データマイニングにおける二値データ解析
——決定木とロジスティック回帰分析

奥喜正　本村猛能　前鶴政和　内桶誠二

1. はじめに

顧客が新商品を購入したかどうか、新薬が患者に有効であったか否か、薬物投与によってマウスが死亡したかどうかという現象を扱う二値データ解析では(Cox and Snell, 1989)、第一選択として使用される統計モデルはロジスティック回帰分析である(Hosmer and Lemeshow, 1989)。ロジスティック回帰分析では、結果に対応する目的変数への説明変数の影響度をオッズ比で明示できるという利点があるが、複数の説明変数間の相互関係を吟味することは難しい。そこで、データマイニングの代表的手法の一つである、決定木(Decision Tree; Tree Models)を用いた二値データ解析が想定できる。決定木は判別、予測を目的とするデータマイニング手法で、経営手法のCRM(Customer Relationship Management)での優良顧客(Loyal Customers)の属性抽出などに活用されてきた(杉田・桜井, 2001)。CRM実現のための要素技術としてのデータマイニングでは、クラスター分析、決定木、相関分析などの諸統計データ解析手法を用いるが、決定木を二値データ解析に用いれば、目的変数に対する説明変数間の相互関係を階層的に捉えることが可能になる。

そこで、本稿では決定木の一手法のCHAID(CHi-square Automatic Interaction Detection)のアルゴリズムを、クロス集計表の独立性のためのカイ二乗検定との関連で現実データを分析しながら説明する。続いて目的変数が二値データであるデータセットに対して、ロジスティック回帰分析と決定木の両手法を併用して、両モデルの長所や短所を検討する。さらにCRMという経営戦略と決定木との関係も統計的データ解析の見地から併せて考察する。すなわち、CRM実現のためのデータマイニングにおける、決定木とロジスティック回帰分析の有効的な併用活用を模索することが本稿の目的である。

2−1. 決定木

決定木は自動交互作用検出法(Auto Interaction Detection; AID)から発展した手法であるので、交互作用の検出に適しており、クラスター分析とともに一般的なデータマイニング手法である。目的変数の注目する属性に関する重要な知識を、木構造によるルールの組み合わせで表現するもので、説明的かつ明示的な解を導き、優良顧客である確率など「予測」、「判別」を目的としたデータマイニング手法である。すなわち、決定木では一連の説明変数の中から、一つの適切な説明変数を選択してデータセットを、より均質な傾向をもつサブセットに分割することを繰り返して、目的変数に強く関連している説明変数や注目したいサブグループを発見することを目的とする。クラスター分析では、データ分割のためにすべての変数を利用するのに対して、決定木では各データ分割は一つの説明変数のみで分割される点が決定的に異なる。具体的な応用例では、顧客データベースから、決定木を利用して優良顧客の属性モデル化という、CRM への適用が挙げられる。なお、CRM を簡潔に説明すると、長期的に収益性のある顧客関係を構築維持するための統制のとれた経営戦略をいう。また、決定木手法としては CHAID が代表的である(Kass, 1980)。現在、広く利用されている決定木手法は CART(Classification and Regression Trees)であって(Breiman et al., 1984)、CHAID とはそのアルゴリズムが異なる(Paolo, 2003)。

次に、CHAID の解析手順を簡潔に示す。

1. 目的変数と各説明変数間で複数のクロス集計表を作成する.
2. クロス集計表にカイ二乗検定を実行する.
3. 最も有意である、すなわち、P-value が最小の説明変数を分割のための変数の候補とみなして、その説明変数によってデータセットを分割する.

このように CHAID ではクロス集計表に対する独立性のカイ二乗検定を繰り返してデータセットの分割を行い、停止条件が成立するまで分割を実行する。決定木の評価には、CHAID では不純度の測度(Impurity Measure)としてピアソンの X^2 検定統計量を使用するわけだが、CART のそれは Gini の分散測度(Gini Impurity; Gini Index)である。分割対象データが全データの 1% 未満というのが、分割停止条件の目安の一つである。

ところで、有意水準を α と設定した検定を n 回繰り返すと、第 1 種の誤りを犯す確率は

$$1-(1-\alpha)^n$$

となる。例えば、有意水準 5% の検定を 4 回実行すると、検定全体の有意水準は 18.5% になってしまう。そこで、有意水準を検定回数 n で割るという、Bonferroni の調整済

