リレーションシップ・データへのデータマイニングの適用

鶴田　育緒，後藤　正輝，香田　正人

1. はじめに

市場の成熟によって新規顧客の獲得が困難になった産業では，One-to-Oneマーケティング[7]導入の必要性が高まっている．これは，製品の大量生産，大量プロモーションによる「市場シェア重視」から，個々の消費者や顧客の嗜好やニーズに合わせて，一人一人個別にマーケティングを展開する「個客価値重視」へのシフトを意味する．

近年企業におけるデータウェアハウスの整備に伴い，One-to-Oneマーケティングに利用可能なデータの蓄積がなされている．特に日々のトランザクションから生成されるリレーションシップ・データ（顧客の商品購入行動に関する情報）は，顧客各人のニーズを直接的に把握できる情報であり，また欠損値やノイズが比較的少ない精度の高い情報であるといえる．

本稿ではリレーションシップ・データへのデータマイニングの適用例として，顧客セグメント識別と顧客スコアリングの2つの事例について述べる．いずれも教師付き学習アルゴリズムを利用する問題であり，データに基づいてアプリケーションを生成するプロセスであるともいえる[1]．

2. 顧客セグメント識別

本事例の分析対象は，あるクレジットカード会社（以下A社）のカード利用データである．データマイニングをクレジットカード業界に適用する際には，顧客の属するセグメントを正確に識別することが重要な役割を果たす．たとえば継続的な高額利用者であれば維持重視のプロモーションを，また離反が危惧される利用者には離反防止のキャンペーンを行い，既存顧客の維持につなげることができる．特に新規顧客の獲得には，既存顧客維持の数倍のコストが必要であるといわれ，顧客セグメントの識別はリスク管理の観点からも重視される．

本事例ではカード利用履歴データから，その顧客の所属するセグメントを識別することを目的とする．セグメント化はビジネス上の視点からあらかじめ実施され，各レコードにはカード利用履歴と共にその顧客の所属セグメントが記録されている．予測には適応リサンプリング（adaptive resampling[8]）と重みつき投票（weighted voting）による，複合分類木モデル（multiple decision trees）を利用し，限られたデータをもとにして，より高い識別精度を得ることを目指した．

2.1 利用データ

本事例では，1998年10月にA社のクレジットカードに入会した顧客16,382名の，1998年12月（入会3ヶ月目）から1999年9月（入会12ヶ月目）までの10ヶ月間のカード利用履歴データを利用する．10月に入会した顧客群は，ボーナス時や年度末利用者の変動の影響を受けにくく，利用状況に特別な要因が含まれにくいといわれている．

分析データには，スクランブル化された顧客IDをキーとして，月別のキャッシング，カードショッピングの利用件数，金額などのリレーションシップ・データ，また年齢，居住形態，勤務形態，勤続年数などのデモグラフィック・データ，そしてキャッシング，カードショッピングそれぞれのセグメント区分の情報などが記録されている．

本事例では，特に顧客のキャッシングの利用状況に着目する．A社では顧客のキャッシングの利用について，以下の4セグメントの分類を行っている（図1）．

セグメントA（13,793名）

キャッシング未利用者

セグメントB（805名）

つるた　いくお，ごとう　まさてる
筑波大学 システム情報工学研究科
（〒305-8573 つくば市天王台1-1-1）
こうだ　まさと
筑波大学 社会工学系
（〒305-8573 つくば市天王台1-1-1）

