ロジスティック回帰モデルを用いた不適切会計企業の早期発見

M2011MM021 堀山洋輔
指導教員：石崎文雄

1 はじめに

近年経済の不況に伴い, 上場企業の不適切会計の増加が目立ってきている. 東京商工リサーチ[1]が行っている上場企業の不適切会計調査の結果, 上場企業の不適切会計は2009 年度から徐々に増え続け, 2011 年度には前年度より 8 社増加の 32 社と過去最多を記録し, 投資家にとって大きな問題となっている. 不適切会計企業が増加していることから, 不適切会計を起こす企業を早期発見し警告することが近年の社会では求められている.

本研究では, 不適切会計を起こした/起こしていない上場企業の公開されている財務諸表と業績のデータを, Shih ら [1] や Shen ら [4] の研究を基に, ロジスティック回帰を用いたモデルで分析し, 分析結果から不適切会計企業の条件を割出す. 分析結果を基に, 不適切会計企業との条件一致率を%で表示し, 50%を超えていた場合は警告を表示し, 50%以下の場合は危険小を表示するモデルを実装する. また, その性能を評価するため, 既存の決定木・ニューラルネットワークによる警告モデルと真陽性率 (True Positive Rate)・偽陽性率 (False Positive Rate)・正確性 ($Accuracy$) において性能比較を行い, どの程度優れているのかを証明する.

2 不正検出の諸研究

不正検出手法を用いた研究を紹介していく. 表 1 は各研究の検出対象と使用している手法, 研究内で最も精度の高い手法をまとめたものである. また, ここでは本研究において比較するアルゴリズムの概要について説明する.

2.1 財務諸表不正検出手法

Shih ら [1] は, 財務諸表不正警告モデルにおける, ロジスティック回帰とバックプロパケーションニューラルネットワーク (BPNN) の性能比較を行っている. データセットは台湾の製造会社の不正を起こした会社 96 件, 起こしていない会社 192 件, 合計 288 件の財務諸表を使用しており, 確率分類において最適なカットオフ値によって性能比較を行っている.

Kirkos ら [2] は, 不正な財務諸表を検出する際における, 3 つのデータマイニング技術, 決定木 (DT), ニューラルネットワーク (NN), ベイジアンネットワーク (BN) の適用性を調査している. データセットはギリシャの製造会社の不正を起こした会社 38 件, 起こしていない会社 38 件, 合計 76 件の財務諸表を使用し, 10-fold cross validation で性能比較を行っている.

2.2 不正取引検出手法

Patil ら [3] は, 電子取引における不正において, ESMC アルゴリズムと決定木の性能比較を行っている. ここでは, 性能比較のための指標として, 真陽性率 (TPR), 偽陽性率 (FPR) と正確性 ($Accuracy$) を使用している.

Shen ら [4] は, クレジットカード取引 (2005 年〜2006 年の取引データ) において, 決定木, ニューラルネットワーク, ロジスティック回帰の性能比較をリフトチャート・デシル分析によって行っている.

2.3 本研究の方針

本研究では, データの入手の容易さの点から, データは不適切会計を起こした/起こしていない企業の財務諸表と業績のデータを使用する. また, 先行研究の技術の性能比較の観点から, ロジスティック回帰が不正検出の研究において多く用いられており, 他の手法と比較した精度も最も高いため, 不適切会計企業早期発見モデルとして最も優れていると推測できる. よって, ロジスティック回帰モデルを用いた不適切会計企業の早期発見モデルを実装し, 実装したモデルの性能評価を行うために, 既存の決定木, ニューラルネットワークによる警告モデルと性能比較を行う. 以下に, 性能比較を行うモデルの概要の説明を行う.

2.4 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に, 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 があり, 本研究でもこれを用いて性能比較を行う.

C4.5 では情報量 $info$,

$$info = \sum p_i log_2 p_i \tag{1}$$

をもとに情報利得 $Gain$,

$$Gain = (\text{分割前の平均情報量} - \text{分割後の平均情報量}) \tag{2}$$

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.

