ロジスティック分析でのステップワイズ法と決定木による属性選択法の実データをもちいた比較

Comparison with two attribute selection methods using actual data, stepwise procedure in logistic regression analysis and selection by decision tree

峰岸 達也　伊勢 昌幸　新美 礼彦　小西 修
Tatsuya Minegishi　Masayuki Ise　Ayahiko Niimi　Osamu Konishi
公立はこだて未来大学　株式会社　インテリジェントウェイブ
Future University Hakodate　INTELLIGENT WAVE INC

Abstract: The development of information processing technique enables us to collect, to process and to take advantage of a large amount of information. Especially it is big problem, to link information with management efficiency or strategy in business. The data mining is effective method in these fields. The data mining is method of discovery new knowledge and patterns from huge amounts of data. The data mining is used in business fields and having an enormous effect on the business management. But whether user can discover beneficial knowledge or patterns in data mining depends on whether user can analyze good data by good algorithm. Therefore data mining requires cleaning unsuitable data. In this paper, we propose the selection method that can select feature with major impact to detect the illegal credit card use from real credit card data.

1　はじめに

近年、大容量記憶媒体の低価格化、計算機処理能力の向上、情報通信技術の発展などからネットワーク社会においてデータの収集・活用が容易になったことをうけデータマイニングが注目されている。大量のデータの中から隠された知識や予想と異なるパターンを発見するデータマイニングは、大量のデータを扱う様々な分野で有益とされている。特にビジネスの分野においては、多くの企業が情報を活用していかに経営に優位に結び付けられるかが重要な課題となっている。

しかし、ネットワーク社会においてデータの膨大性は問題点のひとつであり、いかに有益な情報かそうでないかを判別しなければならない。これはデータマイニングにおいても同様である。大量のデータから知識をマイニングする際に知識発見には結びつかないようなデータが大量に存在してしまう場合があるかもしれない。このような場合ではそれらのデータがノイズとなってしまい、マイニングにおいて悪影響を及ぼしてしまう可能性がある。このような問題を回避するためにはデータ内のノイズの除去や、データのクリーニングが必要となってくる。したがって、データマイニングでは知識発見には結びつかないようなデータをどのように扱い、マイニングするかということが課題とされている。

本稿では実際のクレジットカード利用データに対してデータマイニングを適用している。クレジットカードもデータマイニングが用いられている分野の１つである。クレジットカードの分野ではデータマイニングを用いて近年問題となっているクレジットカードの不正利用の検出がなされている。詳しい説明は２章で示すがクレジットカード利用データは非常に膨大なものとなっており、その中から不正利用を検出するにはデータマイニングが適している。

しかし、実際にはクレジットカードの不正利用検出においても検出には結びつかないようなデータ (属性) をマイニングに用いていることが多い。そこで本研究では決定木を用いた属性選択法を提案することで、既存手法で用いられている属性選択法との違いを検証する。

以下では、まず２章で関連研究について述べ、３章では既存手法と提案手法で用いられているアルゴリズムについて述べる。４章では本研究で提案する手法の説明と、その手法を適用した不正利用検出システムについて述べ、５章でその検証実験と評価を行う。６章では実験の考察を述べ、最後に７章で本稿のまとめと今後の展開について述べる。

2　関連研究

2.1　不正利用検出におけるデータマイニング

クレジットカードの不正利用検出にデータマイニングが用いられる理由としては、近年のクレジットカー

ドの不正利用件数の増加と、その手口の巧妙化があげられる。以前では第三者がカードの取引をチェックし、疑わしい取引に対して警告を行うことで不正利用検出を行っていた。しかし、先にあげた理由によりすべてをリアルタイム、かつ人の手でチェックすることが困難になった。したがって、コンピュータを用いて膨大な数のデータをリアルタイムで処理・分析を行うことができるデータマイニングが用いられている。

また、クレジットカードの不正利用検出におけるデータマイニングのプロセスは以下の通りである。

① クレジットカード利用データの収集・蓄積
② データを処理・変換し、データの前処理を施す
③ 様々なアルゴリズムを用いた分析
④ 分析結果からの不正利用モデルの発見
⑤ 利用データをモデルに当てはめ、疑わしい取引に対して警告

