資料の除籍の予測が可能かどうかを確認した。
- 端的には次の「結果」が得られた。

過去の除籍実績から未来の除籍を予測し，

- ❑ **"除籍対象とならない資料"の予測は，的中率98%だった。**
- ❑ **"除籍すべき資料"の予測は，的中率9%だった。**

以下，方法・結果(詳細)・考察・結論を述べる。

今回の分析の対象データ

Method

- 東邦大学医学メディアセンター本館資料
- 購入した資料　※備品および消耗品
- 図書　※雑誌や視聴覚資料は含まない
- 2011年4月1日-2021年3月31日に登録した資料
  (かつ2011年以降に出版された資料)

【なにを除籍としてカウントするか】

- 「不要」と判断して除籍したものをカウント。
- 紛失・汚損等で除籍したものは，今回の除籍計算の対象には含めない。

分析のSTEP1-4：簡単版全体像

Method

過去(2020年度末まで)の除籍実績から，
未来(2021年度)の除籍を予測し，
その予測がどの程度あっていたかを確認する。

1. 2020年度末までの除籍実績を元にして，資料の**各要素と除籍との関係**を数値化する。
2. 1. の結果を用いて，**各資料の除籍のされやすさ**を数値化(=**傾向スコア**)する。
3. 傾向スコアと2020年度末までの除籍実績とを比べ，予測に使う**カットオフ値**を決める。
4. このカットオフ値から2021年度の除籍がどの程度予測できたか(**的中率**)を確認する。

STEP1:ロジスティック回帰分析

Method

2020年度末までの除籍実績を元にして，資料の要素と除籍との関係をオッズ比で表現。除籍有無(1/0)を目的変数，以下の要素を説明変数として用いる。

生存時間：登録から除籍まで，在籍の場合は登録から2021年3月31日までの日数を年換算で。
書誌情報：「改訂版である」「本学教職員の著作である」「複数人による著作である」「NDC491・496・499番台の資料である（それぞれ）」
所蔵情報：「複本である」

STEP2:傾向スコア算出

Method

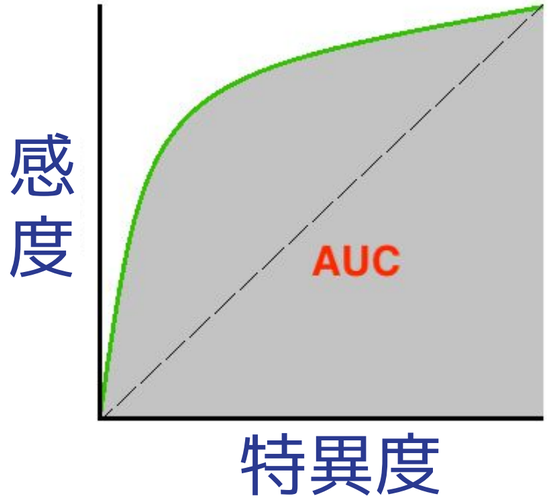
- ロジスティック回帰分析の結果を用いて，各資料の除籍のされやすさ=傾向スコアを算出。
- 0-1で変動。1に近いほうが可能性が大きくなる。
- このスコアで各資料の除籍傾向を数量として比較可能になる。

STEP3:ROC曲線

Method

傾向スコアと2020年度の除籍実績との関係をROC曲線で調べ，予測に使うカットオフ値を決める。

	除籍	在籍	計
カットオフ値以上	a	b	a+b
カットオフ値未満	c	d	c+d
計	a+c	b+d	a+b+c+d



- 感度=a/(a+c):除籍の内，カットオフ値以上の割合
- 特異度=d/(b+d):在籍の内，カットオフ値未満の割合
- ROC曲線:カットオフ値と感度・特異度の変動を図で表現

STEP4:陽性適中率・陰性適中率

Method

カットオフ値から2021年度の除籍が予測できるかを確認する=スコアからの予測と，実際の結果と比較。

	除籍	在籍	計
カットオフ値以上	a	b	a+b
カットオフ値未満	c	d	c+d
計	a+c	b+d	a+b+c+d

- 陽性的中率=a/(a+b):カットオフ値以上で除籍の割合
- 陰性的中率=d/(c+d):カットオフ値未満で在籍の割合

分析結果概要

Result

- 対象件数：14,442件
  - 2020年度末までに除籍：1,611件
  - 2021年度に除籍：524件

STEP1:ロジスティック回帰分析の結果

Result

要素		オッズ比 (95%信頼区間)
生存時間(1年)		0.96 (0.94-0.98)
改訂版である		4.05 (3.62-4.53)
複本である		2.65 (2.32-3.03)
東邦教職員著作		0.58 (0.46-0.72)
複数人著者		1.54 (1.28-1.84)
分類	491(基礎医学)	0.98 (0.82-1.17)
	496(眼科・耳鼻)	0.68 (0.48-0.98)
	499(薬学)	3.41 (2.31-5.03)