2.5 ニューラルネットワーク

ニューラルネットワークは, 人間の脳の神経回路の仕組みを模したモデルで, 教師信号（正解）の入力によって問題を最適化していく教師あり学習と, 教師信号を必要としない教師なし学習に分けられる. 明確な解答が用意される場合には教師あり学習が, データ・クラスタリングには教師なし学習が用いられる. 本研究では明確な答えが必要とされるため教師あり学習のバックプロパゲーションニューラルネットワークモデルを使用する. このモデルは入力層, 中間層, 及び, 出力層から成る多層のネットワークであり, 同じ層にあるユニットや下の層にあるユニットへ接続することはできない.

[1] http://www.tsr-net.co.jp/

表 1 不正検出の対象と精度

文献番号	不正検出対象	使用する手法	最も精度の高い手法
[1]	財務諸表 (製造会社)	ロジスティック回帰・BPNN	ロジスティック回帰
[2]	財務諸表 (製造会社)	DT, NN, BN	BN
[3]	電子商取引	ESMC アルゴリズム・DT	ESMC アルゴリズム
[4]	クレジットカード取引	DT・NN・ロジスティック回帰	ロジスティック回帰

3 提案するアルゴリズム

本研究で提案するアルゴリズムであるロジスティック回帰と実装するモデルの構築案を述べる.

3.1 ロジスティック回帰

ロジスティック回帰は, 事象の発生確率を予測する手法で, 予測結果が 0 から 1 の値の間を取る. 不適切会計の発生の有無 (Yes/No) のような 2 値しか取りえない値を従属変数の実績値として用い, 説明変数を用いてその発生確率を求めるという構造になっている. 一般にある現象の発生する確率を p, その現象の生起を説明するために観測された変数群 x=(x_1, x_2, $\cdots$, x_n) を説明変数とする場合, ロジスティック回帰は,

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)} \quad (3)$$

として求められる. ここで, β_0 は定数, β_n は x_n の回帰係数となる.

例として, 企業の不適切会計をおこす要因として, x_1 を業績, x_2 を棚卸資産, x_3 を総資産とし, これらをデータとして入力し, ロジスティック回帰分析により β_0〜β_3 を求める. また, 不適切会計発生の確率指標として, 分母に非発生確率, そして分子に発生確率をおいて算出したものをオッズと呼ぶ. オッズ比はその比率であり, ロジスティック回帰の分析結果である回帰係数の指数を取り, $\exp(\beta_n)$ で表される. この手法を基に, 不適切会計企業発生に影響を与える要因をオッズ比から求め, 求められた結果から, 不適切会計企業との条件一致率を求める不適切会計企業の早期発見モデルを実装する.

3.2 不適切会計早期発見モデル構築案

不適切会計企業早期発見モデル構築案について説明を行う. ここでは本研究で実装する不適切会計企業早期発見モデルの構築手順を (1)〜(5) で示し, 図 1 に表示する.
(1) まず準備として不適切会計を起こした/起こしていない企業の財務諸表と業績のデータを, WEB 上のサイトから収集しておく. 収集したデータのファイルは全て同一ディレクトリに保存し, C 言語のプログラムを読み込みやすくしておく.
(2) 収集したデータを, 統計分析ツール R を用いた, ロジスティック回帰モデルにより分析を行う.
(3) 分析結果を出力.
(4) 分析結果を基に不適切会計企業の条件を割り出す.
(5) 不適切会計企業の条件を参照し, 不適切会計企業との条件一致率を%で表示し, 50%を超える場合は危険度が高いと判断し警告を表示し, 50%以下の場合は危険小を表示する.