しかし、1 日に数十万〜数百万件の取引データが発生するものにデータマイニングを用いたとしても、モニタリングで 1 日に対応できるのは数百〜数千件であるのが一般的である。したがって、検出件数は多くても全体の 1%程度という厳しい条件の中で、疑わしい取引データを効率的に検出しなければならない。さらに実際の不正利用率は全取引データに対して 0.02〜0.05%程と極めて低い割合であり、膨大な取引データの中から、極めて少ない不正利用を検出しなければならないということが課題とされている。

2.2 ACEPlus

本研究では株式会社インテリジェントウェイブ社が開発したクレジットカード不正検知システムである ACEPlus[1] を用いている。

ACEPlus とはクレジットカード取引データからスコアとルールを組み合わせた分析を行うことでクレジットカードの使用状況をリアルタイムで観察し、疑わしい使用に対して警告を行うことで最小限の被害にとどめるためのシステムである。

2.3 クレジットカード利用データ

今回は実際のクレジットカード取引データを使用した。このデータには 124 の属性があり、84 属性がクレジットカード会社が受信する取引情報や、保有している会員情報などの生データとなっており、残りの 40 属性はクレジットカード利用者の利用挙動から算出された独自のデータとなっていて、振舞いデータと呼んでいる。また属性のうちのひとつにその利用データが正常利用か不正利用かを判断するための属性が存在する。

クレジットカード利用データは非常に膨大であり、1 か月分のデータサイズは約 700MB である。また、2.1 で実際の不正利用率は 0.02〜0.05%程度と述べたが、このデータでは 0.5%程にサンプリングし直している。

3 用いたアルゴリズム

3.1 CS 変換

CS 変換 [2] とは株式会社 数理技研が開発したデータ変換手法であり、属性変換操作として 3.2 に示すロジスティック回帰分析前の下準備として用いられる。具体的には連続属性に対しては区間による離散化を行ってグループを形成し、該当するグループのロジットに変換する。また、離散属性に対してはそれぞれの属性値のロジットに変換する。ロジットに関しても 3.2 で述べる。

3.2 ロジスティック回帰分析

ロジスティック回帰分析とは ACEPlus で用いられている分析手法である。ある事象が起こる確率を予測する統計的分析手法の 1 つで、特に 2 値識別に優れている。一般的に回帰分析とは、目的変数 z と説明変数 x の間に式を当てはめ、目的変数が説明変数によってどれほど説明できているかを定量的に分析することである。線形的な合成関数を用いると

$$\mathrm{z} = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \tag{1}$$

ここでの β は回帰係数と呼ばれ、z を導くためにそれぞれの x に与えられた重みである。つまり x と z が最もよく当てはまるような値のことである。

事象を説明するために観測された説明変数群 $\mathrm{x} = \{x_1, \cdots, x_r\}$(r は説明変数の数を表す) を考える場合、$\mathrm{x} = \{x_1, \cdots, x_r\}$ である事象が発生する条件付き確率を $p(x)$ とする。ここで r 個の説明変数が目的変数に与える影響を式 (1) とした時、$p(x)$ を合成変数 z のロジスティック関数を用いて表した

$$p(x) = \frac{\exp(z)}{1 + \exp(z)} \tag{2}$$

がロジスティック回帰モデルである。式 (2) を変形すると、

$$z = \log \frac{p(x)}{1 - p(x)} = logitp(x) \tag{3}$$

となり目的変数を求める式にすることができる。この時、右辺のような式のことを logit(ロジット) と呼ぶ。

3.3 ステップワイズ法

ステップワイズ法とはロジスティック回帰モデルを検討する際の変数選択法であり、逐次選択法 [3] とも呼ばれる。ステップワイズ法でも変数増加法 (forward)、変数減少法 (backward)、変数増減法 (stepwise) の 3 種類がある。

変数増加法では有意な変数を 1 個ずつモデルに取り入れ、最もよいモデルが作成された部分で止める。逆に変数減少法では最初にすべての変数を取り込んでから、有意でない変数を 1 個ずつ除去していき、最もよいモデルが作成された部分で止める。