み確率をCHAIDでは採用して有意水準を総括的に調整する。この調整方法の着想は次式から理解できよう。αが十分に小さければ

$$1-(1-\alpha)^n \fallingdotseq 1-(1-n\alpha) = n\alpha$$

が成立する。クロス集計表では説明変数の数が多いときには、その組み合わせ数が多くなって集計表の分析結果が煩雑になってデータ構造の解読が困難になるが、CHAIDは解析結果を明示的な木構造にまとめて、ユーザーにとって視覚的にも理解しやすい解やルールを与えてくれる。

CHAIDによる決定木はルートノードから、各ノードで目的変数に対して最良の分割をもたらす説明変数によって排他的な分割が行われ、この手続きを再帰的に適用して停止条件に該当するまでデータセットの分割が継続される。いかなる基準でデータセットが分割されて、最終的にターミナルノードで、どのような属性をもつグループとして纏められるかという、分析結果を明示的に示せるという長所が決定木にはある。

参考までにCARTのアルゴリズムについても簡単に言及する。CARTアルゴリズムの基本は、枝刈り(Concept of Pruning)にある。まず十分に大きな木を生成し、それから必要以上の過学習にあたるリーフやノードの部分木を削除して(Pruning)、最善のサブツリーを枝刈りで発見しようとするのがCARTアルゴリズムの基本方針である。枝刈りは複雑度の増加を考慮するCost-complexity Pruning基準(Breiman, 1984)に従って行われる。CARTアルゴリズムの特徴は、データセットを常に二分岐させて成長させること、分割において統計的検定は利用しないこと、木を成長させてから枝刈りを行うことにある。

2. CHAIDについて

つぎに、タイタニック号乗客生死データに改良を加えた2101名から成る、改良データセットにSPSSのAnswerTreeのCHAIDを適用しながら、そのアルゴリズムを説明する。簡潔に言うと、CHAIDとは独立性のためのカイ二乗検定のp-value値をデータセット分割規準に繰り返し使用して、カテゴリカルデータの決定木をつくるアルゴリズムといえる。ところで、本稿のデータセットでは目的変数は「乗船客の生死」である。説明変数には「性別」、「(大人か子供かという)年齢」、乗船した「等級」が選ばれており、このデータセットの変数はすべてカテゴリカル変数である。このデータは元来、生存・死亡という二値の事故データではあるが、仮に乗客の生死を優良顧客、不良顧客という顧客データに読み替えると、決定木のCRMへの応用が身近なものとして感じられ

るかもしれない。

さて、**分割ステップ1**では、乗船客生死という目的変数と、性別、年齢、等級という説明変数の間で3個のクロス集計表を作成し、独立性のためのカイ二乗検定を実行してそれぞれP-valueを求めた。3個の説明変数のなかで、

生死と性別のクロス表

度数

		性別		
		女性	男性	合計
生死	死亡	118	1302	1420
	生存	332	349	681
合計		450	1651	2101

生死と年齢のクロス表

度数

		年齢		
		子供	大人	合計
生死	死亡	51	1369	1420
	生存	53	628	681
合計		104	1997	2101

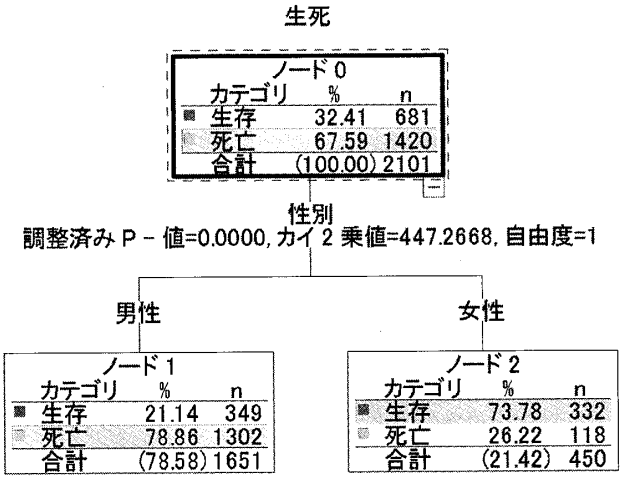
生死と等級のクロス表

度数

		等級				
		乗組員	一等	二等	三等	合計
生死	死亡	651	114	157	498	1420
	生存	203	194	115	169	681
合計		854	308	272	667	2101

	カイ二乗値	自由度	P-value
性別	447.2	1	2.83E－99
等級	187.9	3	1.76E－40
年齢	17.18	1	3.39E－05

一番小さいP-valueの説明変数をデータセット分割のための基準変数とするのだから、説明変数「性別」がこの分割ステップの基準変数に選ばれた。この分割ステップで、女性の方が圧倒的に生存率の高いことが判明した。CRM的に解釈すると女性の方が優良顧客である確率が高いことが期待された。



