財務指標に対するバギングを用いた企業格付けの分析

2011SE024 後藤 謙治 2011SE078 井上 広大 2011SE208 岡 敬宏
指導教員: 河野 浩之

1 はじめに

本研究では，EDIUNET による企業の格付けを正しい出力データとして財務指標を用いて分析する．その中で，高い精度でいかに間違いを少なくしつつ判別できるか，またそこに費やす時間の短縮を目指す．そこで，Weka を用いてデータベースに格納した上場企業約 2300 件の財務データにバギングを利用し，EDIUNET の格付けの通り正しく判別し分析するシステムを実装する．本論文は全 5 章から構成され，2 章では，本研究に類似した先行研究の比較，財務データと財務指標について説明する．3 章では，問題定義とそれに基づく提案内容，バギングの説明をする．4 章では，PostgreSQL へのデータの格納，Weka による格付けの分析手順と得られた結果，生成されたルールの検証について説明する．5 章では，本研究のまとめを説明する．

2 データマイニングの手法と先行研究の比較

本章では，2.1 節で本研究に類似した先行研究，2.2 節で本研究で扱う財務データと財務指標について説明する．

2.1 先行研究の比較

表 1 は，財務分析を決定木によって行う本研究と類似した先行研究の比較である．

表 1 先行研究比較

[4]	・企業の財務データを含むデータベースから学習した決定木を利用し，信用リスクを定量化 ・領域ルール，決定木 ・エラー率は 0.190
[5]	・企業情報から倒産モデルを再構築し，有効性を検証 ・バギング学習，決定木，サンプリング ・Cap50 値は 70%
[8]	・個々の信用リスクの分析における決定木法の特徴・利点を研究し，実証分析を行う ・決定木，サンプリング ・予測精度 81.06%

表 1 の [5] は，企業情報ベンダーが提供する約 42 万件の予測倒産率から倒産モデルを再構築する研究である．企業情報ベンダーの倒産モデルは未公開のため，モデルを再構築し，構造を推定することで中小企業の与信フレームを考える．再構築アプローチによるモデル開発は，企業情報ベンダーから提供された倒産率を目的変数に，企業情報を説明変数とし，モデル構造を推定する方法である．ターゲットは予測倒産率 z% 以上を負事例，x% 未満を正事例にする．再構築モデルを安定化するために，バギング学習により複数の決定木の平均値を予測倒産率とする．再構築モデルは Cap50 値が 70% 前後であり，ビジネス利用に可能な精度を有する．この再構築モデルに企業情報を当てはめることで，中小企業約 120 万社の予測倒産率が得られる．

2.2 財務データと財務指標

企業は会計制度を持つことで，経営に関する情報を貨幣単位で収集，要約，分析し，報告するための会計情報を整備することができる．このような会計情報は，複式簿記の原理に従って作成される財務諸表に要約される．財務諸表は，貸借対照表，損益計算書，株主資本等変動計算書，キャッシュ・フロー計算書等から構成される．貸借対照表，損益計算書，キャッシュ・フロー計算書は 4 半期ごとに作成され，これらは一般に主要財務 3 表と呼ばれる．本研究では，主要財務 3 表の主な財務データを扱う．

次に，財務指標とは財務諸表の数字をもとに割り出した比率のことである．有価証券報告書から会社の様子を知りたいとき，また数社と比較した時は，財務比率を算出してその数値を用いると非常に分かりやすい．財務比率と貸借対照表や損益計算書，キャッシュ・フロー計算書に記載されている会計数値との大きな違いは，財務比率が規模の大小を問わず比較可能であるという点である．例えば，売上高や純資産，利益などは規模の大きい企業の方が規模の小さい企業よりも大きい方が一般的であるから，会計数値の大きさによって業績の良し悪しを測ろうとすると，規模の大きい企業という結果となる可能性が高い．これに対して，財務比率は割合を示すものであるから，規模の大きい企業の比率が大きいとは限らない．したがって，これらの財務比率は，企業の収益性や流動性，効率性などを測定するためには有効な手段であると考えられる．

財務データと財務指標の各項目の解説は [1][2][3][6] を参考にして説明する．

3 決定木を用いた財務分析の提案

本章では，3.1 節で問題提起と提案内容，3.2 節でバギングのアルゴリズムについて説明する．

3.1 先行研究での問題点をふまえた提案

先行研究 [5] での問題点は，判別精度が Cap50 値で 70% とまずまずの値だが間違いが少なくないことである．90% 以上が高精度モデルであるため，より精度を高める必要がある．また格付けを正確に判断するためには景気や業種を加味して判断することが必要だが，先行研究では省略されている．これを踏まえた上での改善点は，動作時間を短くしつつ，分析結果の精度向上をすることである．