これらに対して変数増減法はいったん取り込んだ変数も組み合わせによっては有意でなくなることがあるので、その場合は除去する処理を組み込んだ方法である。

3.4 決定木

決定木とは、意思決定や物事の分類を多段階で繰り返し実行する場合、その多段の分岐過程を階層化して樹形図で表現したグラフ表現である。最もよく知られている決定木のツールとしては C4.5[4] があげられ、本研究でもこれを用いている。

C4.5 では情報量

$$info = -\sum p_i \log_2 p_i \tag{4}$$

p_i : i 番目の事象の生起確率

をもとに情報利得 Gain
$Gain$=(分割前の平均情報量
-分割後の平均情報量) (5)

が最大となる属性を順次選択して決定木を構築する。したがって、C4.5 では、事例の集合をすべての部分集合ができるだけ単一のクラスに属するような決定木を構成し、分類する。しかし、意味的な内容を無視して構造のみを捉えた非常に複雑な木となってしまうことが多く、誤り率が高くなってしまうことがある。そこで誤り率が最小になるように決定木の枝刈りが行われ、より単純でわかりやすい決定木を構築する。図 1 では枝刈り前の決定木と枝刈り後の決定木を表している。左の枝刈り前の決定木におけるノード C の部分木の誤り率がノード B における誤り率よりも低くなった場合に、右の枝刈り後の決定木のようにノード B がノード C の部分木に置き換えられる。

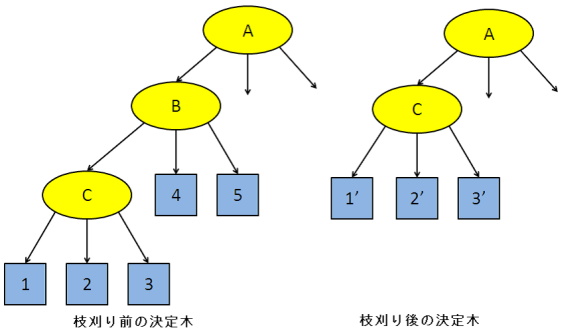

図 1: 枝刈り前の決定木と枝刈り後の決定木

4 提案する手法

3 章までに述べたように関連研究 [5] では多くの属性を分析に用いてしまっているということが言える。そこで本研究ではデータマイニングのプロセスであるデータの前処理部分に対しクレジットカード利用データから決定木を構築し、現れた属性を分析に用いるという属性選択法を提案する。そして、その手法を適用した不正利用検出を行い、既存システムの分析手法であるロジスティック回帰分析で用いられている属性選択法であるステップワイズ法との比較を行う。これにより既存研究で用いられている属性選択法との違いを比較し、選択された属性が分析にどのように影響しているのかを考察する。

また、決定木の全ノードから現われている全属性を分析対象とせず、ルート属性から一定の階層目までに現われている属性のみを分析対象とした場合の結果も比較することで決定木から属性選択を行う際の属性数の違いによる分析精度の違いも検証する。

本研究では提案する手法を適用したシステムを実装し、検証実験を行った。このシステムは 2.2 であげた既存システムである ACEPlus に提案する手法を適用させている。既存システムである ACEPlus と本研究で提案するシステムの違いを図 2 に示す。

ACEPlus では初めにデータのサンプリングを行い、CS 変換を行った後のロジスティック回帰分析に用いる属性の選択法としてステップワイズ法が用いられているのに対して、本研究で提案するシステムではロジスティック回帰分析を行う際にステップワイズ法ではなくサンプリングされたデータから決定木を構築し属性選択を行うことで分析に用いる属性を決定している。

