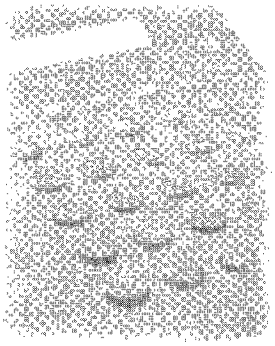
解説



金融ビジネスユースに適したデータマイニング手法: MBR

～融資申込み顧客の信用度の判定～

松本 和宏* ・ 前田 一穂** ・ 柳沼 義典*** （株）富士通研究所
新井 祥一**** ・ 佐藤 文信 （株）富士通

データマイニングにおける分類手法の1つであるMBR (Memory-Based Reasoning) について，そのアルゴリズムと金融与信問題への適用について解説する．まず，MBRについて一般概説を行い，富士通研究所独自に拡張を施したMBRの特長を紹介する．次に，MBRの金融与信問題への適用例を紹介し，MBRがビジネスユース向けの実用化に適したデータマイニング手法であることを議論する．

*matsumoto.ka-12@jp.fujitsu.com **maeda.kazuho@jp.fujitsu.com ***yaginuma@jp.fujitsu.com ****arai.shoichi@jp.fujitsu.com

分類手法とデータマイニング

近年，蓄積された大量のデータに埋没した価値ある情報を発見する「データマイニング」の技術が注目されている．

本稿では，データマイニングにおける分類手法の1つであるMBR (Memory-Based Reasoning：記憶に基づく推論) について，そのアルゴリズムと金融与信問題への応用について解説する．

分類（すなわち，ある対象があらかじめ与えられた分類（正常，事故など）のうちのいずれに属するかを判定する）手法としては，従来，多変量解析，ニューラルネットワーク[1]，決定木[2),3)]などによるものが提案されている．これらは，ある時点で利用可能なデータ（既知事例）をもとに，状況を抽象化して説明するモデルを作成し，分類はそのモデルを用いて高速に処理を行う手法である（モデルベースの手法）．モデルを用いた分類手法の背景には，データから複雑なノイズを除去し抽象化した後には，状況は単純に説明できるであろうという信念がある．さらに，計算機資源に関する制約に大きく依存することなく，分類を高速に処理したいという実用上の動機もある．

MBRは，モデルを事前に用意することなく，予測時にデータを探索して結果を算出する分類手法である．このため，MBRは，モデルベースの手法とはまったく異なったパラダイムに則った手法であるといわれる．MBRの提案の背景には，データマイニングとして議論されるような大量のデータを対象とする場合において，状況は非常に複雑で，抽象化しても簡単に説明できる保証はないというモデルベースの手法についての反省がある．さらに，計算機性能の向上，高速化アルゴリズムの研究開発により，分類時の計算時間が短縮されてきているという計算処理における進展もある．

本稿は前後半の2つからなる．前半では，MBRのアルゴリズムと我々による改良について解説する．後半では，金融与信問題を応用例として取り上げ，MBRを用いたデータマイニングシステムがビジネスユースに向けた実用化に適していることについて議論する．

MBRの特徴と性質

●MBRとは

Memory-Based Reasoning (MBR)[4),5)]とは分類手法の1つである．以下では，分類先を示す変数を目的変数，それ以外の変数を説明変数と呼ぶ．MBRは，目的変数が分かっている過去の事例（既知事例）から，分類したい事例

(テスト事例) に説明変数が類似した複数の事例を探索し，その (重み付き) 多数決により分類先を決定する手法である．探索する類似事例の数 (k) はあらかじめ設定しておく場合が多い．k=3である簡単な例を図-1に示す．テスト事例に類似した3個の事例を探索し，その多数決 (＋が2個，－が1個) により，＋と分類している．図-1では年齢，職業の2次元 (2つの説明変数) で表現したが，実際にはより多くの説明変数を用いて分類するのが一般的である．

MBRでは，説明変数ごとにどれだけ分類に有用かという評価を行い，既知事例とテスト事例間の距離計算の際に，より有用な説明変数を重要視することで正答率を上げようという試みがなされる．この評価値を本稿では影響度と呼ぶ．その例としては，cross-category feature importance (CCF) [4] などが挙げられる．