CHAIDでは不純度測度のピアソンのX^2検定統計量のP－valueで決定木を評価するが、参考までに、このステップのデータ分割を相互情報量(Mutual Information)で評価してみる。

分割前エントロピー：$-\sum_{i=1}^{2} p_i \log_2 p_i = 0.9086$

分割後エントロピー：(男性比率)x(男性エントロピー)＋(女性比率)x(女性エントロピー)

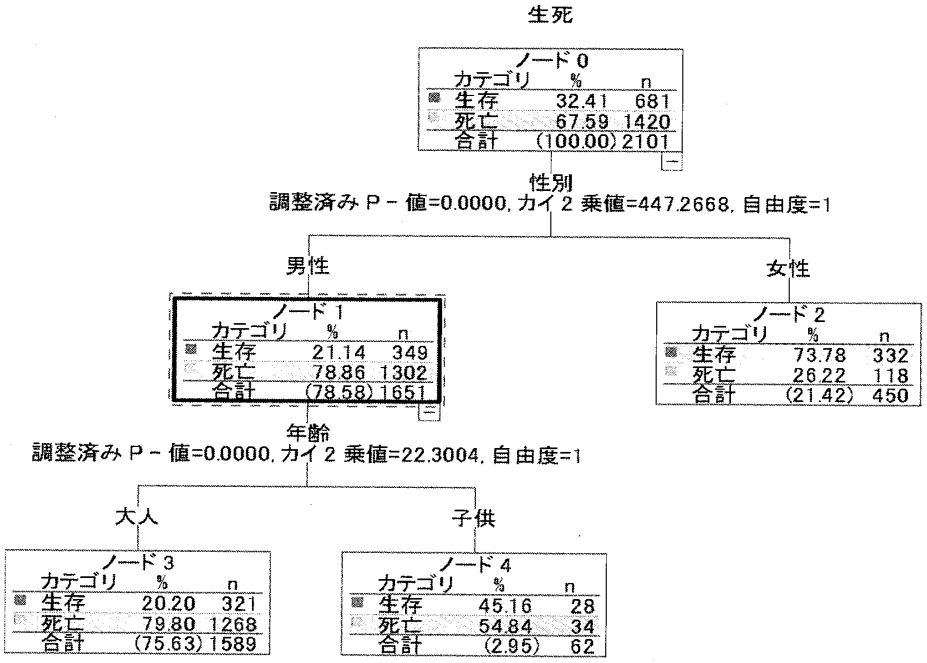
$$= 0.786 \times 0.7470 + 0.214 \times 0.834 = 0.7657$$

よって、相互情報量はエントロピー(Entropy Impurity)の減少量であるので、この相互情報量値は0.9086－0.7657＝0.143となったが、データ分割は相互情報量が大きくなる変数ですべきことになっている(福田他，2001)。