両システムとも属性選択およびロジスティック回帰分析後は不正利用モデルを作成し、そのモデルにデータを当てはめることで不正利用の検出を行う。検証実

験の詳しい設定については 5 章に示す。

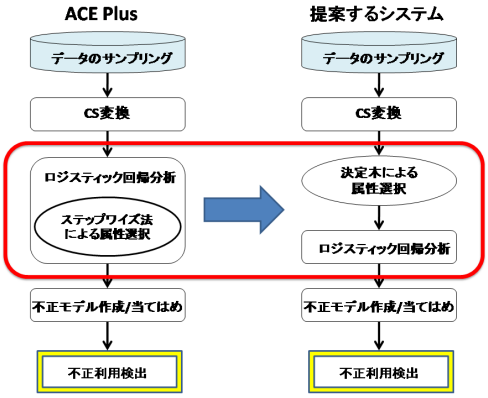

図 2: ACEPlus と提案するシステムの違い

5 実験と評価

ACEPlus に 4 章で示した提案する手法を適用し、不正利用検出の実験を行った。その結果を通常の ACEPlus で不正利用検出を行った場合の結果と比較した。

5.1 決定木の構築

実験で用いたデータは以下の通りである。

- データの属性数
  - 取引データ 57 属性と振舞いデータ 42 属性
- 不正利用のサンプリング比率 (3 通り)
  ① 実際の不正利用率である 0.02%
  ② ACEPlus のサンプリングの際に設定されている不正利用データの割合である 0.5%
  ③ 今回の実験で設定した 10%

通常、実際のクレジットカード利用データには 120 程の属性が存在するが、決定木構築に不向きなものや、不正モデル作成に結びつかないようなものは対象としていない。

これら 3 通りのデータを Weka[6] と呼ばれるデータマイニングツールソフトにおいて決定木構築アルゴリズムである C4.5 を基にした J4.8 と呼ばれるアルゴリズムによって決定木を構築した。

5.2 決定木構築の結果

①〜③のサンプリングデータを 5.1 であげた決定木構築に不向きな属性を削除して決定木を構築した。

①ではルート属性により終端ノードが 2 分されただけの決定木となってしまい、決定木の分類の成功率は 99%以上となったが、ほぼすべての不正利用データが分類失敗となってしまった。これは 0.02%という低い不正利用率により決定木自体の精度は高くなったものの、実際にはほぼすべての不正利用データを分類できていない結果となってしまった。

②では図 3 のような決定木となった。図にはクレジットカード利用データ内の生データである属性名が表示されているため解像度を下げている。

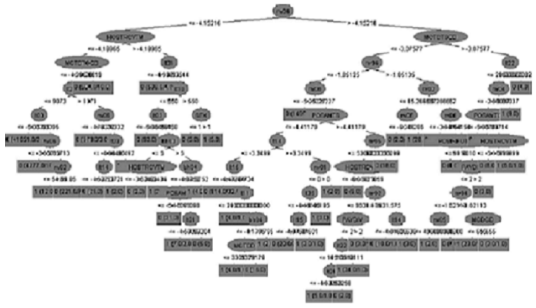

図 3: 構築した決定木②

終端ノード数は 51、決定木のサイズは 101 となった。また決定木の精度は 99.5%となったが、この場合も多くの不正利用データを分類に失敗してしまう結果となった。

③の場合では図 4 のような決定木となった。こちらも同様に解像度を下げている。

図 4: 構築した決定木③

終端ノード数は 611、決定木のサイズは 1,221 となった。また決定木の精度は 95.413%であり、今回は不正利用率が 10%という設定であったので、不正利用データの分類成功数も高かった。したがってこの場合の決定木を ACEPlus の分析において用いる属性を選択する際の決定木とし、サンプリング時に不正利用データの割合を 10%に固定したままデータを再度ランダムにサンプリングし直して同様の決定木を全部で 10 本構築し、それらを観察した。

10 本の決定木の終端ノード数、サイズ、精度に大きな違いは見られなかった。表 1 に 10 本の決定木の交差

検定からの分類における詳細な精度を示す。

表 1: 決定木の結果

		平均値	分散値	最大値	最小値
適合率	正常	0.9723	6.10E-07	0.973	0.971
	不正	0.8078	2.02E-05	0.815	0.801
再現率	正常	0.9798	5.60E-07	0.981	0.979
	不正	0.7526	3.94E-05	0.762	0.742

