

アンサンブル学習による交通機関選択モデルの構築とその評価

著者	長谷川 裕修, 内藤 利幸, 有村 幹治, 田村 亨
雑誌名	土木学会論文集D3 (土木計画学)
巻	68
号	5
ページ	1_773-1_780
発行年	2012
URL	http://hdl.handle.net/10258/00010403

doi: info:doi/10.2208/jscejpm.68.1_773

アンサンブル学習による交通機関選択モデルの構築とその評価

長谷川裕修¹・内藤利幸²・有村幹治³・田村亨⁴

¹正会員 秋田工業高等専門学校助教 環境都市工学科 (〒 011-8511 秋田市飯島文京町 1-1)

E-mail: hasegawa@ipc.akita-nct.ac.jp

²非会員 (株) ドーコン 交通部 (〒 004-8585 札幌市厚別区厚別中央 1 条 5 丁目 4-1)

³正会員 室蘭工業大学助教 大学院工学研究科くらし環境系領域 (〒 050-8585 室蘭市水元町 27-1)

E-mail: arimura@mmm.muroran-it.ac.jp

⁴フェロー会員 室蘭工業大学教授 大学院工学研究科くらし環境系領域 (〒 050-8585 室蘭市水元町 27-1)

E-mail: tamura@mmm.muroran-it.ac.jp

近年、人工知能や機械学習等の知的情報処理分野においてアンサンブル学習というモデル構築手法に注目が集まっている。アンサンブル学習とは、複数の単純なモデルを構築し、それらを用いた分類結果を統合して最終的な分類結果を得る手法であり、高精度だが計算コストが高いニューラルネットワークやサポートベクターマシン等の機械学習手法に劣らない精度を発揮することが知られている。交通機関選択モデルは多種多様な意思決定主体の総体としての判断を表現するものであり、アンサンブル学習によるモデル化が有効であると考えられる。本研究では、平成 18 年に実施された道央都市圏パーソントリップ調査結果を用いてアンサンブル学習による交通機関選択モデル構築を行い、他手法との比較分析によって有用性と課題を検討した。

Key Words : *modal choice analysis, ensemble learning, machine learning, disaggregate behavioural model*

1. はじめに

近年、人工知能や機械学習等の知的情報処理分野においてアンサンブル学習 (ensemble learning) というモデル構築手法に注目が集まっている。アンサンブル学習とは、複数の単純なモデルを構築し、それらを用いた分類結果を統合して最終的な分類結果を得る手法であり、高精度だが計算コストが高いニューラルネットワークやサポートベクターマシン (support vector machine, 以下 SVM と記す) 等の機械学習手法に劣らない精度を発揮することが知られている。交通機関選択モデルは多種多様な意思決定主体の総体としての判断を表現するものであり、アンサンブル学習によるモデル化が有効であると考えられる。

本研究では、平成 18 年に実施された道央都市圏パーソントリップ調査結果を用いてアンサンブル学習による交通機関選択モデル構築を行い、他手法との比較分析によって有用性と課題を検討した。

以下、2. 章でアンサンブル学習の概要と交通機関選択モデルへの適用の意義を述べ、3. 章の実証分析によりアンサンブル学習の有用性を示し、最後に 4. 章で本研究の成果をまとめる。

2. 交通機関選択モデルへのアンサンブル学習適用の意義

本章では、最初にアンサンブル学習の概要を述べ、次に代表的なアンサンブル学習手法であるランダムフォレスト (random forests) について概説し、最後に交通機関選択モデルへのアンサンブル学習適用の意義を示す。

(1) アンサンブル学習の概要

入力に対して分類結果を出力するモデルを分類器といい、分類器を複数組み合わせ (アンサンブル)、それらの結果を統合することによって精度向上を図る手法をアンサンブル学習またはコミッティー学習 (committee learning) という。近年、個々の分類器の精度を向上させるための工夫に限界があることが分かってきたことから、個々の分類器の精度が低くても、全体として高い精度を発揮するアンサンブル学習に注目が集まっている。アンサンブル学習には、訓練データの作り方や学習アルゴリズムの異なるバギング (bagging)、ブースティング (boosting)、確率的属性選択 (stochastic attribute selection)、スタッキング (stacking)、ランダムフォレストなど多くの手法があり、人工知能や機械学習等の知的情報処理分野において研究が進んでいる¹⁾。