分割ステップ2に進む。男性では、男性と年齢、等級の二個のクロス集計表が作成可能である。

男性		自由度	カイ二乗値	P-value
	年齢	1	22.3	2.33E－06
	等級	3	29.05	5.72E－06

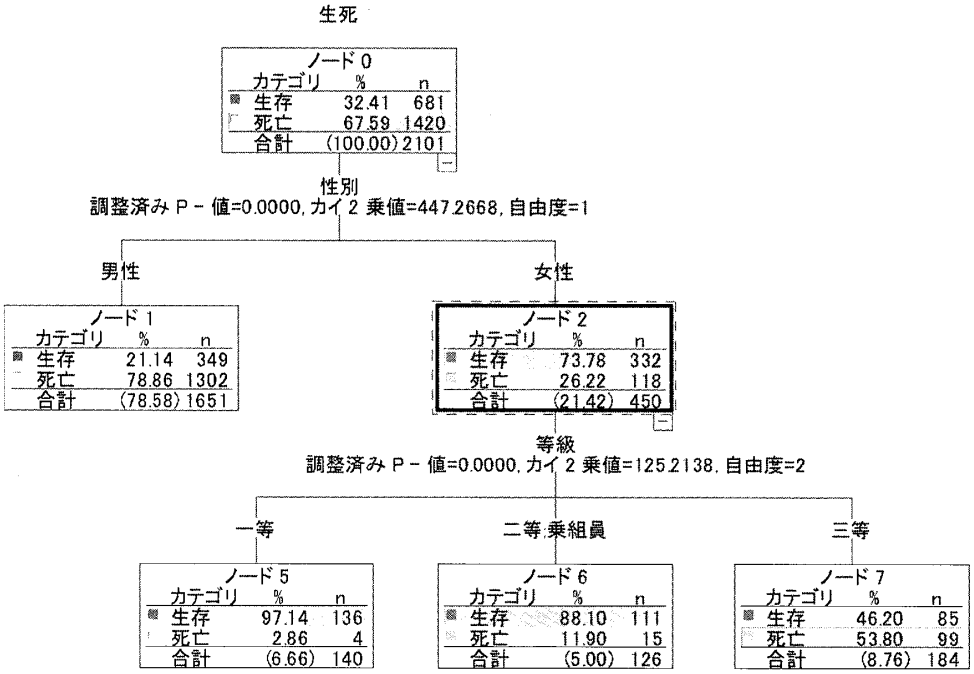
男性サブセット分割では、分割のための説明変数には「年齢」がここでは採用された。



この分割結果で、男性では子供のほうが生存率の高いことが判明した。優良顧客に例えれば、男性では子供の方が優良顧客である確率が高いというルールが成立することになる。

女性		自由度	カイ二乗値	P-value
	年齢	1	4.845	0.027726395
	等級	3	125.28	5.62472E－27

女性サブセットの第二分割では、分割の説明変数にはP-valueの値から「等級」のみが5%有意な変数として採用された。なお、「年齢」のP-valueは0.027(＞0.05／2＝0.025)であるので、5%有意な説明変数とはいえない。女性では高い等級に乗船した乗客ほど生存率が高いことが判明した。優良顧客の例に沿って述べれば、女性では高い等級に乗船する客ほど優良顧客になる確率が高いというルールが作成されたことになる。



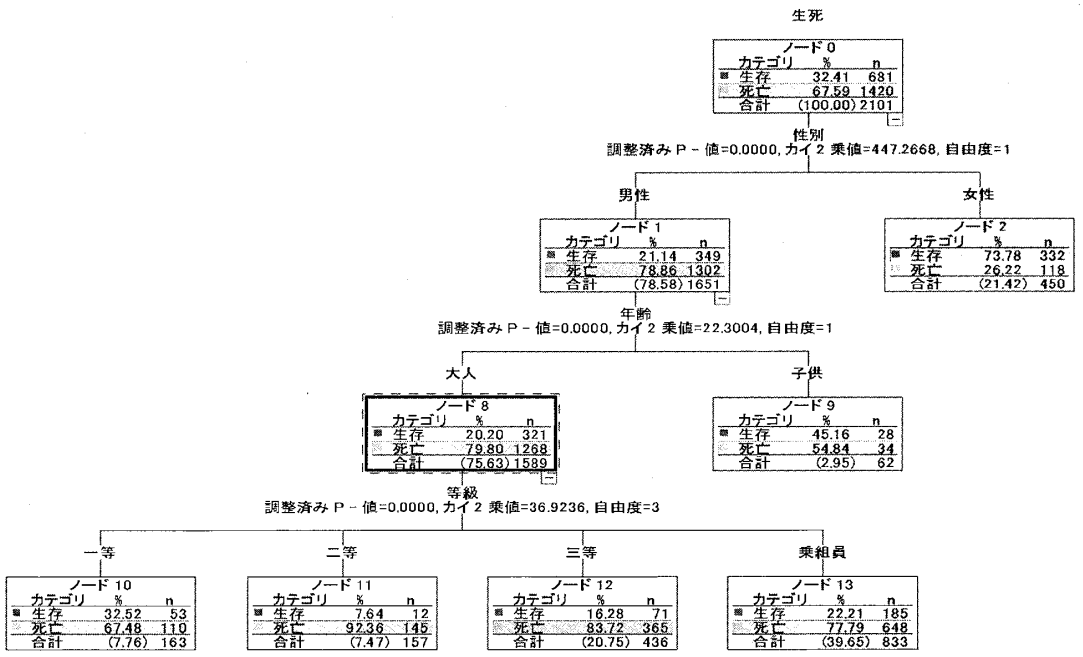
分割ステップ3で男性サブセットの第三分割は、「大人」の層で「等級」が、分割のための説明変数として採用された。子供のサブセットでは、データが希薄(Sparseness)であるのでこれ以上は分割しない。ここで着目すべきことは、二等乗船客より三等乗船客のほうが成人男性では生存率が高かったという事実である。

生死と等級と年齢のクロス表

度数

性別	年齢			等級				合格
				乗組員	一等	二等	三等	
男性	子供	生死	死亡		0	0	34	34
			生存		5	10	13	28
		合格			5	10	47	62
	大人	生死	死亡	648	110	145	365	1268
			生存	185	53	12	71	321
		合格		833	163	157	436	1589