また，MBRは類似事例の目的変数値の分布 (目的変数分布) から各目的変数値の確信度を算出できる．さらに，テスト事例と類似事例間の距離重み付き目的変数分布から算出することで，より正確な確信度を計算する試みもなされ，この場合は影響度も考慮することができる．この確信度は，類似事例の目的変数分布から算出するため，直感的に理解しやすい．

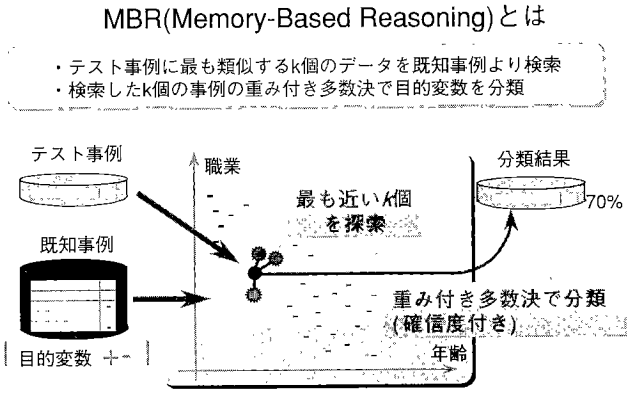


図-1 MBRの概要

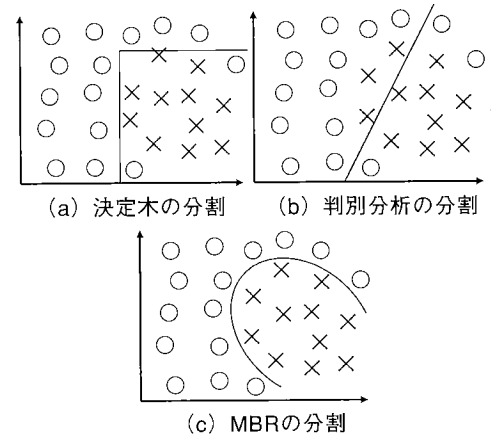


図-2 各分類手法の分割

●MBRの利点と欠点

MBRの利点として，①大量データに対して高い正答率が期待できる，②既知事例の増加・変化に柔軟に対応できる，が挙げられる．

①について図-2を使って説明する．簡単のために，事例は2つの説明変数を持ち，一方を縦軸，他方を横軸として表現するものとする．また，目的変数は○×の2値とする．この事例空間を決定木，判別分析，MBRの各手法で分類する場合を考える．図-2 (a) のように決定木は軸に平行な複数の直線で分割する．また，図-2 (b) のように，判別分析は任意の1本の直線か，あるいは二次曲線で分割する (図では直線の場合を示している)．それらに対して，十分な事例数を伴うMBRは，図-2 (c) のように任意の形状の分割を行うことができ，より高精度な分類が期待できる (ただし，実際には図のように滑らかな分割とはならない)．さらに，ニューラルネットワークも含めた他の手法が事例空間全体で1つのモデルの作成を行うのに対して，MBRはいわばテスト事例ごとに局所的なモデルを作っていることと同等といえ，その点でも高い精度を期待できる．

②は他の一般的な分類手法が，モデル化を行うため，既知事例の増加・変化への追随にモデルの再構築やパラメータの再設定を必要とするのに対し，MBRは複雑なモデル化を行わず事例をそのまま使用するため，これらの変化にも柔軟に対応できることを意味する．

一方，一般に，MBRの欠点として，①類似事例数の決定が困難，②予測に必要な計算量が大きいことが挙げられる．

類似事例数は正答率に対して大きな影響を持つパラメータであるが，最適な類似事例数を解析的に求める方法は知られていない．また，MBRはそのアルゴリズムの特性上各テスト事例ごとにその類似事例の探索を行わなければならないため，予測時の計算量は大きくなる．

●MBRの改良

我々は従来のMBRの欠点を克服し，実在のデータに適用するために改良を行ってきた．大きな改良点は以下の3点である．

①目的変数値の分布の偏りを考慮した，上述のCCFを

改良した新しい影響度計算方法（newCCF）

②ユーザの負担を軽減するための，実験的な方法による類似事例数の自動決定

③並列化による予測の高速化

これらの詳細は誌面の都合上割愛する．興味のある方は文献7），6）を参照されたい．

MBRの金融与信問題への適用

●ビジネスユースとしての実用化

金融与信問題への適用を例として，MBRがビジネスユースに適したデータマイニング手法であることを具体的に議論する．ここで，これからの論点についてあらかじめ整理する．