(2) ランダムフォレストの概要

ランダムフォレストは多数の決定木を用いたアンサンブル学習の代表的な手法であり、分類問題では各決定木による多数決で最終的に分類するクラスを決定する。ランダムフォレストのアルゴリズムを以下に示す²⁾。

- (a) 訓練データからランダムな復元抽出により N 組の訓練集合を作成する。なお、各訓練集合の抽出の際、元の訓練データの 1/3 は抽出対象から除外し、OOB (out-of-bag) データとして保存する
- (b) 個々の訓練集合を用いて枝刈りされていない最大の決定木を N 本作成する。このとき、各決定木の分岐のノードはランダムに選択された変数の中から OOB データを最も精度良く分類するものを選択する
- (c) 各決定木による N 通りの分類結果の中で最も多いものを最終的な結果として出力する

(3) 交通機関選択モデルへのアンサンブル学習適用の意義

交通行動分析における機械学習手法適用の意義を秋山³⁾は『交通行動を現象論的に考えると、人間の空間的移動に対する「問題解決」であるといえる。このとき、人間は過去に処理した問題と類似の問題は、経験をもとにうまく解決するという「学習能力」を有している。したがって、交通行動分析においても、学習能力が付加された行動モデルを構築することは、交通行動に関する知識の蓄積という点で有効である。』と述べている。アンサンブル学習も機械学習の一手法群であり、更に、そのアルゴリズムからも交通機関選択モデルへのアンサンブル学習適用が有用なことが示される。

社会的意思決定において、その主体である人々は家庭・企業・趣味のサークルなど複数の活動場所を持ち、さらに各活動場所における意思決定で重視される要因が異なり、また多くの場面で多数決ルールに基づく意思決定を行うが、これは本章(2)に示したランダムフォレストのアルゴリズムと整合的である。

よって、多種多様な意思決定主体の総体としての判断を表現する交通機関選択モデルへのアンサンブル学習適用には意義があると言える。

3. 実証分析と考察

本章ではアンサンブル学習における代表的な手法であるランダムフォレストを用いた交通機関選択モデル(以下、RFモデルと記す)、単一の学習器によって高い精度を発揮するSVMを用いた交通機関選択モデル(以下、SVMモデルと記す)、交通行動分析において一般的に用いられる多項ロジットモデルを用いた交通機関

選択モデル(multinomial logit model, 以下MNLモデルと記す)の比較分析を行い、アンサンブル学習の都市交通の評価技術としての有用性と課題を検討する。

(1) 分析対象

本研究では、平成18年に札幌市およびその周辺市町村を対象として行われた第4回道央都市圏パーソントリップ調査により得られた通勤交通トリップを用いて交通機関選択モデルの構築を行う。

使用するパーソントリップデータは、目的トリップ毎に個人属性・トリップ属性・調査日など1レコード当たり162項目、全233177レコードを持つ。

モデル構築に先立ち、以下に示す手順でデータの前処理を行った。

(a) 以下の条件に合致するレコードを抽出

- 通勤トリップ
- トリップチェーンの1トリップ目
- 自宅から出発
- 午前中に移動が完結

(b) 公共交通利用のうち、以下に当てはまるレコードを削除

- 身障者パス利用
- 1DAYカード・ドニチカキップ(1日のみ定額で乗り放題)利用
- 定期券利用ありと回答し、定期券料金以外の支払料金が極端に高額

(c) 分析に用いる変数を選択

- 被説明変数：代表交通手段
 - 自動車
 - 公共交通機関(路線バス・鉄道)
 - その他(徒歩・二輪車)
- 説明変数：
 - 年齢(歳)
 - 性別：ダミー変数
 - 免許の有無：ダミー変数
 - 専用自動車の有無：ダミー変数
 - 所要時間(分)：出発時刻と到着時刻の差
 - 所要費用(円)
 - * 自動車：駐車駐輪月額料金を日単位に換算(30で除した)額と駐車駐輪支払料金の合計金額
 - * 公共交通機関：定期券利用分を除く合計支払額
 - * その他：駐車駐輪月額料金を日単位に換算(30で除した)額と駐車駐輪支払料金の合計金額