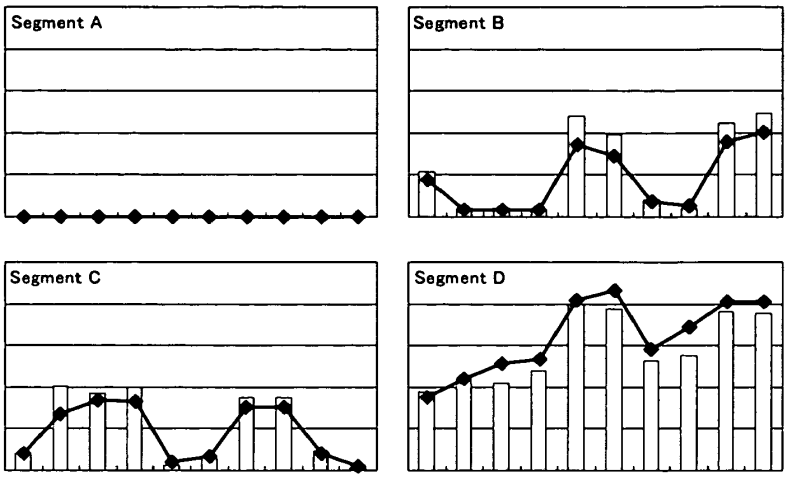

図1　各セグメントのキャッシングの入会3ヶ月目から10ヶ月目までの月別利用状況．折れ線グラフは利用件数，棒グラフは利用金額を示す

ボーナス期，年度末の高額利用者

セグメントC（582名）

ボーナス期，年度末以外の高額利用者

セグメントD（1,202名）

高額利用継続者

キャッシングの10ヶ月間の月別利用件数と月別利用金額の計20の属性値から，この4セグメントのいずれに属するかを予測するセグメント識別モデルを，複合分類木モデルによって作成する．

2.2　複合分類木モデル

顧客のキャッシング利用状況のセグメント分類を確認すると，多くの顧客がキャッシング未利用者のセグメントに属しており，他のセグメントに属する顧客の数は限られていることがわかる．

このようにセグメント間のデータ数が大きく異なる場合，分類木アルゴリズムは，データ数のより多いセグメントから強い影響を受け，結果として出力に偏りが生じることが知られている[3]．したがって，高い汎化性能を持つアルゴリズムを作成するためには，モデル学習時に，データ数の少ないセグメントの情報をより多く提示する必要がある．

本研究では複合分類木モデル（図2）を用いて，望ましい構造を実現した．複合分類木モデルでは，分類木がT回呼び出される．1番目の分類木には，訓練データをそのまま入力してモデルを作成する．対して2番目以降に呼び出される分類木には，直前の分類木における誤分類数に応じて，セグメント毎に抽出される確率が異なった復元無作為抽出（ブートストラップ・リサンプリング）が行われる．T個すべての分類木が作成されたら，それぞれの予測結果の線形結合により，最終予測が行われる（アンサンブル）．

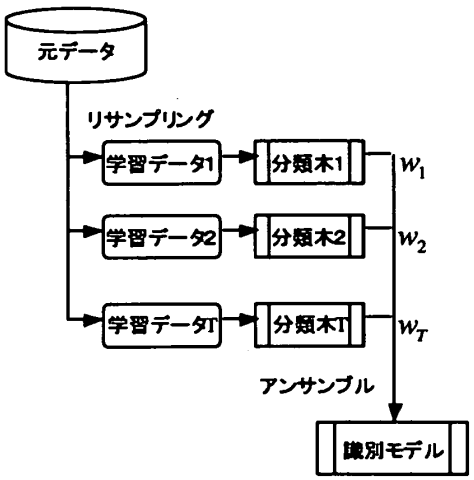

図2　複合分類木モデル

このように限られた数のデータから複数のブートレプリカによる予測モデルを作成し，全体の予測精度を向上させる学習の手法は，アンサンブル学習（ensemble learning）と呼ばれており，他の代表的な手法としては，PAC学習の枠組みに従うAdaBoost（ADAptive BOOSTing[4]）が知られている．

2.2.1　適応リサンプリング

アンサンブル学習の一つにbagging（Bootstrap AGGregatING[2]）がある．これはn件のデータから，大きさnの復元無作為抽出を繰り返すことにより，複数の学習データ（ブートデータ）を生成する手法である．

すべてのデータを$1/n$の確率で抽出する操作を繰り返す代わりに，以前のモデルで誤分類が多く発生した

セグメントのデータを，より高い確率で抽出する手法が適応リサンプリングである．

t 番目の分類木におけるセグメント c の誤分類数を $e(c)$，C をセグメントの総数とした時に，次に与える擬似確率関数に従って，$t+1$ 番目の各セグメントのデータを抽出する．

$$\Pr(c)=\frac{(1+e(c)^r)}{\sum_{i=1}^{C}(1+e(i)^r)}$$

ただし $c=1,\cdots,C$ であり，r は任意の整数である．

2.2.2 重みつき投票モデル

Weiss ら[8]は，各分類木の識別結果の多数決をとり（単純投票），最終モデルの予測を生成した．本事例では，各分類木の識別結果を線形結合する際に，誤分類率の低い分類木の重みを大きく，誤分類率の高い分類木の重みを小さくするような重みつけ（重みつき投票）を行う．

