SGEN(Supervised Group Elastic Net)：
高次元・高関連性データを用いた説明変数発見

Supervised Group Elastic Net: Internal Feature Estimation from High-dimensional and correlative Data

コウ詩敏* 早矢仕晃章 大澤幸生

Shimin Kuang*, Teruaki Hayashi, and Yukio Ohsawa

東京大学 大学院工学系研究科 システム創成学専攻
Department of Systems Innovation, Graduate School of Engineering, The University of Tokyo

Abstract: This paper presents an approach to finding internal features from high-dimensional and -correlative data, which we called Supervised Group Elastic Net(SGEN). SGEN keeps track of the group explained features in the dataset. It focuses on the relatively weak correlative feature in the group. SGEN consists of 3 steps. The first step is grouping. We group the expandable feature into a group, using K-means. The second step is choosing the essential group from Group Elastic Net. The third step is finding a relatively low-correlative group in selected groups. The expandable feature in the relatively low-correlative group may be indiscoverable. Furthermore, we conduct an experience to apply SGEN to financial statement data. As a result, we find the SGEN can find the internal feature that we can relate to a certain event in the real world.

1 はじめに

デジタル社会の発展につれ，多様なデータを活用した意思決定が求められいる. 実社会では, 時々刻々と生み出される多様なデータに対して統計学や情報処理技術を活用した様々なモデルが開発されてきた. さらに，昨今は膨大なデータが蓄積可能となり，ビッグデータ処理としてパターン認識や自然言語処理などに注目度集まっている.

私たちの日常生活もデータ利活用によって変化してきている．例えば，メーカーはデータからサプライチェーンの最適化を行い，様々なプロセスの自動化を実現してきた．また，映画製作会社は視聴者の多様な行動データを解析することでターゲット層の嗜好や傾向を発見し，新たなコンテンツ作りに活かしている．つまり、データは実社会の事象を理解するためのエンティティという側面だけでなく、それによって得られた知識や知見が私たちの生活に影響を与えるという新たな属性を有しつつある.

しかし，データからモデルを学習する際には，2 つの大きな課題がある．1 つ目は，モデルが有用な情報を抽出するほど，十分なデータが蓄積されていないということである. 例えば, 財務諸表データ（企業決算期の貸借対照表，損益計算書，キャッシュフロー計算書）は企業の財務状況を開示するために存在する．しかし，企業がどれくらいの財務状況を開示するかは企業の IR 戦略となるため, 開示されるデータ量が少ない場合がある．そのため，p>>n 問題，すなわち，変数量がデータ量より多くなっていまうという問題が生じうる．この問題に対し，様々な統計学手段が開発されてきた．2 つ目の課題はデータの間の相関性の問題である．説明変数の間の相関性が高すぎる場合，目的変数との関連性を分析する際に，どの説明変数が効果的かを判断することが困難となる．そのため，データの間の相関性を考慮するモデルの開発が重要となる.

以上に述べた課題に対し，データを予測モデルに入れる前に，事前に重要な説明変数を抽出し，モデルの精度を高めるという，目的を持つ「変数選択」の手法が注目されてきている．本研究は従来の変数選択手法が高次元・高相関性データから見かけの相関性を持つ変数を抽出する課題に対し，実社会での事象に対応する説明変数を発見することを目的にする.

2 先行研究

Girish Chandrashekar, Ferat Sahinの調査では，変数選択はFilter methods, Wrapper methods, Embedded methodsの3種類に大きく大別される[1].

Filter methodは変数の相関性評価，情報量評価などの変数のクオリティを評価できる基準を使い，評価基準的によい変数をピックアップし，関連性の低い変数を排除する方法である．相関性評価の式は式１に示す.

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}}$$

式　1：Correlation criteria

式の中，xi は第 i 項の変数であり，Y は目的変数であり，cov()は共分散である．そして，var()は分散を指す.

Wrapper methods は予測モデルの精度で変数のクオリティを評価する．つまり、ランダムに変数のペアを作り，異なるペアを予測モデルに入れ，モデルとしての精度が良いものを選択するという方法である．Wrapper methods は作られたペアに対して計算を実行するので，計算量は膨大となる．そこで，計算量を減らすため，Embedded methods が開発されている．Embedded methods は変数選択をモデルの訓練プロセスの一部として考え，変数選択と予測を同時に行う方法である．Embedded methods は計算量大幅に低減し，利便性が高いため，様々な応用がなされている.