(d) モデル構築に用いる訓練データとモデルの評価に用いるテストデータに分割

表-1 通勤手段別サンプル数

	訓練データ	テストデータ	Total
auto	9039	9039	18078
mass	5157	5157	10314
other	2580	2581	5161
Total	16776	16777	33553

- 全データから交通手段ごとに 1/2 を非復元ランダム抽出し、モデルの構築に用いる訓練データとする
- 残りの 1/2 をモデルの評価に用いるテストデータとする

前処理の結果、被説明変数として代表交通手段を、説明変数として「年齢」「性別」「免許の有無」「専用自動車の有無」「所要時間」「所要費用」を持つ全 33553 サンプルを抽出し、これを訓練データとテストデータに分割した。通勤手段別のサンプル数を表-1 に示す。表中の auto, mass, other はそれぞれ自動車、公共交通機関、その他（徒歩・二輪車等）を表す。

(2) モデルの性能評価方法

モデルの適用領域によってその評価基準は異なるが、交通行動の選択を確率的に取り扱う際には尤度を元に評価を行うことが一般的であり、補足的に訓練データへの的中率が用いられてきた⁴⁾。しかし、モデル構築に用いる訓練データに対する学習結果の適合度が良いモデルであっても、未学習のテストデータに対する予測精度（汎化性能）は悪い場合がある。このようなモデルの訓練データへの過適合は、交通機関選択モデルのように政策の意思決定に大きな影響を及ぼす場合には、特に注意が必要である。

また、ロジットモデルが特定の選択肢に対して過大な推計結果を与える例⁵⁾も報告されており、各選択肢に対する精度も検討することが望ましい。

そこで本研究では、訓練データ・テストデータそれぞれに対して、モデルによる分類結果と実データでの観測値を分割表を用いて整理し、ここから計算される有効性指標から各モデルの性能を評価することとした。本章(1)で述べた通り、本研究で用いる選択肢は自動車、公共交通機関、その他（徒歩・二輪車等）の3種類であるから、分割表は表-2 のようになる。

表中の auto 行・auto 列の aa はモデルが自動車利用と分類し、実際に自動車利用であったサンプル数を、auto 行・mass 列の am はモデルが自動車利用と分類し、実際には公共交通機関利用であったサンプル数を、

auto 行・other 列の ao はモデルが自動車利用と分類し、実際には公共交通機関利用であったサンプル数をそれぞれ表す。以下、mass 行および other 行においても同様である。

分割表から得られる有効性指標には、医療分野のスクリーニングにおいて用いられる感度・特異度・偽陽性率・真陰性率・陽性的中度・陰性的中度・検出率、情報検索や機械学習のシステム評価に用いられる精度・再現率などがあるが、本研究では、選択肢毎の性能を評価する指標として精度と再現率を用いる。また、モデルの全体的な性能は、全サンプル数に対する的中サンプルの割合を表す的中率 H によって評価する。精度 α は、モデルによる分類結果の正確さを示す指標であり、選択肢毎の精度は式(1)~(3)により求められる。

$$\alpha_{auto} = \frac{aa}{aa + am + ao} \quad (1)$$

$$\alpha_{mass} = \frac{mm}{ma + mm + mo} \quad (2)$$

$$\alpha_{other} = \frac{oo}{oa + om + oo} \quad (3)$$

再現率 k はモデルが実データの観測値をどの程度再現したかを表す指標であり、交通手段毎の再現率は式(4)~(6)により求められる。

$$k_{auto} = \frac{aa}{aa + ma + oa} \quad (4)$$

$$k_{mass} = \frac{mm}{am + mm + om} \quad (5)$$

$$k_{other} = \frac{oo}{ao + mo + oo} \quad (6)$$

的中率 H は次式により求める。

$$H = \frac{aa + mm + oo}{aa + am + ao + ma + mm + mo + oa + om + oo} \quad (7)$$