t 番目の分類木の誤分類率を $er(t)$，分類木の総数を T としたときに，t 番目の分類木の重みを次の式で定義する．

$$w(t)=\frac{\exp(-er(t)^2)}{\sum_{i=1}^{T}\exp(-er(i)^2)}$$

ただし $t=1,\cdots,T$ であり，単純投票モデルにおいては $w(1)=\cdots=w(T)=1$ である．また，最終モデルにおいて，2つ以上のセグメントに対して同一の出力値が得られて識別が不能の場合には，誤分類であるものとみなす．なお分類木は S-plus のライブラリ tree を使用した．

2.3 数値実験

A 社の顧客 16,382 件のデータを，非復元無作為抽出により 10,000 件の訓練データと 6,382 件のテストデータに分割する．

単純投票，重みつき投票それぞれについての複合分類木モデルを訓練データから作成し，両者の識別性能を比較する．またテストデータに適用することにより，両モデルの汎化性能を評価する．各モデルを構成する分類木が1個から40個それぞれの場合について比較を行う．

図3，4に，単純投票，重みつき投票それぞれについて，モデルに含まれる分類木の数の増加に伴う，訓練データの誤分類率の推移を示した．単純投票モデルでは，分類木が偶数個の場合に誤分類率が高くなっている．これは，2つ以上のセグメントの得票が同数となり，識別不能になっていることを示している．対して重みつき投票モデルでは，分類木の個数を増やすにつれて誤分類率が低下していくことがわかる．また，テストデータについても同様に，分類木の個数の増加に伴う誤分類率の低下が確認された．

2.4 議論

識別すべきセグメント数が4である本事例では，モデルに含まれる分類木が偶数個，奇数個にかかわらず，識別不能に陥る可能性は存在するはずであるが，図3からも分かるとおり，偶数個の分類木を含む単純投票モデルにおいて誤分類率が高くなっている．

表1，2にそれぞれ1番目と2番目の分類木の混同行列（confusion matrix）を示した．1番目の分類木では，本来セグメントBまたはCに分類されるべき顧客がセグメントDに分類されており，2番目の分類木では逆の傾向がみられる．誤分類の多いセグメントは次のリサンプリングにおいて重視されるため，単純

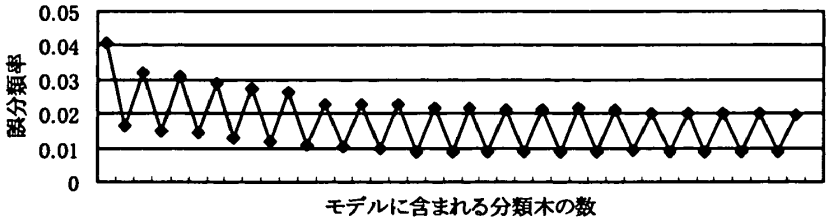

図3　単純投票モデルの誤分類率の推移（訓練データ）

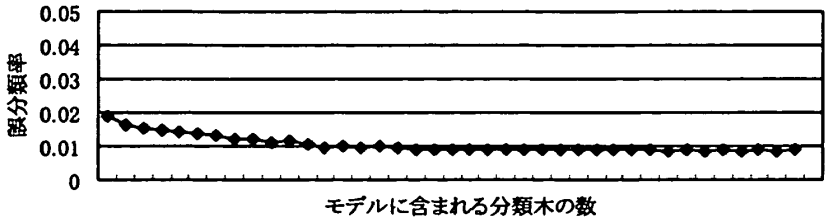

図4　重みつき投票モデルの誤分類率の推移（訓練データ）

表1　1番目の分類木の混同行列

		モデルの予測			
		A	B	C	D
真の分類	A	8387	0	0	0
	B	0	406	0	**90**
	C	0	18	308	**56**
	D	0	28	0	707

表2　2番目の分類木の混同行列

		モデルの予測			
		A	B	C	D
真の分類	A	8387	0	0	0
	B	0	477	13	6
	C	0	17	365	0
	D	0	**144**	**92**	499

