

ランダムフォレストに基づく薬剤師国家試験 合否予測モデルの構築とその活用法に関する検討

松延 千春[†]・井上 寛*
窪田 敏夫[◆]・白谷 智宣[†]

1. はじめに

薬剤師養成を目的とする6年制薬学部・薬学科に在籍している学生は、薬剤師国家試験に合格することが在学中の大きな目標となっている。第一薬科大学薬学部薬学教育推進センター（以下、推進センター）では、6年制薬学部
に在籍する学生の学習サポートに役立つ薬学教育コンテンツの作成・運用
について様々な研究をしている。推進センターにおける研究課題の1つに上
級学年学生の知識習得状況の把握があり、これまでに機械学習アルゴリズム
の一つであるサポートベクターマシン (Support Vector Machine, SVM)¹⁻⁴⁾
を用いた薬剤師国家試験合否予測モデルの構築に取り組み、その予測モデル
の合否予測精度が高いことを明らかにしている⁵⁾。また、合否予測結果を知
識習得状況把握のための指標とすることにより、個々の学生に対する適切な
フィードバックができる可能性についても示唆している⁶⁾。

しかし、SVMを用いた合否予測では、予測に重要な説明変数を算出する
ことができない。もし、合否予測に重要な説明変数を把握することができ
れば、どの時期にどの分野の知識習得を重点的に進めていけばいいのかを示
すことができ、教育的介入を実施する際により適切な学習プランを組み立てる

ことが可能になるのではないかと考えた。

分類予測の機械学習で用いられるアルゴリズムの1つであるランダムフォレスト (Random Forest; RF)⁷⁾はアンサンブル法の一つであり、様々なデータに対して有効であることが多くの文献から示されている⁸⁾。アンサンブル法とは、複数の機械学習モデルを組み合わせてより強力な学習モデルを構築しようとする手法である。RFは、学習データから無作為に作った様々な学習データの部分集合を用い決定木分類を行い、予測するときにはすべての決定木の予測を集め、その結果の平均を取り全体の予測をする決定木のアンサンブル法である。決定木分類の最大の問題点は、学習データに対して過剰に適合し予測の精度を悪くしてしまうことである。この問題点を少しずつ異なる決定木を集めアンサンブル学習することで解消している。また、RFの利点として数値計算が高速であることが挙げられる。さらに、SVMでは算出することができない予測モデル構築に使用した説明変数の重要度を Gini 係数としてモデル構築と同時に算出できる。

そこで本研究では、RFを用いた薬剤師国家試験合否予測モデルを新たに構築し、SVMによる予測結果と比較することでRF予測モデルの精度を確認するとともに、同時に算出される説明変数の Gini 係数から薬学教育におけるRFの活用法について検討した。

2. 既存の要求分析と優先順位の決定のための方法論

本研究では、平成27、28及び29年度に6年生として在籍している学生のうち、平成27、28及び29年度9月、11月及び1月に実施した模擬試験(薬学ゼミナール全国統一模試：第228回～第236回)を受験した学生の成績データ及びその学生らがその年に受験した薬剤師国家試験(第101、102及び103回)の合否結果を解析に用いた。薬学ゼミナール全国統一模試の出題構成及び、各時期の解析対象者数はそれぞれ表1、表2に示す。なお、今回解析に使用したデータのうち未受験による欠損データは解析対象から除外している。

予測モデルの説明変数には出題科目(①物理、②化学、③生物、④衛生、⑤薬理、⑥薬剤、⑦病態・薬物治療、⑧法規・倫理・制度、⑨実務)の各得

表1 薬学ゼミナール全国統一模試の出題構成及び問題数

科 目	必須問題	薬学理論問題	薬学実践問題		計
			複合問題	単 問	
物 理	5	10	5	－	20
化 学	5	10	5	－	20
生 物	5	10	5	－	20
衛 生	10	20	10	－	40
薬 理	15	15	10	－	40
薬 剤	15	15	10	－	40
病態・薬物治療	15	15	10	－	40
法規・制度・倫理	10	10	10	－	30
実 務	10	－	65	20	95