男性		カイ二乗値	自由度	P-value
	子供	24.027	2	0.000
	大人	36.924	3	0.000



一般に説明変数の数が多くなるときクロス集計表による分析だけでは、変数の組み合わせ数が多くなって、数多なクロス集計結果の煩雑さによって的確な判別が困難になる。このようなときにCHAIDが与える木構造は、ユーザーにとって解釈しやすい有効な解を与える。

3. 分析方法と結果

本稿の解析データは、タイタニック号乗船客2201名の生死データに、100名を無作為抽出削除するという改良を加えた2101名から成る改良データである。二値データにSPSSのAnswerTree(SPSS, 2001)のCHAIDを有意水準を5%に設定して適用するとともに(図1)、同時にロジスティック回帰分析でも解析した(表1)。このデータセットでは目的変数は「乗客の生死」で、説明変数には「性別」、「年齢」、乗船した「等級」が選ばれており、本稿のデータセットでは変数がすべてカテゴリカル変数であった。さらに、決定木手法の一つである、CARTでも分析を試みた(図3)。

表1.改良タイタニック号乗客生死データに対するロジスティック回帰分析結果

方程式中の変数

		B	標準誤差	Wald	自由度	有意確率	Exp(B)
ステップ1[a]	等級	−.323	.048	45.785	1	.000	.724
	年齢	−.947	.252	14.061	1	.000	.388
	性別(1)	2.647	.137	373.635	1	.000	14.106
	定数	−.075	.263	.082	1	.775	.928

a. ステップ1：投入された変数等級，年齢，性別

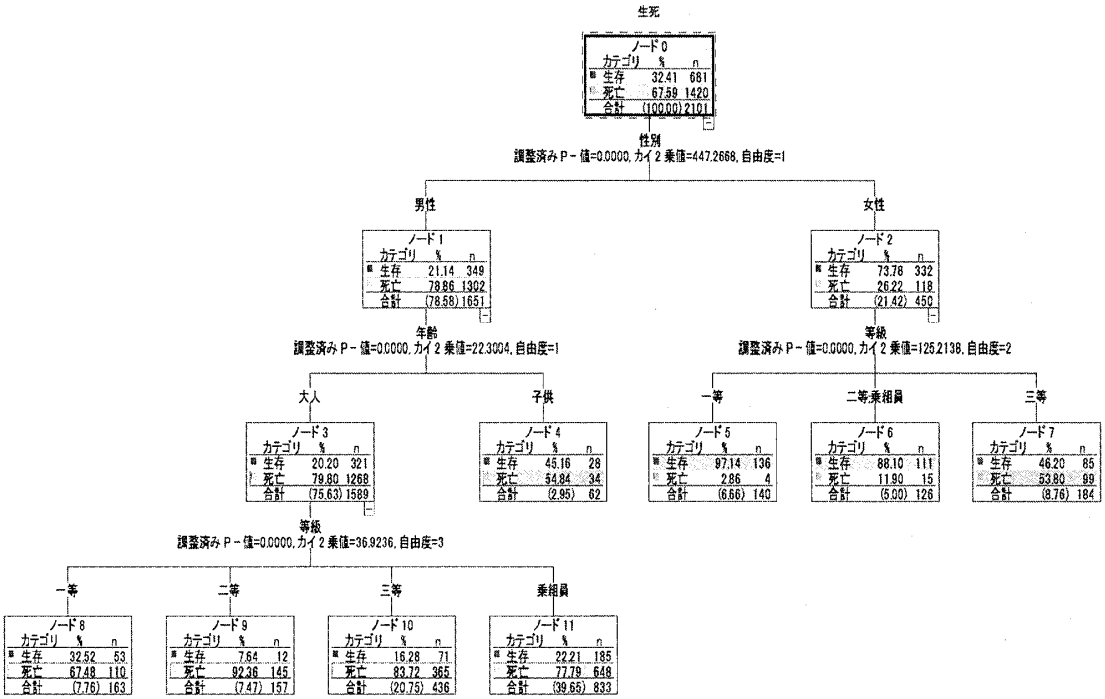


図1. 改良タイタニック号乗客生死データに関するCHAIDによる決定木.