なお、テストデータに対する的中率 H が先に述べた汎化性能を評価する指標となる。

(3) RF モデル

本章(1)で作成した 16776 サンプルの訓練データからランダムな復元抽出により 500 組の訓練集合を作成し、それらを用いて 500 本の決定木からなる RF モデルを構築した⁶⁾。

構築した RF モデルにおける説明変数の重要度を図-1 に示す。図中の横軸は各説明変数がジニ係数の減少に寄与する度合いを示しており、これが高いほど、分類精度への貢献が高いことを表す。高い方から順に「公共交通の所要費用」、「専用自動車の有無」、「所要時間」、「免許の有無」、「性別」、「年齢」、「自動車の所要費用」となり、特に重要度が高い変数においてこれまでの既往研究および一般的な感覚と合致する結果が得られた。

構築した RF モデルの評価のために、訓練データとテストデータの分類を行った。表-3 に RF モデルによ

表-2 性能評価のための分割表

	auto	mass	other	Total
auto	aa	am	ao	aa+am+ao
mass	ma	mm	mo	ma+mm+mo
other	oa	om	oo	oa+om+oo
Total	aa+ma+oa	am+mm+om	ao+mo+oo	aa+ma+oa+am+mm+om+ao+mo+oo

Importance of variables(RF model)

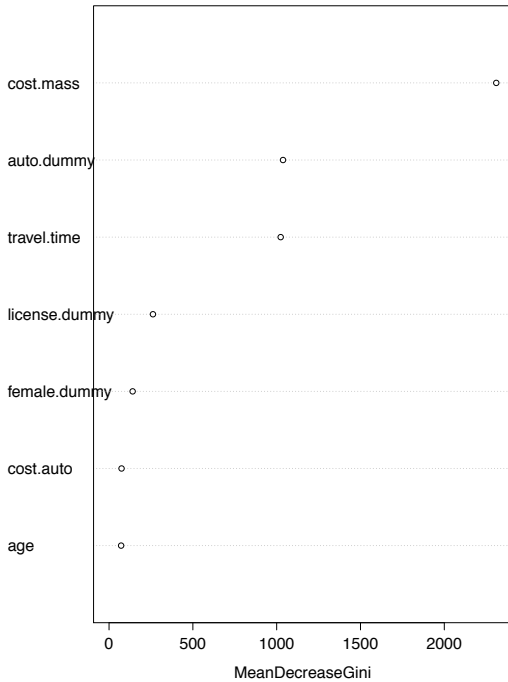


図-1 RF モデルにおける説明変数の重要度

表-3 RF モデルによる訓練データ分類結果

RF (訓練)	auto	mass	other	Total
auto	8136	723	1162	10021
mass	345	4366	174	4885
other	558	68	1244	1870
Total	9039	5157	2580	16776

$$\alpha_{auto} = 0.812, \alpha_{mass} = 0.894, \alpha_{other} = 0.665,$$

$$k_{auto} = 0.900, k_{mass} = 0.847, k_{other} = 0.482,$$

$$H = 0.819$$

る訓練データの、表-4 にテストデータの分類結果を示す。

表-3, 4 より、以下のことが分かる。

表-4 RF モデルによるテストデータ分類結果

RF (テスト)	auto	mass	other	Total
auto	8196	720	1121	10037
mass	322	4354	188	4864
other	521	83	1272	1876
Total	9039	5157	2581	16777

$$\alpha_{auto} = 0.817, \alpha_{mass} = 0.895, \alpha_{other} = 0.678,$$

$$k_{auto} = 0.907, k_{mass} = 0.844, k_{other} = 0.493,$$

$$H = 0.824$$

- (a) 訓練データ・テストデータともにはほぼ同様の分類結果であるが、自動車・公共交通機関の再現率を除きテストデータの方が高性能
- (b) 訓練データでは再現率が高く、テストデータでは精度が高い
- (c) 各手段に分類されたサンプル数を見ると、自動車は過大、公共交通機関は適切、その他は過少に分類された
- (d) 自動車・公共交通機関の精度・再現率ともに0.8を超え良好
- (e) その他は精度・再現率ともに低く、特に再現率は0.5を切っている