データマイニングシステムをビジネスユースとして実用化し，普及させるためには，分析者だけでなく，分析経験のない利用者によっても活用されることが重要である．これを実現するためには，次のような要件が解決される必要がある．

A）［運用面］データマイニングシステム導入後のメンテナンスが容易である（具体的には，データ蓄積に伴うデータ状況変化への対応に関して人的，経済的負担が少なくてすむ）．

B）［分析面］データマイニング手法がデータのどのような特徴を捉えて動作しているのかを知らなくても，利用者はシステムの出力情報を理解することができる．

C）［効果面］データマイニング手法を用いることによって得られる効果について，事前に定量的に推定できる．

これらの要件は，我々のデータ分析やデータマイニングシステムの顧客適用の経験に基づいている．しかし，これらはいずれも一般的な要件であるため，多くの分析者，利用者に同意されるものと想定している．他方，一般の汎用分析システムでは，分析者による利用を前提としているため，これらの要件についての対応が不十分である場合が多い．

●金融与信問題とは

近年，金融ビッグバンなどによる自由化の流れの中で，銀行業，クレジット金融業，消費者金融業などの境界があいまいとなり，競争が激化しているといわれている．たとえば，銀行業が堅実な顧客層に低金利で融資するのに対し，消費者金融業が堅実とはいえない顧客層にも高金利で融資するといった従来の図式がなり立たなくなってきており，これが競争激化の一因となっている．

ここで，銀行業などの貸し手にとって，借り手が融資をする対象としてどの程度，信用を供与することができるのかが関心事となる．与信とは，このような信用の供与のことをいう．

●分析課題

与信には大別して，初期与信と途上与信がある．

初期与信では，新規顧客の融資の申込みに対して，応諾すべきか，謝絶すべきかが分析課題となる．たとえば，初期与信について適切に分析することにより，貸し手は，正常に返済してくれる顧客を確保することができ（機会損失の回避），また，事故になりそうな顧客を謝絶することもできる（貸し倒れの回避）．

途上与信では，契約中の顧客に関する多様なリスクを調べることが分析課題となる．たとえば，契約更新を迎える顧客に対してどのような対応をとるか，各顧客の限度額をどのように増減するか，事故になりそうな顧客を早期にみつけることができるか，などである．

また，与信には，有担保・無担保，法人向け・個人向けなどの金融商品に応じて，さまざまな問題設定があり得る．これは，金融商品ごとに利用可能な顧客情報やリスクに対する考え方が異なるためである．

本稿では，このようにさまざまな金融与信問題のうちで各業界において共通に扱われる，無担保個人ローンの初期与信問題を例に取り上げる．

●データ蓄積に伴う変化への的確な対応

運用面に関して指摘した要件A）の観点から，MBRと他手法の比較について議論する．

金融業界での競合状況において，貸し手としては，借りたお金を誠実に返済してくれる安全な顧客に融資し，事故になりそうな顧客には融資しないといった信用リスクの判定が中心課題となる．特に現代は，景気サイクルや社会構造などに関してさまざまな状況変化が起こる時代である．このため，現状に的確に対応して，信用リスクを判定することが重要となる．現状への的確な対応の例としては，運用中に蓄積したデータを予測に活用し，より現状に即した予測をしたいという期待がある．

従来は，モデルベースの手法，すなわち，多変量解析，ニューラルネットワーク，決定木などの手法によりデータを抽象化したモデルを作成し，そのモデルを用いて顧客の信用リスクを予測するのが一般的であった．モデルベースの手法の場合，現状への的確な対応は，分析者が

会員番号	年収	勤続年数	役職	職種	自社内延滞回数	返済比率	信用ランク
A0001	400	8	一般職	事務	新規	0	正常
A0002	500	10	一般職	営業	なし	20	正常
A0003	500	9	一般職	営業	多	30	事故
A0004	600	5	一般職	営業	新規	10	正常
A0005	300	6	管理職	営業	少	30	事故