財務分析を行う上で，間違った分類をいかに少なくするかということが最も重要であり，企業の信用リスクを正確に判断しなければならない．また，分析時間を短縮するこ

とで使いやすい分析ができるようにしていく．

本研究では，EDIUNET の格付け分析のために多くの財務指標を説明変数とした決定木のアルゴリズムを使用する．また，決定木の精度向上を行うために，アンサンブル学習のバギングを利用する．

また，実験データのクレンジングを行い余分なデータを減らし，バギングで識別器 (決定木) の個数を適切に設定することで処理にかかる時間を短縮する．

また財務分析では業界や企業規模の違いによって，財務データのとる値が大きく変わるので，各業界ごとに分けて分析を行う．そうすることで信用の高い分析結果を得ると同時に，業種の違いによるルールの違いなどが検証できる．

3.2 バギングの適用

bagging とは袋詰めの意味であり，異なるデータ群で複数回分析した結果を統合・組み合わせるものである．アルゴリズムは，まず「教師付き標本」をブートストラップで B 回抽出したデータを判別関数 (決定木) を利用することで B 個の弱学習器を作成する．次に B 個の学習器から最終的な 1 つの判別関数，判別器を決定する．その際，判別問題と回帰問題の 2 つの場合に分類手法が異なる．判別問題の場合は各弱学習器に標本を入れて各サンプルについて多数決を用いて最も多いラベルを正しいラベルとする．回帰問題の場合は，各学習器に標本を入れて各サンプルの出力された平均をとる．

4 Weka を用いて財務分析を行う

本章では実際に Weka のアルゴリズムを用いて財務分析を行った結果を示す．4.1 節で分析に用いるデータの紹介と実験環境，4.2 節で PostgreSQL へのデータの格納，4.3 節では Weka による実験手順，4.4 節で Weka の実験結果，4.5 節で生成された決定木のルール検証を説明する．

4.1 実験データとソフトウェアの実験環境

本研究で使用する財務データは財務分析.jp(http://www.financial-analysis.jp/) の財務諸表一覧 (貸借対照表，損益計算書，キャッシュ・フロー計算書) からダウンロードする．また，該当する企業の格付けを EDIUNET(http://ediunet.jp/) から探して入力し，財務諸表のデータが欠けていたり格付けのない企業に関しては分析が行えないので除外する．実験ではこの作業で揃った約 2300 件の企業の財務データを扱う．

本来であれば複数の格付けを利用したデータにするべきだが，それぞれの格付けに共通して対応する企業が少なく，実験データが極端に減るため本実験では EDIUNET の格付けのみを正しい出力データとして使用する．

また，システムの実験環境は Windows8，Intel Core i3，メモリ 8GB である．表 2 は用いる財務データの一部を表す．

表 2 実験で扱う財務データの一部

売上高	181885
当期純利益	423
資本金	5664
業種	情報通信
市場	東 1
総資本	84937

4.2 PostgreSQL によるデータ格納

本研究では postgreSQL を用いて財務データの格納を行う．はじめに実験に用いるデータの入った EXCEL ファイルを CSV 形式に変換し，コマンドを使ってテーブルに data.csv のデータ約 2300 件を格納する．図 1 のようにテーブルを作り，そこの EXCEL の財務データをコピーすることで格納する．

```
zaimu=# create table data (企業名 text, 格付け text, 流動資産合計 decimal,
  固定資産合計 decimal, 資産合計 decimal, 流動負債合計 decimal,
  固定負債合計 decimal, 負債合計 decimal, 資本金 decimal,
  株主資本合計 decimal, 売上高 decimal, 売上原価 decimal,
  売上総利益 decimal, 営業利益 decimal, 経常利益 decimal,
  特別損益収支 decimal, 当期純利益 decimal, 減価償却 decimal,
  営業キャッシュフロー decimal, 投資キャッシュフロー decimal,
  財務キャッシュフロー decimal, キャッシュフロー decimal, 業種 text, 市場 text)
CREATE TABLE
zaimu=# COPY data FROM 'C:¥¥data¥¥data.csv' WITH CSV;
COPY 2296

zaimu=# SELECT * FROM data;
          企業名          | 格付け | 流動資産合計 | 固定資産合計 |
 資産合計 | 流動負債合計 | 固定負債合計 | 負債合計 | 資本金 | 株主資本合計 |
 売上高 | 売上原価 | 売上総利益 | 営業利益 | 経常利益 | 特別損益収支
 ----------------------------------+--------+--------------+--------------+
 ----------+--------------+--------------+---------+-------+------------+
 ----------+----------+----------+----------+----------+-------
 フォーバルクリエーティブ          | BBB    |           57 |          100 |
      157 |          144 |            1 |          145 |         1536 |           10 |
      116 |           64 |           52 |         -174 |         -179 |           14
```