(4) SVM モデル

SVM^{(7),(8),(9)}は、入力と出力の組からなる訓練サンプルを教師信号として、その背後に存在する入出力関係を学習することで2クラスのパターン分類器を構成する機械学習手法の一つである。SVMは、Vapnikらが1960年代に提案した Optimal Separating Hyperplane を基礎とする分類手法であり、1990年代になって Vapnik 自身によりカーネル関数を組み込むことで非線形分類にも対応できるモデルとして拡張された。

SVMは迷惑メールフィルタのようなテキストデータの分類、手書き文字認識のような画像認識、生物学的データマイニングにおけるDNAマイクロアレイ技術開発への適用というように、大規模データベースからの知

表-5 SVM モデルパラメータ大域的探索結果

γ	<i>cost</i>				
	10^{-2}	10^{-1}	10^0	10^1	10^2
10^{-5}	0.539	0.539	0.541	0.749	0.776
10^{-4}	0.539	0.541	0.748	0.776	0.791
10^{-3}	0.541	0.745	0.777	0.793	0.798
10^{-2}	0.739	0.780	0.797	0.809	0.809
10^{-1}	0.790	0.808	0.809	0.810	0.810
10^0	0.772	0.804	0.809	0.810	0.811
10^1	0.539	0.760	0.800	0.798	0.795
10^2	0.539	0.550	0.773	0.768	0.768
10^3	0.539	0.539	0.746	0.746	0.746
10^4	0.539	0.539	0.746	0.746	0.746
10^5	0.539	0.539	0.746	0.746	0.746

識発見を目指したデータマイニング分野での応用が進んでいる⁷⁾。土木計画分野における SVM 適用例としては、福田らによるもの¹⁰⁾や筆者らによるもの^{11),12),13)}がある。

SVM には大別すると分類誤りを許容するソフトマージン SVM と許容しないハードマージン SVM があり、さらに、使用するカーネル関数によって細分される。カーネル関数は元々のデータを線形分離可能な高次元空間に写像するために用いられ、線形カーネル、多項式カーネル、RBF カーネル、フィッシャーカーネル、TOP カーネル等が知られている¹⁴⁾。本研究では最も一般的に用いられる、ソフトマージン SVM と式 (8) で表される RBF カーネルを組み合わせた SVM を用いる。

$$K(\mathbf{x}_i - \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (8)$$

この SVM の分類能力にはソフトマージンのパラメータ *cost* と RBF カーネルのパラメータ γ の組み合わせが大きく影響する。本研究では以下に示す 2 段階の方法でパラメータを決定した。

- (a) 大域的な探索：*cost* を 5 通り (10^{-2} から 10^2)、 γ を 11 通り (10^{-5} から 10^5) の計 55 通りの組み合わせを試行
- (b) 局所的な探索：大域的な探索で最も精度の高かった *cost* = 100、 γ = 1 の近傍で *cost* を 3 通り (0, 1, 2)、 γ を 3 通り (99, 100, 101) の計 9 通りの組合せを試行

表-5、表-6 に SVM モデルのパラメータ探索結果を示す。表中の数字は、 γ 、*cost* 毎の分類精度を表している。これより、本研究は *cost* = 100、 γ = 1 を用いる。

SVM モデルによる訓練データの分類結果を表-7 に、テストデータの分類結果を表-8 に示す。

表-7、8 より、以下のことが分かる。

表-6 SVM モデルパラメータ局所的探索結果

γ	<i>cost</i>		
	99	100	101
0	0.539	0.539	0.539
1	0.811	0.811	0.811
2	0.808	0.808	0.808

表-7 SVM モデルによる訓練データ分類結果

SVM (訓練)	auto	mass	other	Total
auto	8189	762	1096	10047
mass	344	4316	172	4832
other	506	79	1312	1897
Total	9039	5157	2580	16776

$$\alpha_{auto} = 0.815, \alpha_{mass} = 0.893, \alpha_{other} = 0.692, \\ k_{auto} = 0.906, k_{mass} = 0.837, k_{other} = 0.509, \\ H = 0.824$$