投票モデルにおける誤分類率の振動が繰り返される．

以上より，複合分類木モデルは単独の分類木より高い識別精度を示し，また重みつき投票により，識別不能による精度悪化を防止することが確認された．

3. 顧客スコアリング

次に，ある衣料・雑貨販売会社（以下B社）の通信販売カタログ会員からの受注履歴データをもとにして，将来の購入可能性の高い順に顧客をランク付けする顧客スコアリング問題を考える．

一般に通信販売カタログに対する顧客の反応率は低く（B社の場合は10%程度），高ROIを得るためには反応率の高い顧客に限定したカタログ送付を行う必要がある．実際に米大手消費財企業の約3分の2が，的を絞ったマーケティングを行うために，現在データベースを使用しているか作成中であることが知られている[6]．

3.1 利用データ

分析対象データはB社カタログ会員12,242名の，1997年12月から2001年5月までの取引が記録されている受注明細データに，顧客に関する情報と商品属性についての情報を付加したものである．会員には季節毎に年4回カタログが発行されるが，1998年9月以前の発送スケジュールは不定期的である．

本分析の目的は2001年9月に発行された秋号カタログで高レスポンスが期待される，上位1,000名の優良顧客リストを作成することである．B社の協力をもとにテストマーケティングを行い，作成したリストの有効性を検証する．

表3　入力属性

項目	属性
Key	[顧客ID]
In1-4	期別購入回数
In5-8	期別購入金額
In9-12	期別購入品目数
In13	購入回数合計
In14	購入金額合計
In15	購入品目数合計
In16	購入単価
In17	注文あたり商品数
In18	初回購入からの経過期間

受注明細データから，顧客IDをキーとする分析用データを作成する（表3）．その際カタログ発送スケジュールに従って，98年度秋冬期（以下98AW）から01年度春夏期（以下01SS）までの半年毎に顧客の購買行動を捉える．また，カタログ発送が不定期的である98年9月以前のデータも含め，合計値に関する属性と，単価などの比率を取る属性を作成する．

分析用データを構成する属性には，デモグラフィック・データとリレーションシップ・データの利用が考えられるが，本分析ではB社の顧客層が比較的均一であることと，欠損値の状況からリレーションシップデータのみを使用する．

3.2 分析方法

3.2.1 使用アルゴリズム

スコアリングモデル構築には，教師付き学習アルゴリズムとしてニューラル・ネットワーク，RBF（Radial Basis Function）ネットワーク，回帰木を利用する[5]．

教師信号は，予測すべき期間に顧客からの反応があれば1，反応がなければ0をそれぞれ割り当て，それ以前のリレーションシップ・データから，反応の有無を予測する回帰モデルを作成する．作成したモデルの出力値を購入可能性（確率）であるとみなして，その値をもとにして全顧客をランク付けする．

3.2.2 分析の流れ

通常，データマイニングで用いられるノンパラメトリックな学習手法では，手元のデータを学習用データとテスト用データに分割して，前者でモデルを作成し，後者で精度評価を行う．

しかしB社の事例では，分析時点で入手可能であるデータをもとにして半期先の顧客のレスポンスを予測することが目的である．本例では利用可能な最新期

間の一期手前のデータを，それ以前のデータから予測するモデルを作成し，そのモデルを最新期間のデータを予測するために，入力を一期ずらして適用することによって，モデルの精度評価を行うことにする．

また，作成した分析用データには18の入力属性が含まれているが，これらの中には重複する属性や，不必要な属性が含まれている可能性がある．以下の5種類の入力属性の組み合わせを考える．

・期別受注回数(4)
・期別受注金額(4)
・期別受注品目数(4)
・期別受注回数＋受注回数合計＋受注金額合計(6)
・全属性(18)

ただし括弧内は入力属性の総数である．これらの入力属性の組み合わせそれぞれについて，3種類の学習アルゴリズムを適用し，15種類の予測モデルについて，予測精度の比較を行う．