また、10 本の決定木すべてにおいてルート属性から 5 階層目まではほぼ同じ属性が出現していたので、そこまでを安定とみなした。

10 本の決定木に大きな違いは見られなかったので、その中から 1 つを選択し、その決定木に表れている全属性を不正利用検出の分析に用いる属性とした。また、安定とみなしたルート属性から 5 階層目までに現われている属性を分析対象とした場合も考えることで決定木から選択する属性数の違いによって不正利用検出の精度に違いがあるのかを考察する。

5.3 不正利用検出における実験

5.2 までに決定木から選択した属性を ACEPlus のロジスティック回帰分析部分に用いる属性として不正利用検出を行った結果と、ロジスティック回帰分析に用いる属性の選択法としてステップワイズ法を適用している通常の ACEPlus の処理で分析した場合の結果とを比較した。

1ヶ月間の総不正利用会員件数 2,173 に対し、決定木の全属性を用いた場合では 1,118 件、決定木の上位 5 階層目までに現われている属性を用いた場合では 903 件、通常の ACEPlus の分析を行った場合では 1,065 件を検出するという結果となった。また検出できなかった不正利用による被害額はそれぞれ 213,875,304 円、239,469,222 円、220,675,949 円であった。

6 考察

実験ではクレジットカード利用データから決定木を構築することで、不正利用検出に重要な属性を選択した。決定木構築の際に用いた 99 属性のうち 55 属性が現れた。特に振舞い属性は用いたほとんどが現れており、多くが決定木の上位に現われていた。決定木の分類の精度は 95%以上ということから、決定木の精度は高いと言えクレジットカード利用データから不正利用検出を行う際に決定木による属性選択が可能と言える。

また、決定木のサイズは 1,200 程であるが現われている属性は 55 属性と、決定木構築に用いた属性の半分しか現われていない。これは決定木には同じ属性が重複して現われており、特に上位の属性ほど多数現われているということになる。

既存システムで用いられているステップワイズ法で選ばれる属性と、決定木から選ばれる属性を比較すると、属性数ではステップワイズ法が 45 属性であるのに対し、決定木では 55 属性が選ばれている。また、全 40 の振る舞い属性のうち既存システムでは 31 の属性が選ばれているのに対し、決定木からでは 37 属性となり、振る舞い属性も決定木から多く選ばれていた。さらにステップワイズ法で選ばれた属性はほぼすべて決定木でも選ばれていることから、決定木によって選ばれた属性は不正利用検出に重要とされている属性を網羅していると言える。

次に決定木から属性選択を行い、その属性を用いて不正利用検出を行った結果と既存システムでの分析結果との比較を図 5 に示す。

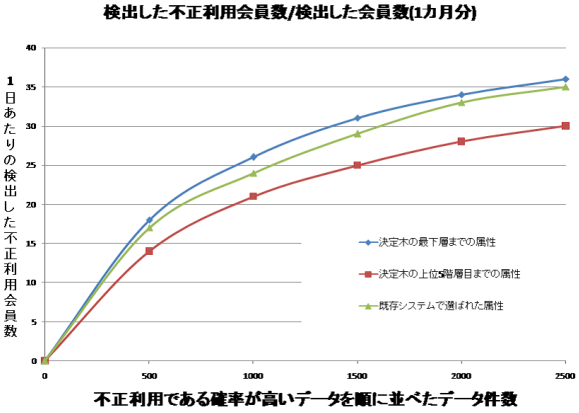

図 5: 実験結果の比較

決定木に現れたすべての属性を分析対象とした場合は既存システムでの分析の場合よりも多くの不正利用を検出できていた。また、決定木の上位 5 階層目までに現われた属性のみを分析に用いた場合が最も検出件数は少なくなった。これは既存システムで用いた属性数よりも多くの属性を用いた場合は検出精度も高くなり、用いた属性数が少ない場合は検出精度も低くなったと考えられる。また、表 2 に 1 か月間の総不正利用会員件数 2,173 件に対するそれぞれの場合の分析対象属性数と不正検知会員数、および被害額の関係を示す。ここで被害額とは検出できなかった不正利用による損失額のことである。決定木に現れたすべての属性を分析に用いた場合の不正検知会員数は 1,118 件であり、51%の不正検出率であった。また被害額は 213,875,304 円であ