表-8 SVM モデルによるテストデータ分類結果

SVM (テスト)	auto	mass	other	Total
auto	8141	814	1098	10053
mass	375	4228	181	4784
other	523	115	1302	1940
Total	9039	5157	2581	16777

$$\alpha_{auto} = 0.810, \alpha_{mass} = 0.884, \alpha_{other} = 0.671, \\ k_{auto} = 0.901, k_{mass} = 0.820, k_{other} = 0.504, \\ H = 0.815$$

- (a) 訓練データ・テストデータともにほぼ同様の分類結果であるが、訓練データの方が全ての性能が高い
- (b) 各手段に分類されたサンプル数を見ると、自動車は過大、公共交通機関は適切、その他は過少に分類された
- (c) 自動車・公共交通機関の精度・再現率ともに 0.8 を超え良好
- (d) その他は精度・再現率ともに低く、特に再現率は 0.5 を若干上回る程度

(5) MNL モデル

機械学習手法であるランダムフォレスト・SVM との比較のため、ランダム効用理論を背景とし、実務においても広く用いられている MNL モデル^{4),15)}による分

表-9 特性変数の特定化

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
自動車	1	0	age	f	l	a	ttime	c_a	0
公共交通	0	1	age	f	0	0	ttime	0	c_m
その他	0	0	age	f	0	0	ttime	0	0

表-10 MNL モデルのパラメータ推計結果

	パラメータ	t 値
定数項 (自動車)	-1.004	-13.8559
定数項 (公共交通)	-0.279	-9.3005
年齢	28.038	0.0401
女性ダミー	-337.080	-0.1206
免許ダミー	1.283	16.3785
専用自動車有無ダミー	2.000	44.1482
所要時間	-18.698	-0.0268
所要費用 (自動車)	9.068	1.6827
所要費用 (公共交通)	15.157	2.0190

$$\rho^2 = 0.413 \quad \bar{\rho}^2 = 0.413 \quad \text{的中率} = 0.760$$

析を行った。本研究では式 (9) に示す線形効用関数を用いる。

$$V_{in} = \sum_{k=1}^K \theta_k X_{ink} \quad (9)$$

ここで、 V_{in} は個人 n が選択肢 i を選択したときの効用の確定部分、 θ_k は k 番目の説明変数のパラメータ、 X_{ink} は説明変数を表す。効用関数の特性変数を表-9 に示す。表中の $\theta_1 \sim \theta_9$ はそれぞれ「自動車の定数項」、「公共交通の定数項」、「年齢 (10 歳)」、「女性ダミー」、「免許ダミー」、「専用自動車有無ダミー」、「所要時間 (10 分)」、「自動車の所要費用 (100 円)」、「公共交通の所要費用 (100 円)」に対するパラメータを表す。

各交通機関の選択確率は式 (10) で表される。

$$P_{in} = \frac{\exp(V_{in})}{\sum_{k=1}^K \exp(V_{jn})}, i = 1, \dots, J \quad (10)$$

ここで、 P_{in} は個人 n が選択肢 i を選択する確率を表す。

訓練データを用いて MNL モデルのパラメータ推計結果を行った。結果を表-10 に示す。

モデルの適合度は $\rho^2 = 0.413$ 、 $\bar{\rho}^2 = 0.413$ と比較的高く、的中率も 0.760 と良好な結果を得た。

表-11 MNL モデルによる訓練データ分類結果

MNL (訓練)	auto	mass	other	Total
auto	8801	1552	1840	12193
mass	0	3205	0	3205
other	238	400	740	1378
Total	9039	5157	2580	16776

$$\alpha_{auto} = 0.722, \alpha_{mass} = 1.000, \alpha_{other} = 0.537, \\ k_{auto} = 0.974, k_{mass} = 0.621, k_{other} = 0.287, \\ H = 0.760$$

表-12 MNL モデルによるテストデータ分類結果

MNL (テスト)	auto	mass	other	Total
auto	8806	1546	1820	12172
mass	0	3190	0	3190
other	233	421	761	1415
Total	9039	5157	2581	16777

$$\alpha_{auto} = 0.723, \alpha_{mass} = 1.000, \alpha_{other} = 0.538, \\ k_{auto} = 0.974, k_{mass} = 0.619, k_{other} = 0.295, \\ H = 0.760$$