図 1 PostgreSQL でのデータ格納

次に，データベースと Weka を連携させるために JDBC を利用する．ダウンロードは (http://jdbc.postgresql.org/download.html) から行う．まず，Weka にデータベースを読み込ませる為，props ファイル内に JDBC URL を記述する．Weka のフォルダ内の weka.jar にデータベース関連の設定が書かれた DatabaseUtils.props の雛形が多数あるので，それを利用する．DatabaseUtils.props を用意し，C ドライブ直下におく．本研究ではデータベースに対応する DatabaseUtils.props.postgresql を編集する．同様にして，weka.jar を開いて中の weka/experiment/DatabaseUtils.props を編集する．

最後に JAVA を動かすために，環境変数を入力する．Weka のエクスプローラーにある「DB を開く」から，URL と PostgreSQL のユーザー名とパスワードを入力したのち接続ボタンでデータベースと Weka を連携させる．

4.3 Weka による実験手順

JDBC によって Weka からデータベースを読み込むことができたら，データベース内の財務データを用いて，クエリーに図 2 の SELECT 文で各財務指標を求めるために必要な計算式を入力し財務指標の計算する．

```
SELECT *, 経常利益/資産合計*100 AS ROA, 当期純利益/株主資本合計*100 AS ROE, 株主資本合計/資産合計*100AS 自己資本比率, 売上総利益/売上高*100 AS 売上高利益率, 営業利益/売上高*100 AS 売上高営業利益率, 経常利益/売上高*100 AS 売上高経常利益率, 資産合計/売上高*365 AS 総資本回転期間, 売上高/資産合計 AS 総資産回転率, 流動資産合計/流動負債合計*100 AS 流動比率, 固定資産合計/株主資本合計*100 AS 固定比率, 負債合計/株主資本合計*100 AS 負債比率, 営業キャッシュ・フロー/固定負債*100 AS キャッシュ・フロー比率, 営業キャッシュ・フロー/流動負債*100 AS 営業キャッシュ・フロー対流動負債比率, 減価償却/(固定資産+減価償却)*100 AS 減価償却率, 営業キャッシュ・フロー/売上高*100 AS キャッシュ・フロー・マージン, フリーキャッシュ・フロー/売上高*100 AS 売上高フロー・キャッシュ・フロー比率, 固定資産合計/(固定負債合計+株主資本合計)*100 AS 固定長期適合率, 営業利益/株主資本合計*100 AS 総資本営業利益率 FROM data WHERE 業種 = '情報通信業';
```

図 2　Weka で財務指標を計算するコマンド

また業種別に分析するために，1 種類ずつ業種でデータを分けた上で Weka の分析を行う．本研究では EDIUNET の格付けに対して決定木やバギングのアルゴリズムを用いて，格付けに対する精度やルールを出す．また，バギングを用いる場合には，1 つの識別器 (決定木) に含まれるデータの総データに対する割合 P と，識別器 (決定木) の個数 I を調整することでより高い ROC Area を求める．

ROC Area とは ROC 曲線の下の面積比率であり，分析精度を表す数値である．

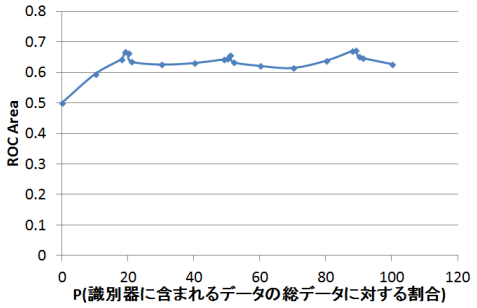

図 3　情報通信業における P の変化による ROC

図 3 は，情報通信業の企業における P の変化によって推移する ROC Area の値を示したグラフである．I は初期値の 10 のまま P を変化させることで，最も高い ROC Area を示す P の値を求める．