表2 各模擬試験における受験者数（上段）および解析対象者数（下段）

		9月模擬試験	11月模擬試験	1月模擬試験
平成27年度	受験者数	130名	130名	106名
	解析対象者数	123名	128名	96名
平成28年度	受験者数	101名	114名	113名
	解析対象者数	99名	114名	113名
平成29年度	受験者数	98名	113名	114名
	解析対象者数	95名	113名	76名

点（1問1点で採点した点数）を、目的変数には薬剤師国家試験の合否結果を用いて、平成27年度モデルと平成28年度モデルをRF及びSVMでそれぞれ構築した（表3）。なお、両年度モデルとも9月、11月、1月それぞれの時点における予測モデルを構築している。

構築した予測モデルの評価は、平成27年度モデルについては平成28年度6年生および平成29年度6年生、平成28年度モデルについては平成29年度6年生が受験した模擬試験（9月、11月及び1月実施）の成績データについて乱数シードを変えて100回予測し、実際の薬剤師国家試験合否結果と照らし合わせて100回中何回誤った判定をしたか（誤判別率）を算出し、誤判別率の平均値を比較することで予測精度を評価した。平成27年度モデルを用いて平

表3 予測モデル構築に使用した模擬試験回および薬剤師国家試験回

実施年度	9月模擬試験	11月模擬試験	1月模擬試験	薬剤師国家試験回
平成27年度	第228回 全国统一模試Ⅰ (9月24, 25日実施)	第229回 全国统一模試Ⅱ (11月26, 27日実施)	第230回 全国统一模試Ⅲ (1月28, 29日実施)	第101回 (2月27, 28日実施)
平成28年度	第231回 全国统一模試Ⅰ (9月29, 30日実施)	第232回 全国统一模試Ⅱ (11月24, 25日実施)	第233回 全国统一模試Ⅲ (1月26, 27日実施)	第102回 (2月25, 26日実施)
平成29年度	第234回 全国统一模試Ⅰ (9月21, 22日実施)	第235回 全国统一模試Ⅱ (11月16, 17日実施)	第236回 全国统一模試Ⅲ (1月25, 26日実施)	第103回 (2月24, 25日実施)

成28年度6年生の予測をした場合の概略を図1に示す。また、RFについては出題科目ごとのGini係数を算出し、各モデルについて合否予測に重要な出題科目を抽出した。

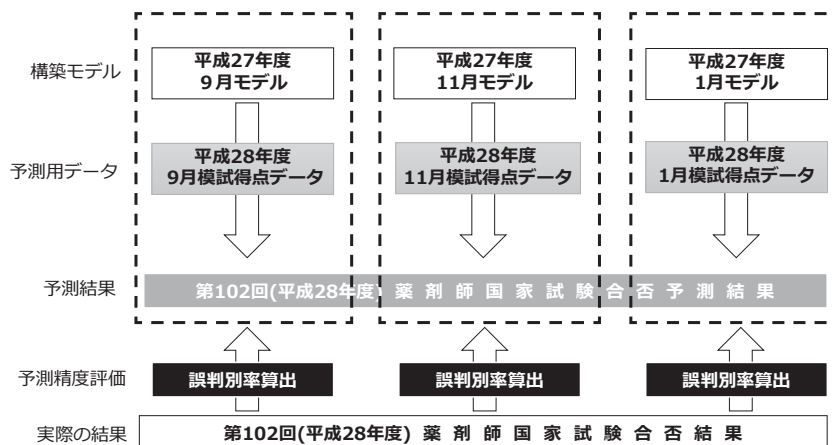


図1 構築した予測モデルの予測精度評価方法概略図
平成27年度モデルを用いて平成28年度6年生の薬剤師国家試験合否予測をした場合

今回の解析では統計ソフトRを使用しており、解析に使用したデータは学籍番号と氏名が削除され個人の特定ができない状態に加工されたものを使用している。また、本研究は本学の倫理委員会の承認(第一薬科大学 18008)