表-1 無担保個人ローンのデータ（既知事例）

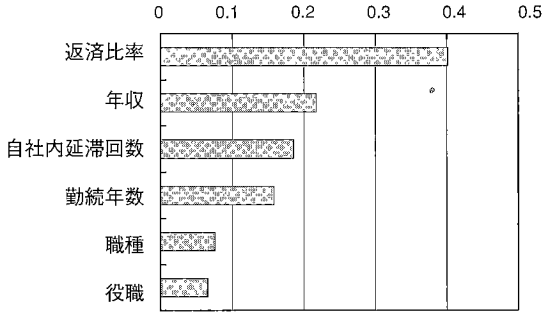


図-3 説明変数ごとの影響度

モデルを更新することによって行う．モデルの更新は，分析者が，分析パラメータの設定や説明変数の選択に関して最適な状況を探索することにより行う．分析者はこの工程に，自らのスキルとノウハウを注ぎ込み，多大な工数をかけて行うのが通常である．また，多変量解析におけるステップワイズ法を用いた説明変数の自動選択や，ニューラルネットワークにおける分析パラメータの自動探索などの機能によって，分析者のモデル作成時の負担を軽減させる工夫がなされる場合がある．この場合であっても，各手法に精通した分析者が，モデルを綿密に検証することは必須である．結果として，貸し手において，分析に関するコンサルティング費用や分析要員を十分に用意できない際には，現状に的確に追随していくことが難しくなる場合が多い．

MBRを用いる場合，このようなモデルの更新に伴う問題を回避することができる．これは，現状への的確な対応が，既知事例の差し替えにより実現されるためである（自動洗い替え機能）．自動洗い替え機能は，我々が独自に拡張を施した影響度計算方法の改良と類似事例数の自動決定を背景としている．すなわち，影響度計算方法に関して，目的変数への影響が少ない説明変数の寄与を小さくする改良を行い，既知事例の時間的変化に依存して有力な説明変数が変化する場合についても，自動的に対応がとれるように工夫した．また，類似事例数の決定を自動化し，利用者による分析パラメータの設定に関する負担を省略できるようにした．後に例を示すとおり，利用者は，影響度により説明されるデータの特徴や，MBRを用いて計算される効果などの結果が妥当かどうかを検証するだけでよい．結果として，自動洗い替えを比較的高頻度に行うことができるようになるため，より的確に現実に対応することができる．

このようにMBRを用いる方法では，モデルベースの手法と比較した場合に，分析者や利用者への負担が少なくてすむため，現実への的確な対応（MBRの場合には自動洗い替え）がより適切（高頻度）に行える利点がある．

●無担保個人ローンの初期与信問題

MBRの無担保個人ローンの初期与信問題への適用例を紹介し，分析面の要件B）と効果面の要件C）が，MBRを用いてどのように解決されるかについて解説する．

データ

表-1に，無担保個人ローンの初期与信問題に関する人工データの一部を示す．

ここでは，このデータに対してMBRを適用して得られる分析結果のイメージについて解説する．なお，以下に述べる分析結果は人工データを用いて得たものであるため，内容自体に意味はなく，また，特定の現実に対応するものではない．

表-1のデータにおける説明変数は，年収，勤続年数，役職，職種，自社内延滞回数，返済比率の6つであり，本稿の説明に十分である程度に少なめに設定してある．実際のデータでは，説明変数は数十変数程度ある場合が多い．

分析課題は，年収，勤続年数などの説明変数をもとに，信用ランク（目的変数：正常（契約履行）か事故（契約不履行）か）を予測し，正常であることのもっともらしさ（スコア：正常に関する確信度）を計算することである．ここで計算したスコアが，信用リスクの判定に用いる基準となる．スコアが，予測対象の顧客が正常であることに関して，既知事例の信用ランクの分布を前提とした確率となるところが特徴である．

会員番号	年収	勤続年数	役職	職種	自社内延滞回数	返済比率	信用ランク
X0001	1500	30	管理職	事務	少	5	?