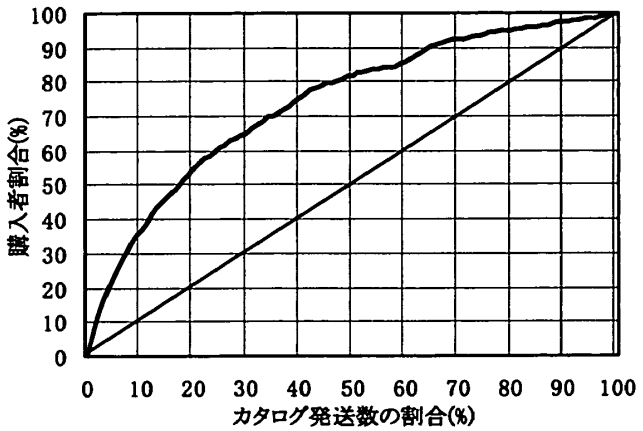

図5　ゲインチャート

3.2.3　予測精度の評価

本分析で利用するアルゴリズムには，AICなどの統計的モデル選択基準を直接的に適用することはできない．また実務への応用の観点からも，モデルの有効性を，そのROIで評価できることが望ましい．

予測全体の有効性を評価するためにゲインチャート(図5)を利用する．ゲインチャートは，高いレスポンス率が期待される順に顧客一覧を並べ替えた時に，予測上位x%の顧客について購入者数の累積割合をプロットしたものである．図中の直線は無作為に顧客を抽出してカタログを発送した場合の反応者数の期待値を示し，プロットされたカーブがどれだけ直線から離れているかがモデルの有効性を示す．

また，予測上位x名の顧客を抽出した場合のモデルの当てはまりの良さを定量的に評価するために，次に示すリフト率を利用する．

$$\text{リフト率}(x) = \frac{\text{予測上位}\ x\ \text{名の顧客のレスポンス率}}{\text{全顧客についてのレスポンス率}}$$

リフト率は無作為に顧客を抽出した場合に比べて，予測モデルを用いることにより何倍の効果が得られるかという倍率を示す．

3.3　結果

3.3.1　予測精度の比較

モデル検討用データ11,515名中，01SSにレスポンスのあった顧客は1,015名であった．すなわち全顧客についてのレスポンス率は8.8%である．表4にそ

表4　リフト率の比較

入力項目	手法※	1000名	2000名	3000名	4000名	5000名
期別受注回数(4)	NN	**3.88**	2.89	**2.37**	2.01	**1.81**
	RBF	3.51	2.63	2.28	1.94	1.75
	RT	2.59	2.66	2.22	1.96	1.70
期別受注金額(4)	NN	3.51	2.72	2.25	1.97	1.77
	RBF	3.30	2.51	2.15	1.95	1.76
	RT	3.13	2.53	2.16	1.88	1.63
期別受注品目数(4)	NN	2.96	2.44	2.11	1.86	1.66
	RBF	3.05	2.54	2.19	2.01	1.79
	RT	3.34	2.60	2.22	1.96	1.70
期別受注回数+受注回数合計+受注金額合計(6)	NN	3.74	2.82	2.33	1.98	1.77
	RBF	3.39	2.64	2.30	**2.02**	1.77
	RT	3.49	2.56	2.17	1.97	1.74
期別受注回数+期別受注金額+期別受注品目数+受注回数合計+受注金額合計+受注商品種類合計+受注単価+受注当り品目数+経過期間(18)	NN	3.65	**2.90**	2.32	1.99	1.79
	RBF	3.52	2.75	2.25	1.99	1.77
	RT	3.49	2.56	2.16	1.95	1.73

※NNはニューラル・ネットワーク，RBFはRBFネットワーク，RTは回帰木を示す．
表中の太字で表した数は最大値であることを示す．

れぞれのモデルについて，上位1,000名から5,000名までのリフト率の値を示した．

予測上位1,000名についてのリフト率が最も高かったのは，期別受注回数のみを入力属性とするニューラル・ネットワークによるモデルであった．11,515名中，予測上位1,000名（上位8.7%）の顧客のレスポンス率は34.2%であり，これは無作為にカタログを発送する場合の3.88倍の高レスポンスが得られることを示している．またこのモデルは，他のリフト率についても同様に高い値を示している．したがってテストマーケティングで用いる最適モデルとして採用する．

3.3.2 テストマーケティング

予測用データとして最新4期間を入力とする12,005名の顧客のデータを用いる．期別受注回数のみを入力属性とするニューラル・ネットワークによるモデルで顧客のスコアリングを行い，上位1,000名のリストを作成した．

2001年9月発行の秋号カタログに対する顧客のレスポンスを約2ヶ月間にわたって調査した．予測スコアを算出した12,005名の顧客の中で，実際にカタログが発送されたのは10,955名であった．また上位1,000名のリストの中で実際にカタログが発送されたのは997名であった．予測上位顧客のレスポンス率と全顧客のレスポンス率の比較を行い，モデルの有効性を検証した．