を得て、第一薬科大学個人情報保護方針に準じて実施した。

3. 結 果

今回、平成27年度モデルについては異なる年度(平成28年度と平成29年度)の予測用データで解析を行った。その結果、RFを用いた予測はいずれも10.0%~20.0%の平均誤判別率を維持していた(図2)。

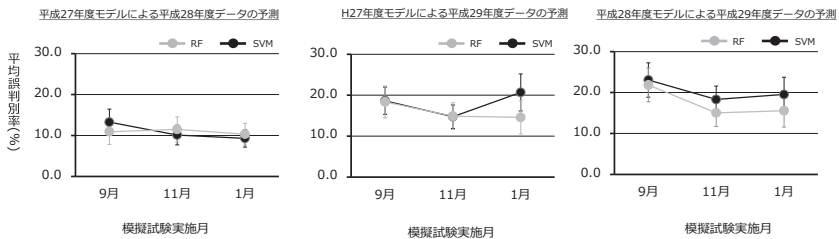


図2 SVMとRFの平均誤判別率(予測精度)比較

また、SVMによる予測と比べても同程度かそれ以下の平均誤判別率であった。さらに、予測用データとして平成29年度データを用いた解析結果から平成27年度モデルと平成28年度モデルの平均誤判別率を比較したが、RFによる予測では20%前後の平均誤判別率を維持しており、モデルの違いにより極端に高い誤判別率を示すこともなかった。

次に、平成27年度モデル、平成28年度モデルそれぞれについて模擬試験実施月ごとに算出された出題科目の Gini 係数結果を図3に示す。

これより、平成27年度モデル、平成28年度モデルの両方において、薬理、薬剤及び実務の Gini 係数が常に上位に位置している一方で、物理および法規・倫理・制度は常に下位に位置している傾向が把握できた。

4. 考 察

先行研究である清水らの報告は、SVMによる平成27年度9月、11月1月モデルを用いた薬剤師国家試験合否予測解析結果にとどまっている^{5,6)}。本研

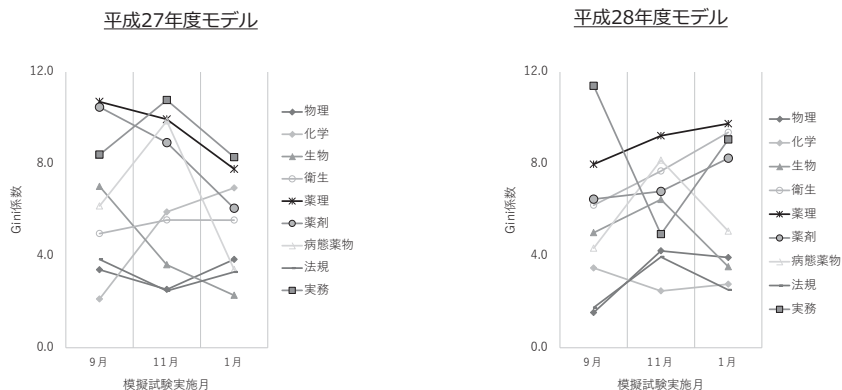


図3 平成27年度モデル，平成28年度モデルにおける模擬試験実施月ごとの出題領域 Gini 係数

究ではSVMに加えてRFによるモデル構築を行い，さらに平成27年度モデルだけでなく平成28年度モデルを新たに構築し，アルゴリズムの違いとモデルを構成するデータの違いにより予測精度がどの程度変動するかを評価した。

今回新たに提案したRFに基づく薬剤師国家試験合否予測モデルの平均誤判別率は，SVMに基づく合否予測モデルと遜色なかったことから，RFでもSVMと同等の精度で予測が可能であることが示された。また，モデルのデータ構成が異なっても極端に高い誤判別率を示すこともなかったことから安定した予測モデルを構築できたといえる。このことから，RFによる合否予測モデルを用いた解析結果についても先行研究⁶⁾と同様に学生の知識習得状況の指標として利用することが可能であると考えられる。