変数選択に有効な方法としてデータ処理時の p>>n 問題を解決し，解にスパース性をもたらす LASSO が提案されている[2]．しかし，LASSO は計算時の ill-condition 現象，すなわち異常値が全体の解に影響する現象を考慮していないという問題があった[3]．そこで，Adaptive LASSO という手法が開発された．Adaptive LASSO はスパース性と、異常値が全体の解に影響する現象を考慮した[4]が，データの間の相関性を考慮できていなかった．そこで Elastic Net では，ノルム１とノルム２を融合したペナルティの性質を研究し，そのペナルティでデータの間の相関性を考慮した[5].Elastic Net は式２に示す.

$$\hat{\beta} = arg\min\left\{||y - x\beta||_2^2 + \lambda \sum_{j=1}^{p}(\alpha|\beta j| + (1-\alpha)\beta j^2)\right\}$$

式２：Elastic Net

ここで,y は目的変数であり,x は説明変数である.

$y = \beta 1X1 + \beta 2X2 + \beta 3X3 + \cdots$ の線形回帰を満たすと仮定する。p は変数の個数，λはペナルティを調整するパラメータである．|β|はノルム１であり，β の２乗はノルム２を指す．Elastic Net は式２の最適化を行うことを目的にする.

Elastic Net はデータ間の p>>n 問題および相関性問題を解決できるが，変数間の相関性が極めて高い場合は，目的変数と相対的に高い関連性を示す変数が抽出されてしまう.例えば,目的変数が「収益」で,説明変数がそれぞれ「収入」,「商品売買からの収入」,「売買によるキャッシュ」,「人件費」の時,「収益」,「収入」,「商品売買からの収入」,「売買によるキャッシュフロー」は同じ事象を表示しているため，極めて高い相関性を持つ.そこで説明変数の「収入」,「商品売買からの収入」,「売買によるキャッシュフロー」の中から変数が抽出されるが，もし実社会でリストラによる収益の変化という事象があったとしても，人件費は抽出されない.

そのため，本研究はこのような高次元かつ高い相関性を持つデータに対し，実社会での事象に対応する説明変数を発見する手法を提案する.

３　提案手法

本章は高次元かつ高い相関性を持つデータから，見かけの相関性だけでなく，実社会での事象に対応する変数を発見する手法 Supervised Group Elastic Net（SGEN）を提案する．提案手法の概念について Feature Concept を用いて説明する．Feature Concept はデータを利活用しようとするときに，得た知識あるいは情報を概念モデルとして描いた図である．variable, predicate, function, logical clasue など, Feature Concept は様々な形で表現できますが,図で表すのが最も多い[6]．そのため，Feature Concept の図（以下 FC 図）を用いて説明する.

SGEN は 3 つのステップに構成される。

ステップ 1 では，説明変数のグループ化を行う．グループ化には K-means を用いている。K-means の FC 図は図 1 に示す.

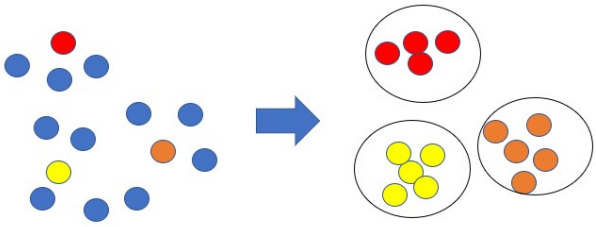


図１：K-means　Feature Concept 図

各点はデータを示し，赤、黄色、橙色で表された点群はランダムに生成した他データとのユークリッド距離が相対的に近いデータのグループとなる。

ステップ 2 では, Group Elastic Net で目的変数と相関性を持つ説明変数のグループを抽出する．Group Elastic Net は式 3 に示す.

$$\hat{\beta} = \arg\min \frac{1}{2m}\left\{\left\| y - \beta 0 - \sum_{j=1}^{P} Aj\beta j \right\|_2^2 + m\sum_{j=1}^{P}\sqrt{nj}(l1\|\beta j\| + l_2\|\beta j\|^2)\right\}$$

式　3：Group Elastic Net

$Aj(m \times n)$ は行列であり，説明変数の第 j グループを示す．P はグループの個数である。y は m 次元のベクトルである．l1,l2 はノルム 1 と 2 の調整パラメータである．Group Elastic Net では式 3 を Proximal gradient descent で最適化する．Proximal gradient descent は凸関数の最適化方法の 1 つとして, proximal operator で目的変数との二乗距離による罰則項とペナルティの和を最小化する点に写像される．微分不可能なところについて proximal operator で処理する方法である.