このモデルで多変量解析した際の各変数(要素)の除籍への影響度がオッズ比で算出されたもの。

1より大きければ除籍されやすく，1より小さければ除籍されにくくなる。

STEP2:傾向スコア算出の例

Result

- 「3年経過」「改訂版」「複本がある」「東邦著者でない」「複数人による著作」「NDC491・496・499のどれでもない」

  → **傾向スコア：0.411**

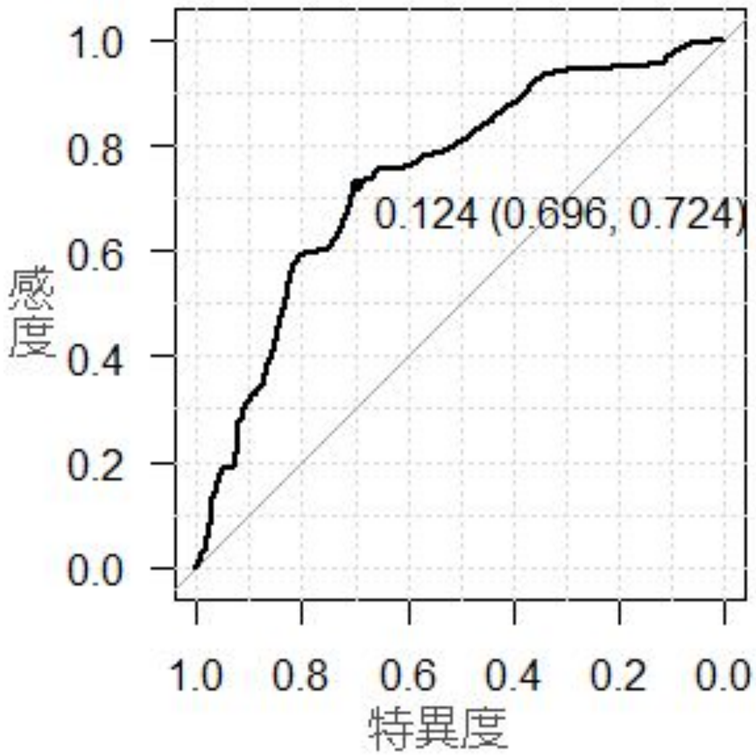
- 「1年経過」「改訂版でない」「複本ない」「東邦著者」「単著」「NDC496」

  → **傾向スコア：0.017**

STEP3:ROC曲線でカットオフ値を探る

Result

- 曲線下面積 0.741(95%CI 0.729-0.754)
- 「感度+特異度」が最も大きくなるのは，傾向スコア=0.124の時。



傾向スコア	2020年度末の状態		計
	除籍	在籍	
0.124以上	1,167	3,910	5,077
0.124未満	444	8,921	9,365
計	1,611	12,831	14,442
	72.4%	69.5%	
	↑感度	↑特異度	

STEP4:2021年度除籍予測の的中率

Result

- 算出したカットオフ値(=傾向スコア0.124)と2021年度の除籍との一致具合を確認したものが下表。

傾向スコア	2021年度末の状態		計		
	除籍	在籍			
0.124以上	335	3,575	3,910	8.6%	←陽性的中率
0.124未満	189	8,732	8,921	97.9%	←陰性的中率
計	524	12,307	12,831		

この結果をどう解釈するか1

Discus.

- 実際の除籍作業時の判定に感覚的には類似している。
- 10年程度の追跡だと年数の除籍影響が見えない。
- 資料の平均的な所蔵期間はもっと長いので，長期所蔵時の年数の持つ除籍へのインパクトは不明。

➔ **あくまで比較的若い資料に限定した傾向として見る。**
➔ **一方で，比較的若い資料の除籍選定こそ，単純に年数で判断できないので，こういった補足的な情報が役に立つ可能性がある。**

この結果をどう解釈する2

Discus.

- 除籍する資料の予測はあてにならない(9%)。
- 除籍対象外の予測精度はそれなり(98%)で，資料の7割が対象となる。
- 除籍対象外の中にも，そこそこの量(除籍全体の1/3)の除籍すべき資料が含まれる。

➔ 除籍をするしないの判断は人が下す必要がある。
➔ が，除籍作業の中にこういった機械判定をうまく組み込むことで，確認対象を絞り込むなど，除籍作業の労力を抑えられる可能性がある。

限界・適用可能性

Discus.

- 分析範囲が10年では短い（が，拡大も難しい）
- より精度の高い結果につながる変数選択・処理について検討の余地がある。
- 今回の結果は本学特有のもので，その数値的な結果に普遍性はない。(内的妥当性はある)

❑ 手法自体は他の機関にも適用可能。他の機関で実行すれば各機関の事情にそった結果が得られる。
(過去の除籍作業の偏りチェックなどにも使える)

❑ 機械的に計算できるので，システムに組み込めばワンクリックで計算することも。

注意点

Discus.

- 過学習・オーバーフィッティング

  説明変数を増やせば学習データ内での精度は上がるが，実データでの汎用性が失われる可能性がある。(今回のモデルは検証データとの比較でモデル精度の差が大きくないため大きな問題はなさそう。)

- モデルと現実の反転

  モデルだけに頼って作業を繰り返すと，モデルが現実を”説明する”のではなく，モデルが現実を”決める”ようになってしまう。(要:人的チェック)

結論：やってみる価値はありますぜ

Con.

- 単純な予測としては芳しい結果ではない
- 使いようによっては業務改善に活かせる
- 分析それ自体として面白かった(個人の感想です)

このようなデータ分析に興味のある方はご連絡ください。

ご清聴ありがとうございます。

ご質問，ご意見，

お待ちしております。

ツールなど

猫：とほにゃん

声：音読さん

動画：DaVinci Resolve

音楽：DOVA-SYNDROME

YouTube Audio Library