6N-7

多段決定木構築による属性選択法を用いたクレジットカードの不正利用検出システムの提案

峰岸　達也[1]　伊勢　昌幸[2]　新美　礼彦[3]　小西　修[1]

公立はこだて未来大学　システム情報科学部　複雑系科学科[1]

株式会社　インテリジェントウェイブ[2]

公立はこだて未来大学　システム情報科学部　情報アーキテクチャ学科[3]

1. はじめに

近年、現金を持ち歩かずに支払いができることやインターネットショッピングでの支払い、ポイントや割引サービスなどに代表されるクレジットカード利用のシーンの増加に伴い、クレジットカード発行枚数も増加している。しかし、その一方で偽造カードなどによる不正利用の犯罪が多発し、その被害額は決して少ないとは言えないのが現状である。

そこで本研究ではデータマイニングを用いて不正利用を検出するシステムの提案を行う。

2. 提案手法と関連研究における違い

本研究では株式会社インテリジェントウェイブ社(以下 IWI とする)が開発しているクレジットカード不正検知システムである ACE Plus[1]を用いている。このシステム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、怪しい使用に対して警告を行うことで最小限の被害に留めるためのシステムである。このシステムのデータサンプリング方法や分析方法を改良した研究がなされているが、多くのものがクレジットカード利用データに存在する多くの属性を分析に用いてしまっているものがほとんどである。

そこで本稿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[2]の部分において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、構築した決定木の上位に現れる属性を不正利用検出の分析に重要な属性と考え、分析対象とする属性数を減らす属性選択法を提案する。

またクレジットカードの不正利用率が非常に低いというなかでも、不正利用に関した決定木が構築できるよう決定木構築の際に分類に失敗したデータのみを用いて再度多段に決定木を構築し、属性選択を行うことを検討する。決定木を構築し属性選択後は ACE Plus で用いられている分析手法であるロジスティック回帰分析を行い、不正利用モデルを作成することで不正利用検出を行うシステムの提案をする。

3. 実験

ACE Plus はサンプリングしたクレジットカード利用データをロジスティック回帰分析し、モデルを作成し、そのモデルをもとに不正利用を検出する。

本研究では ACE Plus の工程であるデータのサンプリングからロジスティック回帰分析までの処理のサンプリング後に決定木構築により分析に用いる属性の選択を行うプロセスを追加する。

まず ACE Plus のサンプリング処理から 1 か月分のクレジットカード利用データをサンプリングし、CSV 形式のファイルにした。しかしこのままでは 700MB ほどとサイズが大きいためにデータ数を 50000 件ほどとして 30MB 程度のファイルに変換した。このデータはまず 1 か月分のデータの中から不正データをすべて取得し、その後、全不正データと正常データを合わせて 50000 件ほどになるよう正常データを無作為に抽出した。これによって決定木構築に用いるデータの割合は 10:1 程度のものになった。このようなファイルを 10 個作成し実験に用いた。

作成したデータファイルから決定木を構築した。今回はデータマイニングツールソフトである Weka[3]において決定木構築アルゴリズムである C4.5[4]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。作成したデータファイルを Weka で使用する際にいくつかの属性を削除している。これは ACE Plus 自体の分析から独自にスコアとして付加している属性や、海外の端末情報データなどで数値データの中にアルファベットなどの文字データが存在していて Weka でノイズとされて認識してくれないようなデータが多く混在している属性など決定木構築に不向きな属性である。最終的に決定木を構築するために用い

A proposal of abusing credit cards detecting systems using attribute selection method with multistage decision tree construction

1 Tatsuya Minegishi, Osamu Konishi・Future University Hakodate

2 Masayuki Ise・INTELLIGENT WAVE INC.

3 Ayahiko Niimi・Future University Hakodate

た属性数は113属性であった。これらの処理を行い 10 本の決定木を構築した。得られた決定木の上位の属性について分析を行った。C4.5 では情報利得による属性選択が行われるので、木の上位の属性は分類に大きな影響を与える属性であると考えられる。