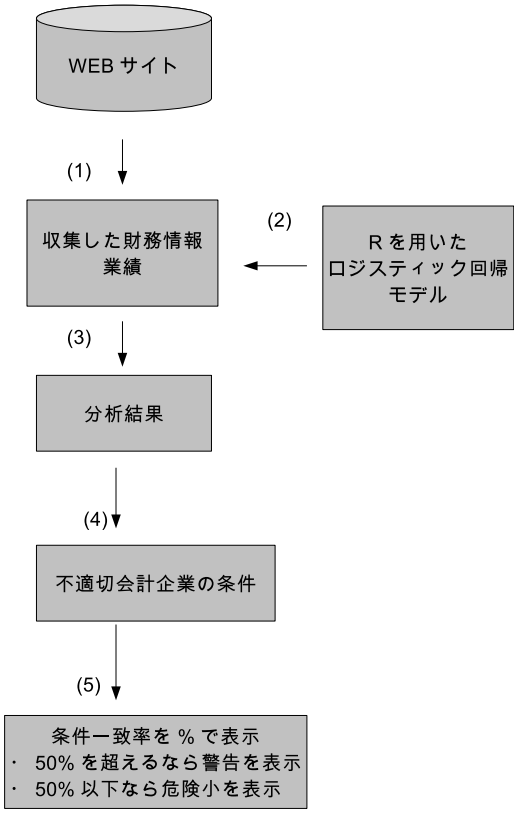

図 1 不適切会計企業早期発見モデル構築案

4 不適切会計企業早期発見モデルの実装

第 3 章で述べたモデルの機能を実現するために, 使用するデータセット, ロジスティック回帰モデル出力結果, 不適切会計企業早期発見モデルの実装について述べる.

4.1 使用するデータセット

本研究で使用するデータとして, 2011 年度に不適切会計を起こした上場企業 32 件, 起こしていない上場企業の 320 件, 合計 352 件の 2010 年から 2012 年までの財務諸表と業績のデータから, 3 年間の変動率を求めロジスティック回帰を行う. これは, 大企業と中小企業では値に大きな差があることから, 比較するのが困難なためである. ロジスティック回帰は統計分析ツール R を用いて結果を求め, それを基にオッズ比を求める. 求められたオッズ比から不適切会計を起こす企業に影響のある項目を調査し, 不適切会計企業の条件を割り出す. 表 2 は取得したデータと 3 年間の変動率から作成した不適切会計企業のデータセットの一例である. これらのデータは, 投資家のための銘柄

分析サイト「shares」[2]から入手し分析を行う．このサイトからデータを取得する理由として, 上場廃止した企業の財務諸表・業績のデータでも入手できること, 財務諸表・業績の 3 年分のデータを無料で入手できることがあげられる．使用する財務諸表と業績の項目は, 財務諸表が 現金等・売上債権 (単位:百万円) 等の合計 15 項目, 業績が, 売上成長率 (3 年平均) 等の合計 3 項目であり, これらの項目を実験のための変数とし分析に用いる．変数選択の理由として, できるだけ多くの項目で分析を行うことで分析結果を正確にしたかったこと, また, 少数株主持分を用いなかった理由は, データが少なく正確な分析ができないと考えたことがあげられる．

4.2 ロジスティック回帰モデル出力結果

不適切会計を起こした 32 件の企業のデータと起こしていない企業 32 件のデータ, 合計 64 件のテストセットを 10 組作り実験を行った結果, ロジスティック回帰の分析結果とオッズ比は表 3 のように得られる．回帰変数は, β_0 が定数, 表 2 の項目の現金等が β_1, 売上債権が β_2, 棚卸資産が β_3, その他流動資産が β_4, その他固定資産が β_5, 有形固定資産が β_6, 無形固定資産が β_7, 投資等が β_8, 仕入債務が β_9, その他流動負債が β_{10}, 短期借入金等が β_{11}, 長期借入金等が β_{12}, その他固定負債が β_{13}, 株主資本が β_{14}, 総資産が β_{15}, 売上成長率が β_{16}, 営業利益成長率が β_{17}, 純利益成長率が β_{18} とそれぞれ対応している．実験結果から, 現金等・棚卸資産・その他流動資産・その他固定資産・有形固定資産・短期借入金等・株主資本・売上成長率・純利益成長率のオッズ比は全て 1 より小さい値を取るため, 変動率が不適切会計企業の負の値の最高値以下であれば不適切会計発生の確率が高くなり, 投資等・その他流動負債・長期借入金等・その他固定負債のオッズ比は全て 1 より大きな値を取るため, 変動率が不適切会計企業の正の値の最低値以上であれば不適切会計発生の確率が高くなる．この結果を基にモデルの実装を行う．