ステップ 3 は Group Elastic Net に選択されたグループの中の説明変数間の相関を計算し，他グループ内の説明変数間の相関と比較する．グループ内の説明変数間の相関性は他グループと比べ，低いグループ内の変数は目的変数と関連しながらも，他の説明変数と関連性が低いため，高関連性による影響を受けない説明変数となる.

以上の 3 つのステップを踏まえる手法を Supervised Group Elastic Net と呼ぶ．当該手法の FC 図は図 2 に示す.

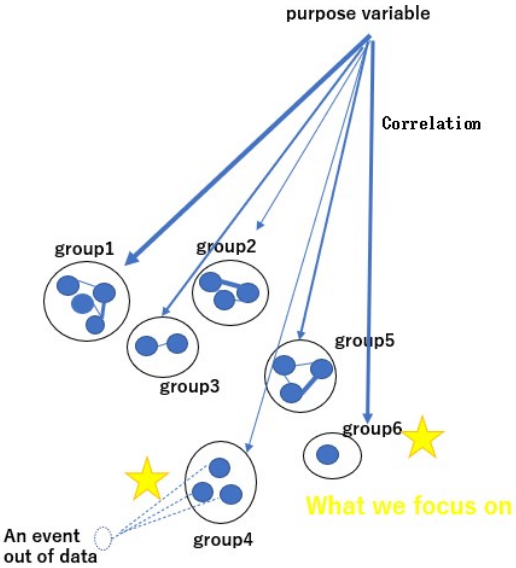


図　2：Supervised Group Elastic Net の FC 図

4　実験

本研究の実験には，企業の財務諸表データを用いた．企業の財務諸表データは企業の経営・財務活動を表し，企業活動をより具体的に記録するため，同一活動を異なる財務項目で表現している場合が多い. また，財務諸表データは一般的に四半期ごとに（3 か月）一回発表されるので，データ量が少ないが，項目数が多い．そのため，財務諸表データは多次元かつ極めて高い相関性を持つ.

次に，実験に使うデータとデータ処理について説明する．本研究は中国に上場する企業の財務諸表データを用いる. また，財務諸表データとは，企業の貸借対照表，損益計算書，キャッシュフロー計算書に記載する項目の決算データである. 2011 年 3 月決算から 2020 年 9 月決算までの時系列データに対し, 式 4 から差分を計算し，分析を行う.

$$\Delta x = x(i+1) - x(i)$$
x:各財務項目データ
i：i時点

式　4：差分の計算

財務諸表の主要項目は主な場合では記載されるが, 企業の IR 方針により, 主要項目以外の財務項目を使って公表する場合もある. 表 1 に示すのは貸借対照表の主要項目である. 企業の資産、負債及び純資産の毎期のストックが記載している. 各財務項目は資産, 負債，純資産への分別が中国の会計基準に従う. 表 2 に示すのは損益計算書の主要項目である．損益計算

1 本研究では会計利益とは、損益計算書の当期純利益（Net profit）である

書は企業のフローを記録する．企業の売上高から売上原価，営業外費用などの項目を計算して当期純利益を計算する．表３はキャッシュフロー計算書の主要項目を示す．キャッシュフロー計算書は経営活動と財務活動，投資活動が記録されている．

Cash	Transactional financial assets	Derivative financial assets
Advances to suppliers	Other receivables	Inventories
Long-term receivables	Long-term equity investments	Investment
Construction in progress	Intangible assets	Goodwill
Other non-current assets	Total non-current assets	Total assets
Accounts payable	Employee benefits payable	Taxes payable
Total current liabilities	Long-term borrowings	Provisions
Total liabilities	Capital surplus	Surplus reserve
Notes Receivable	Long-term prepaid expenses	Total equity attributable to equity holders of the Company Minority interest
Other current assets	Short-term borrowings	Accounts Receivable
properties	Current portion of non-current liabilities	Total current assets
Long-term prepaid expenses	Deferred tax liabilities	Fixed assets
Deferred tax assets	Total non-current liabilities	Other non-current liabilities
Notes payable	Total liabilities and owners equity	