4. 結果・考察

Weka で構築した決定木の一部を図1に示す。

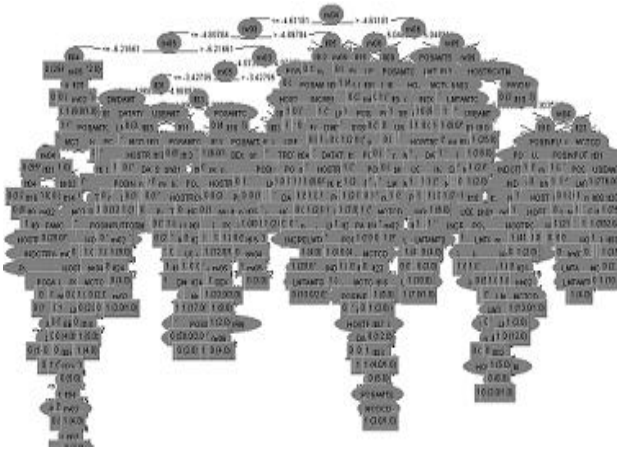

図1 構築した決定木

終端ノード数は約 600、決定木のサイズは 1200 ほどとなった。分類の成功率は平均で 95.68%であった。また、分類における詳細な精度を表1に示す。

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

このままの決定木ではサイズが大きすぎて、結果を考察することが困難であったため、10 本の決定木において上位のほうから出現している属性を比較した。その結果、10 本の決定木を比較したところ根ノードから5階層目まではほぼ同じ属性が現れていたので、そこまでを安定とみなし、5 階層目までに現れている属性を集計した。集計した属性を IWI が独自に行った分析により不正利用検出に強く関連している属性とみなされている属性と比較した。

IWI が行った分析とは、12 ヶ月分のデータを使用し、そのうちの1ヶ月分のデータをテストに用い、残りの 11 ヶ月分のデータを 1 月ごとに学習データとしてモデルを構築し、テストを行うという分析である。この分析結果から分析を行った 11 ヶ月中に何の属性が何ヶ月現れたのかを集計したものである。その結果、1 回以上現れた属性は 55 属性であった。

10 本の決定木に現れていた属性を IWI の分析結果であげられていた属性と比較したところ、55 属性中 38 属性が同じものであった。

また、10 本の決定木に出現していた属性の 23 属性は IWI の分析で 11 ヶ月すべてに出現していた属性と一致していた。

決定木に出現した属性は出現頻度の高いものにクレジットカード内に初めから存在する生データではなく、ACE Plus において分析に用いられているクレジットカード利用者の利用挙動から算出された ACE Plus 独自の属性が多く出現した。

また、この決定木において5階層目までで分類に失敗しているデータを用いて再度決定木を構築した。決定木を多段に構築することにより1回目の試行の際には出現していなかった属性が数は少ないが出現した。

5. おわりに

本研究ではクレジットカード利用データから決定木を構築し、属性の選択をおこない、既存システムである ACE Plus の分析に用いる属性数を減らすことを目的とした不正利用検出システムの提案をおこなった。今回の実験では決定木を構築し、分析に重要な属性を選択することはできた。しかし、今後の課題としては選択したデータのみを用いて ACE Plus の分析を行った場合に不正検出の精度においてどれほどの差があるのかを検証するための実験を行う必要がある。

謝辞

本研究・実験・論文の執筆を進めていくにあたり、実験データの提供や、様々な助言を下さった株式会社 インテリジェントウェイブの関係者方々に深く御礼申し上げます。

参考文献

[1] ACE Plus インテリジェントウェイブ http://www.iwi.co.jp/product/ace.htm
[2] 元田 浩・津本 周作・山口 高平・沼尾 正行、『データマイニングの基礎』、オーム社、P21～29、2006
[3] Ian H. Witten・Eibe Frank、『DATA MINING』、MORGAN KAUFMANN PUBLISHERS、P187～199・P365～425、2005
[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、P17～25、1995