表-2　予測対象の顧客のデータ（テスト事例）

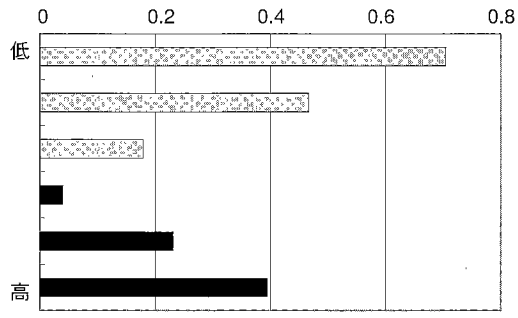


図-4　返済比率の影響度

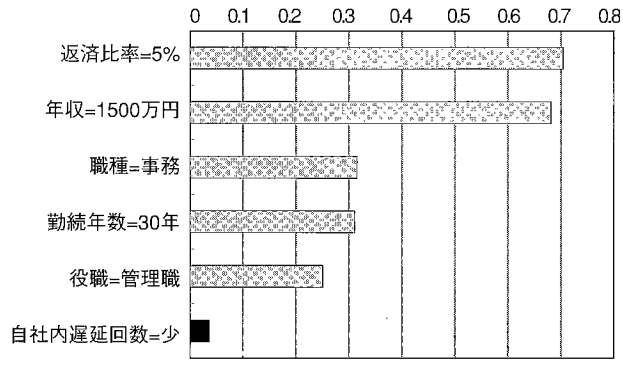


図-5　予測対象の顧客の影響度

変数ごとの影響度

図-3は，説明変数の各々が目的変数にどの程度影響しているかについて示したものである．縦軸に説明変数を示す．横軸には，説明変数の値ごとにnewCCFを用いて計算した影響度をもとに，説明変数ごとに平均を計算した値を示している．横軸の値の範囲は，0から1である．値が1に近いほど，その説明変数が目的変数に影響していることを示す．図-3の例では，「返済比率」「年収」「自社内延滞回数」…の順に説明変数の目的変数への影響が大きいことを示している．

変数値ごとの影響度

図-4は，説明変数の1つである「延滞比率」について，説明変数の値が目的変数にどの程度影響しているかについて示したものである．縦軸に説明変数値，横軸にnew CCFを用いて計算した影響度の値を示している．横軸の値の範囲は，0から1である．値が1に近いほど，その説明変数の値が目的変数に影響していることを示す．棒グラフの色は，グレーが正常の傾向，黒が事故の傾向を示している．ここで，正常・事故の傾向は，説明変数値に該当する顧客層の正常顧客の割合と，既知事例全体の正常顧客の割合との大小関係を基準に決定している．図-4の例では，「返済比率」が低ければ正常の傾向が，高ければ事故の傾向があることを示している．

スコアと予測対象の顧客の傾向

表-2は，予測対象の顧客に関するデータの例である．表-1のデータを既知事例とし，表-2のデータをテスト事例として設定し，MBRによりスコアを計算する．この顧客のスコアは，0.93点であった．スコアの範囲は0から1である．スコアが1に近いほど，正常の傾向が強いことを示す．

表-2と影響度を関連付けることにより，図-5のように予測対象の顧客の傾向について説明することができる．図-5の例では，「返済比率＝5%」，「年収＝1500万円」には強い正常の傾向があり，「自社内延滞回数＝少」には非常に弱い事故の傾向があることを示している．

これらが，分析面に関して指摘した要件B）についての答えである．すなわち，MBRでは，既知事例をどのように捉えているかについて，変数ごと，変数値ごとの影響度により，利用者に提示することができる．また，予測対象の顧客の傾向についても，その顧客に関する影響度として出力することができる．利用者は，影響度とデータに関する利用者自身の経験的知識を比較し，MBRによってデータがどのように捉えられているかを確認することができる．影響度と利用者の経験的知識が一致しない場合，利用者はその原因をデータ集計やOLAPなどの手段で調査することができる．newCCFによる影響度は，説明変数の値に該当する顧客についての正常・事故の件数をもとに計算されるため，データ集計により簡単に検証することが可能である．

他方，多変量解析やニューラルネットワークのようにモデルベースの手法で，モデルが数式で表される場合，データのどのような特徴に基づいて動作しているのかを説明するのは一般に簡単でない．特に，データ集計などの簡単な手段で，利用者の経験的知識と手法によるデータの捉え方の違いを検証できるようにするためには

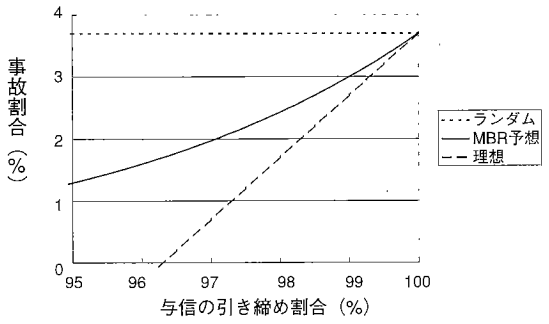


図-6　MBR予測による効果

かなりの工夫が必要である.