しかし、各パラメータおよび t 値を見ると、

- (a) 一般に負の効用を持つと考えられる「所要費用」の符号が正である
- (b) 「年齢」・「女性ダミー」・「所要時間」・「所要費用 (自動車)」は 5% 水準で有意でない

という問題がある。そこで 5% 水準で有意でない変数のうち、所要時間以外の変数を除去して再度パラメータ推定を試みた。その結果、適合度は $\rho^2 = 0.413 \rightarrow 0.404$ 、 $\bar{\rho}^2 = 0.413 \rightarrow 0.404$ と若干減少し、的中率は元と同じ値であった。よって以下の分析では、他手法との比較のため変数除去前の MNL モデル (表-10) を用いる。

MNL モデルによる訓練データの分類結果を表-11 に、テストデータの分類結果を表-12 に示す。

表-11, 12 より、以下のことが分かる。

- (a) 訓練データ・テストデータともにほぼ同様の分類結果となった
- (b) 各手段に分類されたサンプル数と観測値の差が大きく、自動車は過大に、公共交通機関とその他は過少に分類された
- (c) 公共交通機関の的中率は 1.0 で非常に良好だが、実数に対して過少な値であり、高い性能を発揮しているとは言い難い
- (d) その他は精度・再現率ともに低く、特に再現率は

0.3 を下回る

(6) 考察

各モデルの訓練データ・テストデータ分類結果(表-3~8, 11, 12)より, 以下のことが明らかとなった.

- (a) 訓練データ・テストデータともに機械学習である RF モデル・SVM モデルの的中率が MNL モデルを上回った
- (b) RF モデルのみテストデータの的中率が訓練データの的中率を上回り, 高い汎化能力を示した
- (c) RF モデル・SVM モデルともに全ての交通手段の再現率で MNL モデルを上回った
- (d) 各手段に分類されたサンプル数を見ると, RF モデル・SVM モデルは実際の手段別シェア(表-1)に近い値を示したが, MNL モデルは大きく乖離した値となった

以上より, 分類性能から評価すると, 特にテストデータへの分類性能に優れる RF モデルの有用性が明らかとなった.

次に, RF モデルと MNL モデルの分類結果に対する各変数の重要度の違いについて考察する. RF モデルでは変数の重要度をジニ係数の減少への寄与の度合いによって測り, 大きい方から「所要費用(公共交通)」、「専用自動車の有無」、「所要時間」、「免許の有無」、「性別」、「所要費用(自動車)」、「年齢の順」であった. 一方, MNL モデルでは変数の有意性を t 値によって測る. よって t 値の高い順に「専用自動車の有無」、「免許の有無」、「定数項(自動車)」、「定数項(公共交通)」、「所要費用(公共交通)」の順に有意となる. 定数項が上位に入っていることから, 今回構築した MNL モデルで使用した変数では不十分であることが分かる. また, MNL モデルでは自動車利用が過大に推計されたが, これは自動車に関する変数が高い有意性を示したことと関係すると考えられる.

以上の分析結果から, アンサンブル学習手法の一つである RF モデルの有用性が明らかとなった.

4. おわりに

本研究では, 平成 18 年に実施された道央都市圏パースントリップ調査結果を用いてアンサンブル学習による交通機関選択モデル構築を行い, 他手法との比較分析によって有用性と課題を検討した.

本研究により以下の 5 点が明らかとなり, アンサンブル学習による交通機関選択モデルの有用性が示された.

- (a) 訓練データ・テストデータともに機械学習である RF モデル・SVM モデルの的中率が MNL モデルを上回った

- (b) RF モデルのみテストデータの的中率が訓練データの的中率を上回り, 高い汎化能力を示した
- (c) RF モデル・SVM モデルともに全ての交通手段の再現率で MNL モデルを上回った
- (d) 各手段に分類されたサンプル数を見ると, RF モデル・SVM モデルは実際の手段別シェア(表-1)に近い値を示したが, MNL モデルは大きく乖離した値となった
- (e) RF モデルは分類結果に対する各変数の重要度を解釈しやすい形で出力する

今後の課題は以下の 3 点である.