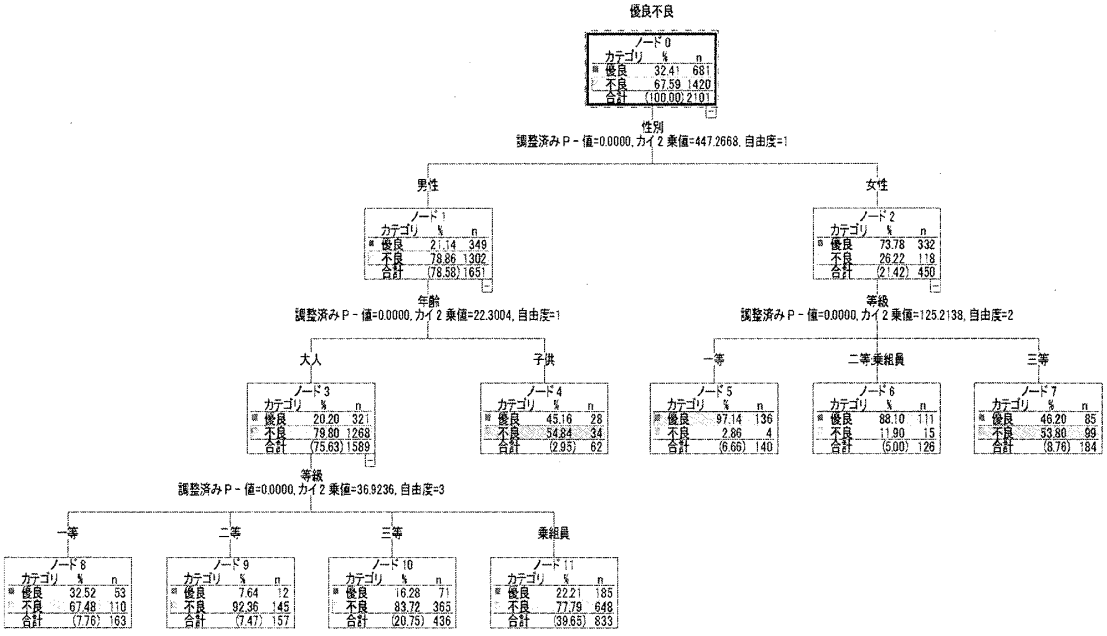


図2. 改良タイタニック号の優良顧客データに関するCHAIDによる決定木.

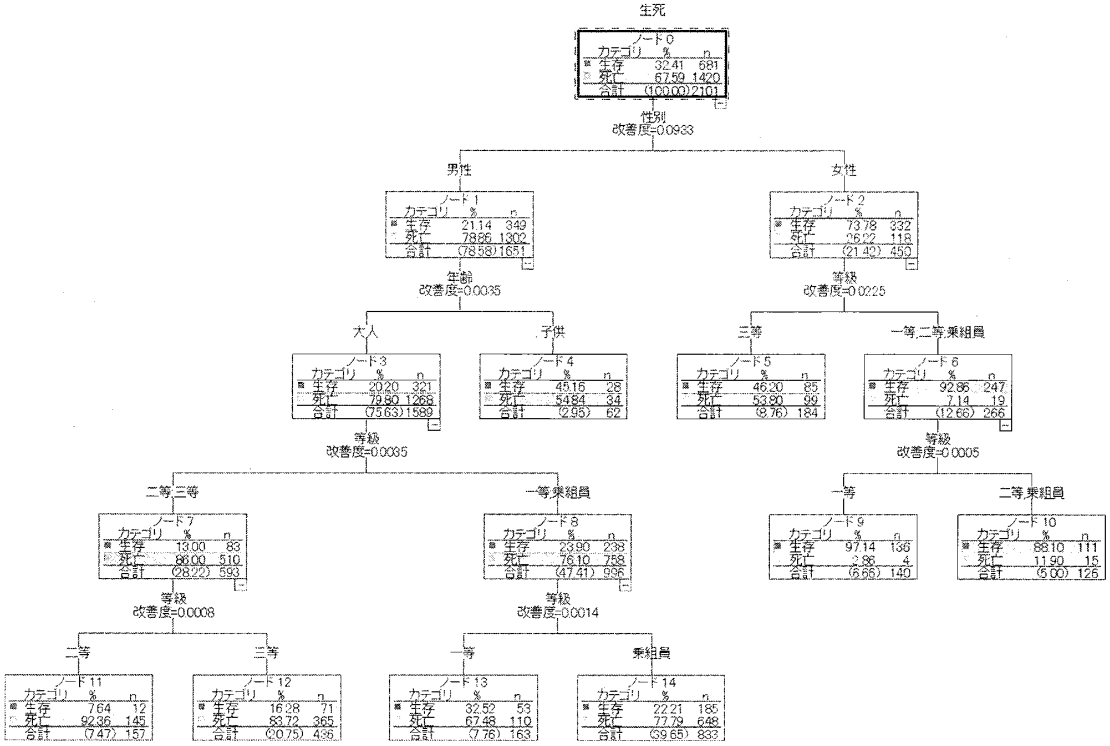


図3. 改良タイタニック号乗客生死データに関するCARTによる決定木.