また，RFで算出されるGini係数についても経時的な比較をすることで，ある時期までにどの出題科目の知識習得を高めておく必要があるかという指標として活用できると考えられる。例えば，本研究の解析結果からGini係数が常に上位であった薬理，薬剤及び実務については，6年生9月までに知識習得を目指すような介入を計画することが可能となり，漫然とした学習計画を見直すことができる。このようにある時点までの目標を明確にできれば，学生の主体的な学習の手助けになることが期待できる。また，それぞれの時期に個々の学生に見合った教育コンテンツを提供するための根拠としても利用可能になると考えられる。

今回は6年生を対象として予測モデルを構築したため、この予測結果を基に学習計画の介入を講じてもその成果について6年次にならないと評価することができない。しかしながら、1～6年次の各学年で実施される様々な試験についてRFを用いた分類予測を適用できれば、6年間におよぶ薬学教育の各ポイントで個々の学生の知識習得状況を早期に把握できることになる。そのため、1～6年次の様々なポイントでRFを用いた予測を行い知識習得度別に学生を抽出し、それぞれのグループの知識習得状況の推移や各時期で重点的に学習すべき科目の傾向をつかむことができれば、入学から薬剤師国家試験に合格するまでの具体的な知識習得ロードマップを学生の特徴別に作成できると考えている。以上より、RFによる分類予測は薬学教育において有効に活用できるツールになるといえる。

一方で、本研究は第一薬科大学に所属する学生の成績データを用いた解析のため一般的な結果とは言いがたい。日本の薬学教育において予測モデルを利用した学習方略の開発について報告している事例は少ないが、弓削田らは多重ロジスティック回帰分析を用いた国家試験合否に影響する科目の分析を行っており、その結果から衛生、薬理、薬剤、病態・薬物治療、実務が有意に影響する可能性を示している⁹⁾。この結果は我々の研究結果と類似している部分もあるため、今回の解析結果についてある程度の一般性は保たれていると解釈できる。可能であれば6年制薬学部・薬学科を設置している大学間で情報を共有し解析を行い一般化できることが望ましい。また、本研究の解析で利用した機械学習の分野では、近年ディープラーニングを筆頭に様々な手法が提案されている。そのため、今後も予測精度を向上させるために様々な手法の比較を行っていく必要があると考えている。しかしながら、今回取り扱った学生の成績データはサンプル数に限りがあり、ビッグデータを使う大規模学習で予測精度を向上させるような手法は不向きである。より予測精度の高い解析法を構築する上で、この“サンプル数に限りがある”という問題を解消できる手法を検討していく予定である。

【所属】

- † : 第一薬科大学 薬学教育推進センター
- ◆ : 第一薬科大学 地域医療薬学センター
- * : 九州産業大学 経済学部経済学科

【参考文献】

- 1) Vapnik V. Statistical Learning Theory. New York: John Wiley & Sons.; 1998.
- 2) Vapnik V. The Nature of Statistical Learning Theory. New York: Springer.; 1996.
- 3) Wahba G, Lin Y, Zhang H. GACV for support vector machines, in A. Smola P, Bartlett B, Schölkopf and D. Schuurmans (eds), Advances in Large Margin Classifiers: MIT Press, Cambridge, MA. 2000; p. 297-311.
- 4) Bishop C. Pattern Recognition and Machine Learning. New York: Springer.; 2006.
- 5) 清水典史, 井上寛, 松延千春, 他. サポートベクターマシンに基づく試験合否予測モデルの構築. 薬学教育. 2018; 2 : 2017-023. Doi : 10. 24489/jjphe. 2017-023.
- 6) 清水典史, 井上寛, 松延千春, 他. 薬剤師国家試験合否予測モデルを利用した学生の学修状況の把握. 薬学教育. 2019; 3 : 2018-037. Doi : 10. 24489/jjphe. 2018-037.
- 7) Breiman L. Random forests. Machine Learning, 45 (1), 1-11 (2004).
- 8) Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd ed. : Springer.; 2009.
- 9) 弓削田祥子, 西丸宏, 加瀬義夫. 薬学教育. 2022; 6 : 2021-026. Doi : 10. 24489/jjphe. 2021-026.