MBR予測による効果

MBR予測による効果は，テスト事例として，既知事例と同じ形式で，信用ランクが分かっている検証用のデータを用いて計算する．実際にMBRによる予測を行い，あるスコア以上の顧客の数とその範囲に含まれる事故顧客の数をもとに，図-6のMBR予測による効果をグラフ化することができる.

図-6において，横軸は，現状を基準にしてどの程度与信を引き締めるかについての値を示す．たとえば，横軸の値がX%であるとは，スコアの高い方からX%について応諾し，残りの100－X%について謝絶する場合に相当する．縦軸は，与信を横軸の値に従って引き締めた場合に，どの程度の事故割合となるかを推定した値を示す.

図-6の例では，95%に与信を引き締めた場合，事故割合をもともとの3.7%から1.3%に低下させることができる．これが，効果面に関して指摘した要件C）についての答えである.

ここでは，与信を引き締めることにより，事故割合を低下させ，貸し倒れを防ぐことができることについて説明した．他方，実際の運用においては，データマイニングを用いない経験的な方法では低いスコアしか計算されないが，MBRでは高スコアが計算される潜在的な優良顧客層を見出すことができるケースがしばしばである．このような顧客層に融資することにより，機会損失を防ぐことができる.

●今後の可能性（マーケティングへの応用）

ここまででは，MBRを用いてスコアを計算し，スコアを信用リスクの判定の基準として用いることについて解説した.

ここでは，MBRを用いたマーケティングの可能性について提案する．目的変数に融資利用額を設定して予測を行うことにより，予測対象の顧客がどの程度の金額を利用してくれそうかに関する予想額を計算することができる．この利用予想額が高い顧客層に対して，キャンペーンを実施することにより，貸し手は借り手の利用額増加による収入増を狙うことが可能である.

まとめ

データマイニング手法の1つであるMBRについて，一般アルゴリズムを紹介し，その利点と欠点，我々による独自の改良について簡単に説明した.

データマイニングシステムをビジネスユースとして実用化することに関して，運用面，分析面，効果面からの要件を説明した．MBRを用いて，これらの要件が満足されることについて，金融与信問題を例として具体的に示した.

すなわち，運用面として，MBRがモデルを用いない手法であるという観点から，データマイニングシステムの導入後に必要とされるメンテナンスに関する負担が少なくて済むことを背景に，現状への的確な対応が適切に行えることについて説明した.

分析面として，MBRの影響度を用いて，データのどのような特徴を捉えて動作しているかについて説明した.

効果面として，MBR予測を用いた場合の効果に関して，与信を引き締めることにより事故割合を低下させることができることを示し，従来の方法では見逃されていた潜在的な優良顧客層を見出せる可能性があることを説明した.

謝辞　本稿は，富士通（株）のスコアリングソリューションを導入していただいているお客様からのご意見，ご指導を参考に執筆しました．特記して感謝します.

参考文献
1) Rumelhart, E. et al.: Parallel Distributed Processing 1, The MIT Press (1986).
2) Quinlan, R.: Rulequest Research Home Page, http://www.rulequest.com/
3) Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J.: Classification and Regression Trees, C.J. Champion & Hall, International Thomson, Publishing (1984).
4) Stanfill, C. and Walz, D.: Toward Memory-Based Reasoning, Communications of ACM, Vol.29, No.12, pp.1213-1228 (1986).
5) 毛利隆夫: Nearest Neighbor法と記憶に基づく推論，人工知能学会誌，Vol.12, No.2, pp.188-195 (1997).
6) Maeda, K., Yaginuma, Y. et al.: A Parallel Implementation of Memory-Based Reasoning on Knowledge Discovery System, Parallel Computing Workshop'97 (PCW97), P1-I (1997).
7) 柳沼義典，前田一穂: 高性能データマイニング～知識発見システムの実現に向けて～，Parallel Computing Workshop '98 Japan (PCW98Japan), DC-1 (1998).

（平成13年6月1日受付）