4.3 不適切会計企業早期発見モデルの実装

現金等・棚卸資産・その他流動資産・その他固定資産・有形固定資産・短期借入金等・株主資本の 2010〜2012 年のデータを項目番号 1〜7 のデータとし, 投資等・その他流動負債・長期借入金等・その他固定負債の 2010〜2012 年のデータを項目番号 8〜11 のデータとし, 売上成長率・純利益成長率の 2010〜2012 年の変動率を項目番号 12〜13 のデータとしてファイル読み込みを行うと, 不適切会計企業との条件一致率を%で表示し, 50%を超える場合は危険度が高いと判断し警告を表示し, 50%以下の場合は危険度が低いと判断し, 危険小を表示するモデルの実装を行った．実装したモデルにより, 企業の不適切会計発覚による投資家の損害を未然に防ぐことができる．例として, 表 2 のデータを使用してモデルの実行を行った場合,
警告:不適切会計企業の条件に 63%一致しています.
不適切会計を起こす可能性が高いです.
を出力する．

[2]http://www.shares.ne.jp/

表 2 データセット

データ	2010	2011	2012	変動率
現金等	206,783	213,561	200,088	-1.6
売上債権	166,638	158,465	169,482	0.8
棚卸資産	89,959	92,929	102,493	6.7
その他流動資産	68,765	68,579	54,495	-10.9
その他固定資産	0	0	0	0
有形固定資産	143,561	141,341	127,808	-5.6
無形固定資産	216,030	205,979	197,145	-4.4
投資等	205,821	125,795	111,636	-26.3
仕入債務	74,074	68,715	75,330	0.8
その他流動負債	151,199	136,522	132,948	-6.2
短期借入金等	113,973	127,535	112,115	-0.8
長期借入金等	547,508	521,252	530,311	-1.5
その他固定負債	54,643	49,557	67,794	11.3
株主資本	155,672	112,477	44,770	-46.3
総資産	1,104,528	1,019,160	966,526	-6.4
売上成長率 (3 年平均)	-	-	-	-2.0
営業利益成長率	-	-	-	-23.8
純利益成長率	-	-	-	0

表 3 ロジスティック回帰・オッズ比結果

回帰変数	回帰係数	オッズ比
β_0	-0.3228512	0.751906724
β_1	-0.0669093	0.978442752
β_2	-0.0321661	1.031426711
β_3	-0.0476957	0.994106436
β_4	-0.0523327	0.987391166
β_5	-0.0496896	0.936371481
β_6	-0.0661108	0.993618449
β_7	-0.0110551	1.02962947
β_8	0.0066735	1.013468899
β_9	0.0156141	0.971174612
β_{10}	0.0111369	1.015188186
β_{11}	0.0037008	0.982674837
β_{12}	0.0046833	1.00435144
β_{13}	0.0055873	1.008840852
β_{14}	-0.0005509	0.985313908
β_{15}	0.1162156	1.000802322
β_{16}	-0.0008666	0.986654847
β_{17}	-0.0103021	1.010918173
β_{18}	-0.0207340	0.979089716

5　モデルの比較・評価

実装したモデルの性能評価を行うため, 第 2 章で述べた決定木, ニューラルネットワークによる警告モデルと性能比較を行う. 性能比較については, Patil ら [3] の研究を基に, 性能評価に適した真陽性率 (TPR), 偽陽性率 (FPR), 正確性 ($Accuracy$) について比較を行う. 本研究では, TPR は不適切会計を起こした企業を正確に判別できる割合であり, 不適切会計を起こした企業を不適切会計を起こした企業であるとして正しく判別した場合を a, 不適切会計を起こしていない企業を不適切会計を起こした企業であるとして誤って判別した場合を c とすると,

$$TPR = \frac{a}{a+c} \tag{4}$$

で表される. 一方, FPR は不適切会計を起こしていない企業を不適切会計を起こしたものとして誤って判別してしまう割合であり, 不適切会計を起こした企業を不適切会計を起こしていない企業として誤って判別した場合を b, 不適切会計を起こしていない企業を不適切会計を起こしていない企業として正しく判別した場合を d とすると,

