土木学会論文集 No.400/I–22, pp.1–6, 2011.12

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　博（工）　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　工修　（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　博（工）　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　工博　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

バックキャスティングを伴う低炭素都市構造を設計するためには，精度が高く，汎用性に優れた交通機関選択モデルが要求される．交通機関選択モデルは多種多様な意志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．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をいい，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蓄積が進んでいる．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，都市交通の評価技術としての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

バックキャスティングを伴う低炭素都市構造を設計するためには，精度が高く，汎用性に優れた交通機関選択モデルが要求される．交通機関選択モデルは多種多様な意志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（ensemble learning）によるモデル化が有効であると考えられる．

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をいい，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蓄積が進んでいる．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，都市交通の評価技術としての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都市交通評価技術としての有用性を示し，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

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方法をアンサンブル学習またはコミッティー学習 (committee learning) といい，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．アンサンブル学習が着目される理由は，これまでに行われてきた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からである．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，上記の見解が適用出来る．

秋山の見解に加えて，本章 (2) に示したランダムフォレストのアルゴリズムの社会的意志決定へのアナロジーにより交通機関選択モデルへのアンサンブル学習適用が意義づけられる．すなわち，

(a) 社会に属する個人が家庭・企業・趣味のサークルなど複数の活動場所を持つことを表す
(b) 各活動場所における意志決定で重視される要因が異なることを表す
(c) 多数決ルールに基づく意志決定を表す

よって，多種多様な意志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サポートベクターマシン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
- トリップチェインの 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
- 定期利用で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（自動車利用・公共交通機関利用・その他に集約）
- 説明変数：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

(d) モデル推定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
- 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，訓練データとする
- 残りの 1/2 を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．

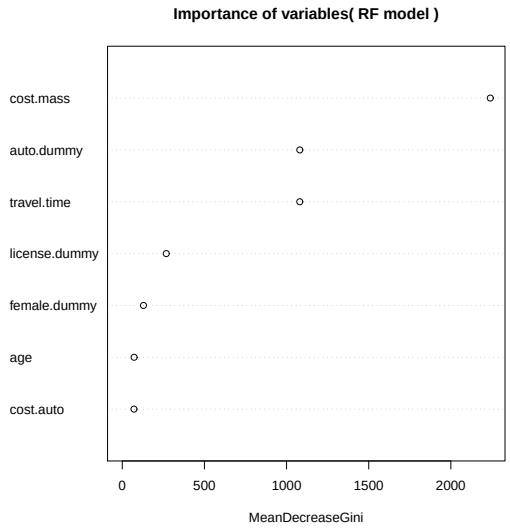
表–1 通勤手段別サンプル数

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

(2) 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[4].

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一般的な感覚と合致する結果が得られた．



図–1 RF モデルにおける説明変数の重要度

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–2** に RF モデルによる訓練データの分類結果を示す．表中の auto 行・auto 列の数字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 列の数字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 列の数字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表す．以下，mass 行および other 行においても同様である．なお，**表–3〜5**，**8**，**9** も同じ表し方で整理している．**表–3** に RF モデルによるテストデータの分類結果を示す．

表–2 RF モデルによる訓練データ分類結果

	auto	mass	other
auto	8136	723	1162
mass	345	4366	174
other	558	68	1244

誤分類率 = 0.181

表–3 RF モデルによるテストデータ分類結果

	auto	mass	other
auto	8196	720	1121
mass	322	4354	188
other	521	83	1272

誤分類率 = 0.176

表–2，**3** 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であるが，全体にテストデータに対する精度の方が高い