ロジスティック回帰分析では、男性の死亡危険度が女性の 14.1 倍というように相対危険度の近似値であるオッズ比が表示されるが、三個の説明変数間の相互関係や階層関係は図解できない。これに対して、決定木では性別が生存に影響を与える最重要な要因であることが示され、続いて男性では年齢、等級という要因が明示された。女性でも等級が生死に影響を与える重要な要因であることが階層的に示された。決定木のターミナルノードを見ると、成人男性では、一等乗客は生存率 32.5%、二等乗客は 7.64%、三等乗客は 16.3% で、二等乗客よりも三等乗客のほうが、生存率が高いことが判明した(図 1, 2)。他方、女性の生存率は、一等では 97.1%、二等乗客は 88.1%、三等乗客では 46.2% であって、乗船した等級が高い程、女性の場合では生存率が高いことが分かった(図 1, 3)。

ルートノードでの生存率は 32.4% であったが、ターミナルノードの一つの女性一等乗船客では、生存率は 97.1% まで上昇した(図 1, 3)。CRM の例に喩えてみると図 2 のように生死データを顧客データに読み替えた例では、一等、二等女性乗客 266 名では、生存率の読み替えの優良顧客(Loyal Customers)である優良率は 92.9% となった。しかるに女性一・二等乗客セグメントは優良顧客になる確率が極めて高いと解釈でき、このセグメントを優良顧客として抽出すれば効率的な CRM 活動ができることが期待され、CRM における決定木の活用方法が窺える。このセグメントを優良顧客としてデータベースから抽出するためには、関係データベースを操作するための世界標準言語である SQL(Structured Query Language)の記述は以下のようになる。

```
／＊ノード 5, 6＊／
    SELECT＊FROM＜TABLE＞
    WHERE (性別＝0)　AND　((等級＝1)OR(等級＝2 OR 等級＝0))；
```

この優良顧客抽出によって、例えばダイレクトメールのコストは、266／2101＝0.127(12.7%)に抑えられるのに対して、メール反応数は 244／681＝0.358(35.8%)確保できるので、決定木による顧客抽出、顧客絞込みで、対コスト利益率は 2.82 倍改善されることが予測できる。全体を対象にする営業活動よりも特定のセグメントを対象にした方が、ダイレクトメールなど営業コストを大幅に削減でき、ダイレクトメール反応率向上で総利益もそれほど減らないという効率的経営が期待できる。

4. 考察

決定木とロジスティック回帰分析を併用してデータ解析を実行することで、データが内在的にもつ含蓄やデータ構造を深く解釈できることを本稿にて実証できた。

ところで、統計的データ解析の観点からCRM(Customer Relationship Management)という経営理念を捉えてみる。CRMとは顧客の価値観を充足させ続けることで、顧客にとって必要とされる企業との関係の構築と維持を継続する経営手法を一般的には指していて、マーケティング活動と見做すべきではない。一回だけの顧客購買を目的とせずに、優良顧客との関係維持強化に主眼をおいたうえで、顧客シェアと顧客維持率の向上を通じて、顧客生涯価値(Life Time Value; LTV)の最大化を目指し、顧客離反(Churn)に注意する経営戦略である。ここで、顧客生涯価値とは簡潔に言えば、顧客に対して継続的に自社の顧客であり続けてもらうということである。CRMの目的は顧客生涯価値の最大化であり、その戦略は顧客ロイヤリティの形成維持、その戦術は顧客との関係の形成維持のためのコミュニケーション強化である(藤田, 2001)。マーケット規模の膨張が望めない現在は顧客シェア、あるいは顧客内シェアを大きくすること、すなわち、特にロイヤリティ形成がなされている優良顧客をターゲットに自社製品を、顧客離反を防いで継続的に一人一人の顧客に多く売ることをCRMでは目指すのである。自社にとっての優良顧客に十分な顧客満足度(CS; Customer Satisfaction)を提供することで、長期的に顧客一人からの収益を向上させる経営情報システム活動とも見做せる。顧客生涯価値の最大化、顧客内シェアの拡大、顧客データ重視による顧客タイプの分類と優良顧客へのサービス付与、通常の販売促進コストによる既存顧客の維持、顧客ロイヤリティ形成促進と顧客離反防止、顧客を個客と見做すというコミュニケーション重視姿勢という諸項目がCRMという経営戦略を実施するに際した重要な基本方針である。