$$FPR = \frac{b}{b+d} \tag{5}$$

で表される. また, $Accuracy$ は不適切会計を起こした企業を不適切会計を起こしたもの, 不適切会計を起こしていない企業を不適切会計を起こしていないものとして正しく判別できる割合であり,

$$Accuracy = \frac{a+d}{a+b+c+d} \tag{6}$$

で表される.　性能比較において, まず TPR, FPR, $Accuracy$ を算出するために必要な a〜d の値を, それぞれのモデルにおいて求める. 表 4 は各モデルの値を示したものである. 決定木・ニューラルネットワークの性能は Weka において算出し, パラメータは最高の性能のものとなるように設定を行った. TPR, FPR, $Accuracy$ の値は上記の計算式によって計算し, 得られた結果を有効数値 4 桁とし, 表 5 に示す.

表 5 から, TPR, FPR, $Accuracy$ 全ての割合において, 本研究が勝っていることが分かる. このことから不適切会計企業検出に最も優れているモデルは, ロジスティック回帰モデルであることが証明できた. また, 本研究の主目的である不適切会計企業の検出の観点においては不適切会計企業の検出率が, 32 件中 24 件の 75%と比較的高い割合で判別することができるが, 条件一致率 50%以上を不適切会計企業であると判別できるようにすれば, 検出率は 32 件中 26 件の 81.25% となるので, TPR, $Accuracy$ の性能は低くなるがそちらを採用するか, もしくは使い分けられるようにするという方法があげられる. 一方, Shih ら [1] の研究で使用したロジスティック回帰の不正会計企業の検出率は 87.5% と本研究のモデルと比較して検出率が高かった. 理由として, 一つの分野の財務諸表でのみ分析を行っていることがあげられるので, 更に性能を向上するための参考とする.

表 4 各実験の値

モデル	a	b	c	d
本研究	24	8	10.4	21.6
決定木	21.4	10.9	11.4	20.6
ニューラルネットワーク	17.3	14.7	11.4	20.6

表 5 性能比較

モデル	TPR	FPR	Accuracy
本研究	69.76%	27.02%	71.25%
決定木	65.24%	34.60%	65.31%
ニューラルネットワーク	60.27%	41.64%	59.21%

6　結論と考察

本研究では, 不適切会計を起こした/起こしていない上場企業の財務諸表と業績のデータをロジスティック回帰モデルで分析し, 不適切会計企業の条件を割出し条件を当てはめ分析していくことにより, 不適切会計企業の早期発見を行えるようにした. 性能比較において TPR, FPR, $Accuracy$ について比較を行った結果, 本研究で実装したモデルは, TPR69.76%, FPR27.02% , $Accuracy$71.25%と, 決定木, ニューラルネットワークと比較してすべての面で勝っていることが分かり, 不適切会計企業検出においてロジスティック回帰モデルが最も優れていることが証明できた.

また, 更に性能を上げるために, 10 年分のデータを入手し分析を行う, Shih ら [1] らの研究のように製造会社のみの財務諸表での分析や金融会社のみの財務諸表での分析等, 分野ごとの不適切会計企業の早期発見を行えるようにすることで, より性能が向上するように改良する等の方法があげられる.

参考文献

[1] Kuang-Hsum Shih, Ching-Chan Cheng, Yi-Hsien Wang, “Finacial Information Fraud Riskworning For Manufacturingindustry - Using Logistic Regression And Neural Network,” Romanian Journal of Economic Forecasting, pp.54-71, 2011.

[2] Efstathios Kirkos, Charalambos Spathis, Yannis Manolopoulos, “Data Mining techniques for the detection of fraudulent financial statements,” Expert Systems with Applications 32, pp.995-1003, 2007.

[3] Dipti D. Patil, Sunita M. Karad, Vijay M. Wadhai, J A Gokhale, Prasad S. Halgaonkar,“Efficient Scalable Multi-Level Classification Scheme for Credit Card Fraud Detection,” IJCSNS International Journal of Computer Science and Network Security, Vol.10, No.8, pp.123-130, 2010.

[4] Aihua Shen, Rencheng Tong, Yaochen Deng,“Application of Classification Models on Credit Card Fraud Detection,” Service Systems and Service Management, International Conference on, pp.1-4, 2007.