(b) 各手段に分類されたサンプル数を見ると，自動車は多く，公共交通機関は適切，その他は少なく分類された

(c) 公共交通機関の誤分類率は 0.1 程度で良好だが，その他の誤分類率は 0.3 を超えており低い

(3) SVM モデル

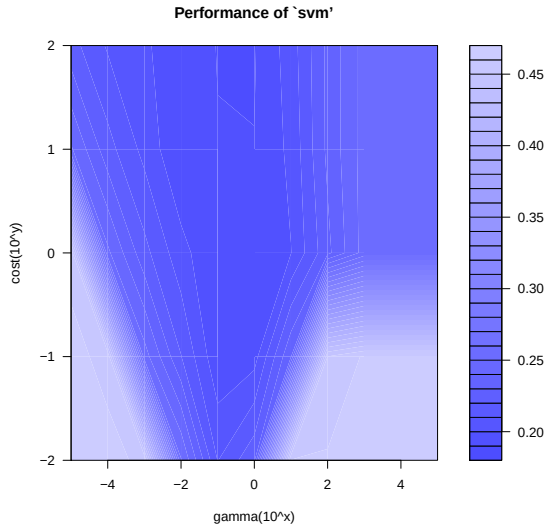
サポートベクターマシン[5),6),7)]（support vector machine，以下 SVM と記す）は，入力と出力の組からなる訓練サンプルを教師信号として，その背後に存在する入出力関係を学習することで 2 クラスのパターン分類器を構成する機械学習手法の一つである．SVM は，Vapnik らが 1960 年代に提案した Optimal Separating Hyperplane を基礎とする分類手法であり，1990 年代になって Vapnik 自身によりカーネル関数を組み込むことで非線形分類にも対応できるモデルとして拡張された．

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[5]．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[8]や筆者らによるもの[9),10),11)]がある．

本研究では RBF カーネルを用いた非線形ソフトマージン SVM を用いるが，この分類能力には二種類のパラメータ $cost$ と γ の組み合わせが大きく影響する．本研究では $cost$ 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し，最も精度の高かった $cost = 100$，$\gamma = 1$ の組み合わせを用いることとした．

図–2 に SVM モデルのパラメータ探索結果を示す．



図–2 SVM モデルパラメータ探索結果

SVM モデルによる訓練データの分類結果を**表–4** に，テストデータの分類結果を**表–5** に示す．

表–4 SVM モデルによる訓練データ分類結果

	auto	mass	other
auto	8189	762	1096
mass	344	4316	172
other	506	79	1312

誤分類率 = 0.176

表–4，**5** 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であるが，全体に訓練データに対する精度の方が高い

(b) 各手段に分類されたサンプル数を見ると，自動車は多く，公共交通機関は適切，その他は少なく分類された

(c) 公共交通機関の誤分類率は 0.1 程度で良好だが，その他の誤分類率は 0.3 を超えており低い

表–5 SVM モデルによるテストデータ分類結果

	auto	mass	other
auto	8141	814	1098
mass	375	4228	181
other	523	115	1302

誤分類率 = 0.185

(4) 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[12),13)]による分析を行った．本研究では式 (1) に示す線形効用関数を用いる．

$$V_{in} = \sum_{k=1}^{K} \theta_k X_{ink} \tag{1}$$

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–6** に示す．表中の θ_1 ～θ_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 (2) で表される．

$$P_{in} = \frac{exp(V_{in})}{\sum_{k=1}^{K} exp(V_{jn})}, i = 1, \ldots, J \tag{2}$$

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–7** 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．

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

表–6 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_a	f_a	l	a	$ttime_a$	c_a	0
公共交通	0	1	age_m	f_m	0	0	$ttime_m$	0	c_m
その他	0	0	age_o	f_o	0	0	$ttime_o$	0	0

表–7 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

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

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–7**）を用いる．

MNL モデルによる訓練データの分類結果を**表–8** に，テストデータの分類結果を**表–9** に示す．

表–8 MNL モデルによる訓練データ分類結果

	auto	mass	other
auto	8801	1552	1840
mass	0	3205	0
other	238	400	740

誤分類率 $= 0.240$

表–9 MNL モデルによるテストデータ分類結果

	auto	mass	other
auto	8806	1546	1820
mass	0	3190	0
other	233	421	761

誤分類率 $= 0.240$

表–8，**9** 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を見ると，自動車は過多に，公共交通機関とその他は過少に分類される結果となった
(c) 公共交通機関の誤分類率は 0 で非常に良好だが，実数に対して過少な値であり，精度が高いとは言い切れない
(d) その他の誤分類率は 0.5 弱とかなり低い