現実的にはCRM実施の第一ステップは、優良顧客の属性抽出から始まるが、まさに優良顧客の判別において必要不可欠となるのがデータマイニングツールの決定木である。それと同時に、大規模顧客データベース構築が昨今、容易に実現可能になったことがCRMという経営戦略が諸企業でますます普及した要因とも思われる。データマイニングに基づくマーケティング戦略がCRMの基幹にある。なお、CRMとデータベースマーケティング、one to oneマーケティングは、ほぼ同義語とみなしてよい。

本稿の最後に、データマイニングと統計的データ解析の関係について言及する(Giudici, 2003)。統計的データ解析とデータマイニングの相違点の一つは、解析対象のデータ数であってデータマイニングでは最小でも1千件以上が要求される。一方、統計的データ解析での対象データ数は、近代統計学の祖、フィッシャー(Fisher R.A.)以来、基本的には実験によって得られるデータ、すなわち、実験データがその主たる対象であるので、精密に計測された少数のデータが対象となる。相違点の第二は、統計的データ解析の目的が、実験データ解析による仮説検証であるが、他方、データマイニングは別名KDD(Knowledge Discovery in Database)と呼ばれるように、大規模データベース

からの知識発見、あるいは、仮説やルールの発見がその目的と言える。統計解析ではデータ分析前に仮説が存在しているのが通常であり、その仮説が本当に成立しているかどうかを現実のデータを使用して判断すること、すなわち、仮説検証が統計的データ解析の役目である。ところで、統計的データ解析の分野にも、従来からデータマイニングに方向性が近い探索的データ解析(Exploratory Data Analysis; EDA)という立場があった(Tukey, 1977)。EDAの主な目的はデータの縮約、外れ値の発見にあるが、データマイニングと統計的データ解析には重複する部分もあるということである。大規模データベース構築がIT技術の発展により容易に実現可能になったことで、統計的データ解析もデータマイニングを包含しつつ発展していく状況に現在はあって、データの流れの上流から下流まで一貫して科学するデータサイエンスという枠組みを考慮することも肝要である(柴田, 2001)。

なお、決定木の応用分野は、糖尿病合併症発症要因の研究など疫学や臨床医学への応用例もあって(Miyaki et al., 2002)、マーケティング関連だけではないことを指摘しておきたい。データマイニングはCRMを実践するための不可欠なツールであり、その有効なるデータマイニング手法の一つ、決定木(Tree Models)を二値データ解析手法のロジスティック回帰分析と関連させて本稿では説明した。本稿では決定木のモデル検証(Validation)については言及しない(Hjorth, 1994)。決定木のモデル検証方法には、データを二分割する方法と交差検証法(Cross Validation)とがあるが、詳細な記述は次稿に譲りたい。

参考文献

Berry, M and Linoff, G.(2004). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. 2nd ed. Wiley, New York.

Breiman, L., Friedman J.H., Olshen, R.A. and Stone, C.J.(1984). Classification and Regression Trees. Belmont, Calif, Wadsworth.

Cox, D.R. and Snell, E.J.(1989). Analysis of Binary Data, 2nd ed. Chapman & Hall, London.

Giudici, P.(2003). Applied Data Mining: Statistical Methods for Business and Industry. Wiley, New York.

Hjorth, J.S.U.(1994). Computer Intensive Statistical Method. Chapman & Hall, London.

Hosmer, D.W. and Lemeshow, S.(1989). Applied Logistic Regression. Wiley, New York.

Kass, G.V.(1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics,* **29**, 119－127.

Miyaki, K., Takei, I., Watanabe K., Nakashima H., Watanabe K. and Omae K.(2002). Novel Statistical Classification Mode of Type 2 Diabetes Mellitus Patients for Tailor-made Prevention Using Data Mining Algorithm. *Journal of Epidemiology*, **12**, 243-248.

SPSS（2001）．Answer Tree 3.0 J User's Guide, SPSS Inc.
Tukey, J.W.（1977）．Exploratory Data Analysis. Addison-Wesley.
アクセンチュア・村山徹・三谷宏治（2001）．CRM 顧客はそこにいる．東洋経済新報社．
朝野熙彦（1998）．消費者行動の予測を目的としたマーケティング・セグメンテーション．
マーケティングサイエンス，6, 45－66.
SPSS 著，杉田善弘・桜井聡訳．（2001）．マーケティングのためのデータマイニング入門．
東洋経済新報社．
福田剛志・森本康彦・徳山毅（2001）．データマイニング．共立出版．
藤田憲一（2001）．図解よくわかる CRM．日刊工業新聞社．
柴田里程（2001）．データリテラシー．共立出版．