り不正検知会員数、被害額ともに最も良い結果となった。これはロジスティック回帰分析に用いる属性数が最も多かったために、分析の精度が高くなったからであると考えられる。したがって、精度の面では既存システムで用いられているステップワイズ法よりも決定木から属性選択を行ったほうが優れているという結果がでた。

また、決定木の上位 5 階層目までに現われている属性のみを分析対象とした場合の結果を見ると、分析に用いている属性数は 13 属性と最も少ないにも関わらず、不正検知会員数は 903 件、検出率は 42%となった。これは最も精度の高かったすべての属性を用いた場合の 51%と比較しても、分析に用いた属性数は 1/4 ほどであるのに、検出率は 80%以上である。

表 2: 不正検知会員数と被害額

	対象属性数	不正検知会員数	被害額
すべての属性	55 属性	1,118 件	213,875,304 円
上位 5 階層目まで	13 属性	903 件	239,469,222 円
既存システム	45 属性	1,065 件	220,675,949 円

図 6 に分析対象属性数と検出精度の関係を示す。

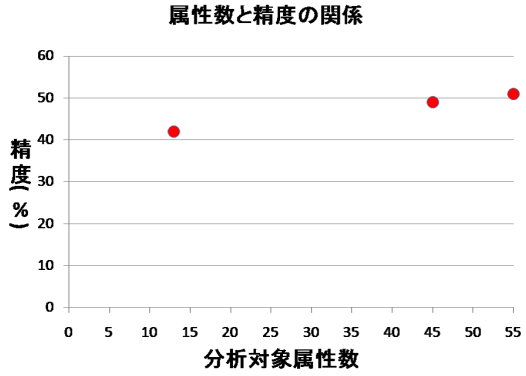

図 6: 属性数と精度の関係

図 6 では決定木の上位 5 階層目までに現れた 13 属性を分析対象とした場合までで精度は向上し、それ以降は分析対象属性数を増やしてもそれに伴って精度は向上していない。したがって、分析に用いる属性数を増やせば、それに応じて検出精度が向上するということではなく、決定木の上位に現われている属性が検出精度に強く影響していると考えられる。

7　まとめと今後の展開

本研究では属性選択という面で決定木から属性選択を行い、既存研究で用いられているステップワイズ法と比較を行った。決定木からの属性選択法ではステップワイズ法と比べ多くの属性を選択し、それらを分析に用いることで不正利用検出精度を向上させることができた。したがって、決定木からの属性選択は有効であるということが言える。しかし、実験結果から決定木のすべての属性を用いずとも、決定木の上位の属性のみを用いても不正利用検出精度は 80%を保っていた。これは分析に用いる属性数の他に、属性によって分析精度が左右されるということである。したがって、多くの属性を用いずとも分析に重要な属性を用いれば属性数の削減、すなわちデータサイズの縮小につながり、分析時間の短縮を考えることができる。

謝辞

株式会社インテリジェントウェイブの関係者の方々には、実験データの提供、また実験を進めていく上での様々なアドバイスなど細部にわたるご指導をいただきました。ここに感謝いたします。

参考文献

[1] (株) インテリジェントウェイブ: ACEPlus http://www.iwi.co.jp/product/ace.htm

[2] 数理技研：CS 変換 http://www.suri.co.jp/products/index05.html

[3] 丹後俊郎　山岡和枝　高木晴良: ロジスティック回帰分析-SAS を利用した統計解析の実際-, 朝倉書店, pp. 198-200, 1996

[4] J.R. キンラン 翻訳:古川康一: AI によるデータ解析, トッパン, pp. 17-25, 1995

[5] 都築学　新美礼彦　小西修: カーネル法による現象データマイニングの試み, 電子情報通信学会第 18 回データ工学ワークショップ, 第 5 回日本データベース学会 年次大会 (DEWS2007), 8pages(in Web), 2007

[6] Ian H. Witten・Eibe Frank: DATA MINING, MORGAN KAUFMANN PUBLISHERS, pp. 187-199, 2005

連絡先

公立はこだて未来大学　峰岸達也
E-mail: g2109043@fun.ac.jp