表　１：貸借対照表項目

Total income
Interest income
Total Cost
Operating cost
Interest expenses
Business tax and surcharge
Operating expenses
Administration expenses
Financial expenses
Gains on the changes in the fair value
Investment income
Operating profit
Non-operating revenue
Non-operating expenses
Total profit
Income tax
Net profit

表　２：損益計算書項目

Cash from selling commodities or offering labor
Refund of taxes and surcharges
Cash received relating to other operating activities
Sub-total of cash inflows from operating activities
Cash paid for goods and services
Cash paid to and on behalf of employees
Payments of taxes and surcharges
Cash paid relating to other operating activities
Sub-total of cash outflows from operating activities
Cash received from disposal of investments
Cash received from returns on investments
Sub-total of cash inflows from investing activities
Cash paid to acquire fixed assets, intangible assets and other long-term assets
Sub-total of cash outflows from investing activities
Net cash flows from investing activities
Sub-total of cash inflows from financing activities
Cash repayments of borrowings
Cash payments for interest expenses and distribution of dividends or profits
Sub-total of cash outflows from financing activities
Net cash flows from financing activities
Add: Cash and cash equivalents at beginning of year
Cash and cash equivalent at end of year

表　３：キャッシュフロー計算書項目

各企業の財務データは qianzhan の有料データベースから抽出して，欠落値は NA とした．各企業の財務項目データに関して，会計基準に即する数字は記録されるかは当該企業の IR 情報と比較して，データ

は実際の IR 情報と一致するかを精査した.

本研究では，Midea Group Co.，Ltd（銘柄コード：000333.SZ）の財務諸表データを使う．なお，Midea Group Co.，Ltd は中国の著名な白物家電メーカーである.

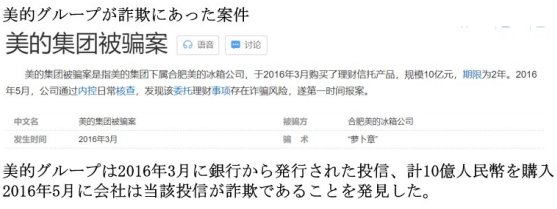
5　結果と考察

本研究では，まず Media グループの財務諸表データに対して分析を行った．SGEN により，発見された項目を表 4 にまとめる.

Other non-current assets（そのほか非流動性資産）
Current portion of non-current liabilities（1 年以内の非流動性負債）
Financial expenses　（金融費用）
Inventories　（在庫）
Fixed assets　（固定資産）
Capital surplus　（資本準備金）

表　4 ：SGEN で発見した財務項目

続いて，それぞれの財務項目に対し解釈を行う。表 4 の「そのほか非流動性資産」は一般的に土地，建物などの非流動性資産以外の資産を指す.「そのほか非流動性資産」は企業が自ら開示しない限り，中身が分からないため，分析する際に注意が必要であることが知られている．Midea グループが 2020 年に開示した年次報告書によると,「そのほか非流動性資産」の中身は銀行から発行された投信である．行から発行された投信は企業の投資業務に値するが，実社会において「そのほか非流動性資産」という項目は当該企業の財務諸表分析に対する影響があまり討論されていない．ただし，当該企業の決算資料や関連ニュースを精査すると，当該企業は 2016 年にそのほか非流動性資産の中身となる銀行から発行された図　3 に示したように，投信にかかわる詐欺にあったと報道された．また，図　4 に示した最近の年次決算報告書に「主要資産の重大変化状況」の中に，「そのほか非流動性資産」はほぼ毎年書かれている．2019 年の年次決算報告書では 800%の増加率が書かれている.