テストマーケティングの期間中に01年度秋号のカタログに反応したのは10,955名中573名であり，レスポンス率は5.2%であった．また予測上位顧客997名中194名の反応があり，レスポンス率は19.5%であった．したがって10,955名中，予測上位997名（上位9.1%）についてのリフト率は3.74である．モデル検証段階での11,515名中，予測上位1,000名（上位8.7%）の顧客のリフト率が3.88であったことと比較すると，ほぼ同程度の有効性が得られたといえる．

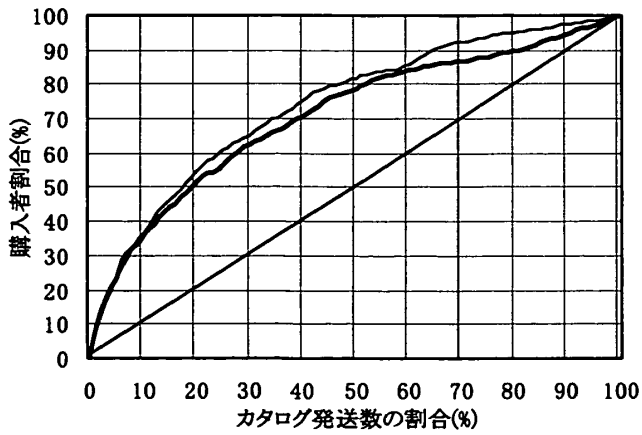

図6 ゲインチャートの比較．太線がモデル検証段階，細線がテストマーケティング結果を示す

モデル検証段階と，テストマーケティング結果のゲインチャートの比較を図6に示した．テストマーケティング結果では，モデル検証段階と比べてモデルの精度は低下しているが，予測上位20%の顧客が全反応者数の50%を占め，上位50%の顧客が全体の80%を占めていることが確認できる．

4. おわりに

本稿では，リレーションシップ・データへのデータマイニングの適用例として，顧客セグメント識別と顧客スコアリングの2つの事例について述べた．精度の高いモデルを構築するためには，二乗誤差などの要約された統計量のみを評価するのではなく，セグメント識別問題では混同行列，スコアリング問題ではゲインチャートにあたるような，モデルの効果を直接的に評価できる方法をあわせて利用すべきであるといえる．

謝辞 データの提供を頂きましたA社，B社に感謝いたします．

参考文献

[1] Bigus, J.: Data Mining with Neural networks, McGraw-Hill Companies, 1996.

[2] Breiman, L.: "Bagging predictors", *Machine Learning*, 24, 123-140, 1996.

[3] Dupret, G. and M. Koda: "Bootstrap re-sampling for unbalanced data in supervised learning", *European Journal of Operational Research*, 134, 141-156, 2001.

[4] Freund, Y. and R. Schapire: "A decision-theoretic generalization of on-line learning and an application to boosting", *Journal of Computer and System Sciences*, 55, 119-139, 1997.

[5] Hastie, T., R. Tibshirani, and J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2001.

[6] Kotler, P. and G. Armstrong: Marketing: An Introduction, Prentice-Hall, 1999.

[7] Peppers, D. and M. Rogers: The One to One Future: Building Relationships One Customer at a Time, Currency Doubleday, 1997.

[8] Weiss, W., C. Apte, F. Damerau, D. Johnson, F. Oles,

T. Goetz, and T. Hampp: "Maximizing Text-Mining Performance", *IEEE Intelligent Systems and their applications*, 14(4), 63-69, 1999.

[9] 後藤正輝, 村山一穂, 門間公志, 香田正人："データマイニングによる顧客スコアリング", 2002年度日本オペレーションズ・リサーチ学会春季研究発表会アブストラクト集, 144-145, 2002.

[10] 後藤正輝, 村山一穂, 門間公志, 香田正人："データマイニング手法によるスコアリングモデルの開発—リレーションシップ・データによる顧客のレスポンス予測—", *Direct Marketing Review*, 1, 19-32, 2002.

[11] 山部浩司, 八巻智, 山本良次, 香田正人："決定木を用いた複合学習モデルについて", 2000年度日本オペレーションズ・リサーチ学会秋季研究発表会アブストラクト集, 222-223, 2000.