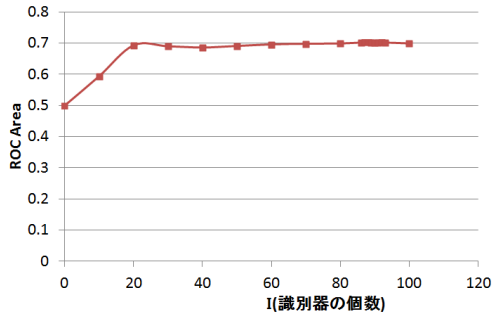

図 4　情報通信業における I の変化による ROC

図 4 は，情報通信業の企業における I の変化によって推移する ROC Area の値を示したグラフである．P は図 3 で ROC Area が最も高くなった値に固定し，I を変化させ最も高い ROC Area を求めた．I に上限はないが，あまり大きな値にすると過学習によって精度が下がったり，分析時間が長くなるので本研究では 100 を上限とする．

4.4　Weka の実験結果

表 3 は業種ごとの J48 とバギングで出た ROC Area の分析結果である．件数はその業種の企業数，P は 1 つの識別器 (決定木) に含まれるデータの総データに対する割合，I は識別器 (決定木) の個数，J48 と bagging はそれぞれの ROC Area を表す．

表 3　決定木とバギングの業種別の ROC Area

業種	件数	P	I	bagging	J48
情報通信	191	89	87	0.703	0.603
小売	227	79	71	0.645	0.533
陸運業	49	39	9	0.655	0.63
サービス	221	41	10	0.627	0.527
その他製品	78	90	29	0.67	0.562
建築	131	50	8	0.679	0.568
食料品	101	41	22	0.68	0.545
卸売	205	81	90	0.623	0.531
金属	66	11	11	0.566	0.457
科学	155	19	10	0.606	0.468
不動産	63	79	8	0.598	0.589
機械	171	39	10	0.556	0.536
電子機器	189	80	9	0.533	0.522
平均				0.626	0.544

業種別に財務指標によって分析した結果，決定木のアルゴリズム J48 で行った分析とバギングのアルゴリズムで行った分析の実験結果には差が出た．情報通信の場合には，決定木のアルゴリズムに比べてバギングのアルゴリズムの方が約 10%ROC Area の数値が大きい．つまりバギングを行ったことで，より精度の高い結果を得られた．

また，業種によって精度のバラつきが大きく，最大で ROC Area70.3% という結果を出すものもあれば，ROC Area 約 50% のランダムで精度が良くないものも数業種ある．

先行研究はこの値が約 70% であるが，評点というものを分析に含めたときの精度である．評点とは，財務情報や会社情報から独自の基準で企業を評価したもので，説明力が強いので説明変数に含めるとモデル全体が引きずられる．先行研究は評点を含めないと精度が 60% 以下のランダムモデルになっているため，評点を含まず行った本研究は業種によって精度の向上ができている．

なお，表 3 に載っていない業種は企業のデータ数が少なく，バギングを行うことが難しいため分析できなかった．

図 5 は，情報通信業の分析において，P を 89 に固定し，I を増加させたときのグラフである．このグラフから I と分析時間は比例することが分かる．本研究では I の上限を 100 にしたため，分析時間は平均で約 1 秒だった．この分析時間はシステムを実務で扱う場合にも，滞りなく作業できる時間である．

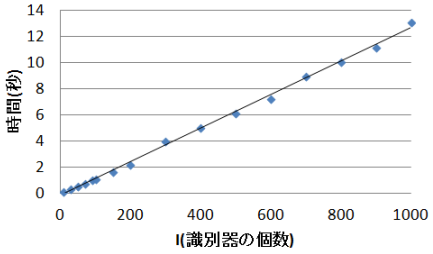

図 5 I の変化による分析時間の増加

4.5 生成された決定木のルール検証

今回分析した結果からできた決定木は，細かい分類まで入れるとかなり複雑なものになってしまうため，minNumObj(葉節点に含まれる最小データ数) の項目を大きめに設定する．情報通信業ではこの値は 20 である．