美的グループは2016年3月に銀行から発行された投信、計10億人民幣を購入した。
2016年5月に会社は当該投信が詐欺であることを発見した。

図　3 ：実社会で起きた Midea グループの詐欺案件

1、主要资产重大变化情况

主要资产	重大变化说明
在建工程	同比减少 42%，主要系工程完工所致
货币资金	同比增加 154%，主要系定期存款增加所致
应收票据	同比减少 62%，主要系重分类至应收款项融资所致
可供出售金融资产	同比减少 100%，主要系重分类至交易性金融资产及其他非流动金融资产所致
长期应收款	同比增加 3,370%，主要系融资租赁业务增加所致
其他非流动资产	同比增加 799%，主要系一年以上到期的结构性存款增加所致

そのほか非流動性資産　　前年比800%，主に銀行から発行された投信が増加したため

図　4 ：決算報告書の重大資産変化状況

さらに，図　5に示したように，当該企業の公式 IR 質問コーナーに，投資者はそのほか非流動性資産という項目の開示内容に対して，疑問の声を上げたが，会社からの良い返答をいまだにもらえていない.

図　5 ：投資者からの質問内容

次に，選出された 1 年以内の非流動性負債に関して，1 年以内の非流動性負債は 1 年以内に返済すべき長期負債のことである.本項目は会計利益と関連性を示していると本手法を通じてわかった.負債の額と利益がかかわるとのことである.Midea グループは社債などの長期負債で経営し，その資本コストが利益に影響している.当該企業は長期負債を多用することが一般的な財務諸表分析の安全性と流動性分析からわかるが，実際的に資本コストと会計利益が関連しているという実社会のファイナンスの視点から納得性のある結果である.

金融費用は融資などの金融活動による費用である.当該企業が行う金融活動は会計利益と関連性を示している.一般的に営業費用以外の費用が収益とかかわらないと考えるので，当該企業の関連情報を精査した.そこで，当該企業は自身がサプライチェーンの中の優位性を強みとして使い，サプライヤーに融資活動を含む金融会社を設立したと分かった.

在庫は売上と明らかに関連していて，そこで利益と関連性を示すのが当たり前の結果となるので，解

釈を行わない．固定資産は当該企業の会計利益と関連性を示すのが、固定資産による収入は企業の利益に影響していると考えられる．そこで，当該企業の事業セクターを調べた結果，グループ傘下に不動産会社があると分かった.

最後の資本準備金について解釈する．資本準備金株式発行などの資金調達を行う際に，一部の金額が会社法の規定により，資本金に繰り入れるが，残りの金額は資本準備金に入る．また，資本準備金を資本準備金は会計利益と関連性を示すことから，当該企業の株式発行が活用されることが分かった. 実際，経営層に対する奨励金として株式が発行される制度がある.

６　結論

本研究は高次元かつ高い関連性を持つデータに対して，見かけの相関性ではなく，実社会での事象に対応する変数を発見する手法となる SGEN を提案した. 本手法を利用して, 説明変数間極めて高い関連性を持ち，かつ高次元の財務諸表データを分析した. 実験において, 中国の著名な白物家電メーカーMideaグループの財務諸表データを用いた.

その結果, SGEN で発見した財務項目は当該企業の実社会での事象に対応していることに加え，それらの項目は従来まだ発見されていない.

今後は SGEN で選択された変数を予測モデルに入れ，予測精度を従来の手法と比べることにより，変数選択手法としての有用性を確かめる．また，他の会社の財務項目データや，遺伝子データを分析することで，適用するデータのバラエティーを増やす.

参考文献

[１] Girish and Ferat , “A survey on feature selection methods, ” Computers&Electrical Engineering, pages 16-18 , (2014) .

[２] Tibshirani R,“Regression shrinkage and selection via the lasso, ” J. R. Statist. Soc. B , pp. 267-288 (1996).

[３] Tibshirani , “The lasso problem and uniqueness, ” Electron.J.Statist, 1456-1490, (2013).

[４] Hui Zou, “ The Adaptive Lasso and Its Oracle Properties, ” Journal of the American Statistical Association, 1418-1429,(2012).

[５] Zou H, Hastie T, “Regularization and Variable Selection via the Elastic Net,”J. R. Statist. Soc. B, pp. 301-320 (2005).

[６] Ohsawa et al, “Feature Concepts for Data Federative Innovations, ” Computer Science, Machine Learning, arXiv:2111.04505, (2021).