- (a) 利用の少ない「その他」の交通手段の再現率を向上させる方法の検討
- (b) RF モデルの移転可能性の検討
- (c) 経路選択や居住地選択等へのアンサンブル学習適用可能性の検討

謝辞: 本研究の分析は全てオープンソースの統計計算システム R¹⁶⁾とその拡張パッケージ^{6),17)}を用いて行った. 特に, R による MNL モデルのパラメータ推計は東京海洋大学の兵藤哲朗先生が Web 上に公開されている文書を参考にした. また, 本研究の遂行にあたり科研費(23760469)の助成を受けた. ここに記して感謝の意を表す.

参考文献

- 1) 元田浩, 山口高平, 津本周作, 沼尾正行: データマイニングの基礎, オーム社, 2006.
- 2) Breiman, L.: Random forests, *Machine Learning*, Vol. 45, pp. 5-32, 2001.
- 3) 秋山孝正: 知的情報処理を利用した交通行動分析, 土木学会論文集, No. 688/IV-53, pp. 37-47, 2001.
- 4) 北村隆一, 森川高行, 佐々木邦明, 藤井聡, 山本俊行: 交通行動の分析とモデリング, 技報堂出版, 2002.
- 5) 秋山孝正, 奥嶋政嗣: 交通機関選択分析のためのファジィ決定木手法の比較検討, 土木学会論文集, Vol. 63, No. 2, pp. 145-157, 2007.
- 6) Liaw, A. and Wiener, M.: Classification and Regression by randomForest, *R News*, Vol. 2, No. 3, pp. 18-22, 2002.
- 7) Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*, Cambridge University Press, 2000. 大北剛訳: サポートベクターマシン入門, 共立出版, 2005.
- 8) 小野田崇: サポートベクターマシン, オーム社, 2007.
- 9) C. M. ビショップ: パターン認識と機械学習下巻: 理論による統計的予測, シュプリンガー・ジャパン, 2008.
- 10) 福田大輔, 庭田美穂, 屋井鉄雄: 疑問型表現自由回答データを用いた社会資本整備に対する市民の関心の抽出方法に関する基礎的研究, 土木計画学研究・論文集, Vol. 24, No. 1, pp. 131-140, 2007.
- 11) 有村幹治, 長谷川裕修, 藤井勝, 田村亨: 非線形最適化へのサポートベクターマシンの応用に関する考察, 土木計画学研究・論文集, Vol. 24, No. 3, pp. 421-426, 2007.
- 12) Hasegawa, H., Fujii, M., Arimura, M. and Tamura, T.: A Basic Study on Traffic Accident Data Analysis Using Support Vector Machine, *Journal of the East-*

- ern Asia Society of Transportation Studies*, Vol. 7, pp. 2873-2880, 2007.
- 13) 長谷川裕修：識別分析手法による土木計画データからの知識発見に関する研究，博士論文，室蘭工業大学，2009.
 - 14) 津田宏治：カーネル設計の方法，日本神経回路学会誌，Vol. 9, No. 3, pp. 190-195, 2002.
 - 15) 土木学会土木計画学研究員会（編）：非集計行動モデルの理論と実際，土木学会，1995.
 - 16) Team, R. D. C.: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.
 - 17) Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A.: *e1071: Misc Functions of the Department of Statistics (e1071)*, TU Wien, 2010. R package version 1.5-24.

(2012. 2. 25 受付)

MODAL CHOICE ANALYSIS USING ENSEMBLE LEARNING METHODS

Hironobu HASEGAWA, Toshiyuki NAITO, Mikiharu ARIMURA and Tohru TAMURA

This study aims to determine an appropriate method for modal choice in the urban area. In our research fields, we usually use disaggregate behavioural models for modal choice analysis. The random forest method as ensemble learning technique is applied to describe the decision mechanism in the travel behaviour. As the result of this study, it is confirmed that ensemble learning method is more effective than other conventional methods.