図 6 は情報通信業の企業を J48 のアルゴリズムで分析し，生成した決定木である．

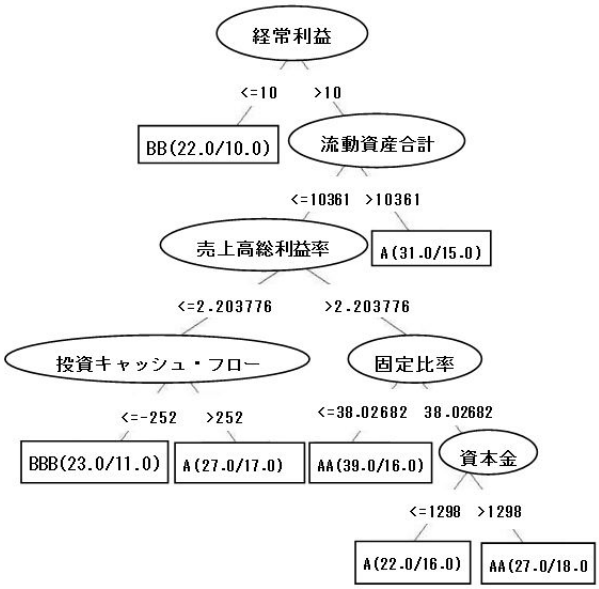

図 6 J48 のアルゴリズムで生成した決定木

この決定木が本当に正しいルールを示しているのかを検証する．はじめに経常利益で分かれており，企業の利益が高いほど格付けが良くなるのは正しい．次に流動負債合計によって分かれるが，負債の合計が大きいほど格付けが良くなっている．本来なら負債が少ないほうが企業のリスクが少なく格付けは高くなるはずだが，企業データが大企業ばかりになったための判別だと考えられる．企業の規模に応じて負債の額も大きいため，負債が大きいほど大企業で安定性があると判別する可能性がある．売上高総利益率は高いほど良く，情報通信業の平均は約 40% である．判別では約 2% とかなり低い値にはなっているが高いほど格付けが良くなっている．投資キャッシュ・フローのマイナス値は小さいほど良く，格付けが正しくできている．固定比率は，株主資本に対する負債の割合なので値が小さいほど良く，分類は正しい．資本金は大きいほど大企業で安定力があり，正しく格付けできている．この結果，情報通信業において生成された決定木のルールは信頼できると確認できた．また，他の業界においてもルールの信頼性が確認できた．

決定木のルールは業種ごとに大きく違うが，重要であることが多かいのは経常利益，市場，キャッシュ・フロー，固定負債，売上高経常利益率，売上高総利益率である．

5 まとめ

本研究では，企業の財務データから財務指標を求め，Weka によって EDIUNET の格付け精度分析を行った．分析には Weka におけるバギングのアルゴリズムを用いることで ROC Area やルールを求め，格付けの信頼性やルールを検証した．そして，JDBC を用いることで，PostgreSQL に格納されたデータを Weka から読み込むことができるように連携させた．

その結果，業種によっては ROC Area が最大 70.3% という値を出すことができ，先行研究に比べて 0.3% 精度が向上した．また，Weka でのバギングの分析時間は約 1 秒であり，システムを実務で扱う場合にも滞りなく作業できる時間である．そして，生成された決定木のルールは，財務的な面から見てもおおよそ正しいルールを得ることができた．

参考文献

[1] 船橋健二, 辻達博, 藤井邦明, 長谷川和彦, “ 図解 中小企業の経営分析, ” 税務経理協会, pp.26-71, 2005.

[2] 飯田信夫, “ 21 世紀のスタンダードがわかる 35 の財務指標, ” 中央経済社, pp.22-191, 1999.

[3] 倉田 三郎, 藤永 弘, 石崎 忠司, 坂下 紀彦, “ 入門 経営分析, ” 同文舘出版, pp.1-68, 2008.

[4] 森本康彦, 福田剛志, 松澤裕史, “ 領域分割決定木を利用した信用リスク管理, ” 電子情報通信学会研究報告, データ工学, Vol.93, pp.1-8, 1998.

[5] 小野潔, “ データマイニングを用いた中小企業の信用リスクの推定モデル, ” 社団法人日本オペレーションズ・リサーチ学会シンポジウム, Vol.55, pp.13-22, 2006.

[6] 斎藤 孝一, “ ケースで学ぶ財務諸表分析 -基本戦略と財務指標の関係-, ” 同文館出版, pp.1-78, 2013.

[7] 株式会社 ALBERT 巣山剛, データ分析部, システム開発・コンサルティング部, “ データ集計・分析のための SQL 入門, ” 株式会社マイナビ, pp.56-82, 2014.

[8] Yu Yanping, Qian Zhengming, Yang Min, Guan Rui, Fang Liting, Guo Penghui, “ Research on the Application of Decision Tree to the Analysis of Individual Cresit Risk. ” International Conference on Information Engineering Lecture Notes in Information Technology, Vol.25, pp.209-214, 2012.