(5) 考察

結果に対する考察を行う前に，機械学習と非集計モデル両者の方向性の違いについて述べる．MNL をはじめとする非集計モデルにおけるモデル推定は，最尤法によって『理論モデルに含まれる未知パラメータを，実験や観測データを最も良く再現するように同定することである[13].』よってパラメータ推定に用いたデータを用いて的中率やその他の適合度指標を算出し，モデルの妥当性を評価する．すなわち，現象記述・理解を主な目的としたモデルであると言える．一方，機械学習におけるモデル推定は，入出力の関係が既知の訓練集合を用いて未知の入力に対して正しく出力を返す関数を求めることであり，この過程を訓練または学習と呼ぶ．よってモデルは訓練で使ったものとは異なるテスト集合を正しく分類する能力（＝汎化能力）によって性能が評価される．MNL モデルとは違い，予測を主な目的としたモデルと言える．

以上を踏まえて各モデルの訓練データ・テストデータの分類結果（**表–2〜5**，**8**，**9**）を考察すると，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 モデルの分類精度が MNL モデルを上回る結果となった．また，RF モデルのみテストデータの精度が訓練データの精度を上回り，高い汎化能力を示した
(b) RF モデル・SVM モデルともにテストデータで最も悪い誤分類率はその他の 0.3 程度であり，MNL モデルに比べて高い分類精度を示した
(c) 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

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「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」,「自動車の所要費用」,「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，分類結果（効用関数）への影響の大きさをパラメータの絶対値から測る．よって t 値は高いもののパラメータ・変数の値がともに小さい「専用自動車の有無」,「免許の有無」は分類結果に対する重要度はそれほど高くないと言える．分類結果に対する重要度は 99% の信頼度で有意かつパラメータも大きい「公共交通の所要費用」が最も高いと言え，この結果は RF モデルの結果とも一致する．しかし,「公共交通の所要費用」に続く重要度をどの変数が有するかを明瞭に示すことは難しい．よって，分類結果に対する各変数の重要度を把握するには，MNL モデルよりも RF モデルを用いた方が適当である．

以上の分析結果から，都市交通の評価技術としてのアンサンブル学習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，都市交通の評価技術としての有用性と課題を検討した．

本研究により以下の 4 点が明らかとなり，都市交通の評価技術としてのアンサンブル学習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 モデルの分類精度が MNL モデルを上回る結果となった．また，RF モデルのみテストデータの精度が訓練データの精度を上回り，高い汎化能力を示した

(b) RF モデル・SVM モデルともにテストデータで最も悪い誤分類率はその他の 0.3 程度であり，MNL モデルに比べて高い分類精度を示した

(c) 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

(d) RF モデルは分類結果に対する各変数の重要度を明瞭に出力する

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 R[14] とその拡張パッケージ[4),15)] を用いて行った．また，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．ここに記して感謝の意を表します．

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, IT Text, 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) Liaw, A. and Wiener, M. : Classification and Regression by randomForest, *R News*, Vol. 2, No. 3, pp. 18–22, 2002.
5) 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.
6) 小野田崇: サポートベクターマシン, オーム社, 2007.
7) C. M. ビショップ: パターン認識と機械学習 下 - ベイズ理論による統計的予測, シュプリンガー・ジャパン株式会社, 2008.
8) 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.
9) 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.
10) HASEGAWA, H., FUJII, M., ARIMURA, M. and TAMURA, T. : A Basic Study on Traffic Accident Data Analysis Using Support Vector Machine, *Journal of the Eastern Asia Society of Transportation Studies*, Vol. 7, pp. 2873–2880, 2007.
11) 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.
12) 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.
13) 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.
14) R Development Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.
15) 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.

(? 受付)