# テキストマイニングを用いたコンサルティングサービス支援手法に関する研究

渡邉, るりこ

(Degree)
博士（システム情報学）

(Date of Degree)
2020-09-25

(Date of Publication)
2021-09-01

(Resource Type)
doctoral thesis

(Report Number)
甲第7884号

(URL)
https://hdl.handle.net/20.500.14094/D1007884

2020 年 度

# 博 士 論 文

テキストマイニングを用いた
コンサルティングサービス支援手法に関する研究

神戸大学大学院 システム情報学研究科
システム科学専攻

渡邉　るりこ

2020年7月

# 目 次

# 図 目 次

# 表 目 次

# 第1章　緒論

## 1.1　研究背景

### 1.1.1　中小企業とコンサルティングサービスの現状

政府は2010年に「中小企業憲章」を閣議決定し，中小企業の活性化は日本経済にとって重要な課題であるとしている[1]．一方で，中小企業の多くは資金や人材などに制約があり問題を自社内だけで解決することが困難であるため，中小企業支援機関が相談を請け負っている．中小企業を活性化させるために，中小企業支援機関の支援体制の重要性は高いといえる．

中小企業庁は中小企業支援を行う支援事業担い手の多様化･活性化を図るため，2012年に「中小企業経営力強化支援法」を施行した[2]．このように中小企業の支援体制が確立している一方で，中小企業は資金的余裕がない場合が多く一つの中小企業支援機関があらゆる経営相談に対応しなければならないため，相談内容は専門性が高く幅広い分野である．中小企業支援機関の中でも特に広い分野の相談業務を行っているコンサルティングサービスでは，相談業務を行う際の課題として，相談に対応するために確保できる時間や費用，人材の不足や相談業務を行う者の能力不足が，課題全体の四割を占めている[3]．このようにコンサルタントの人数不足や相談業務における経験の必要性から，コンサルティングサービスはコンサルタントの勘や経験に依存し，幅広い問題に対応する専門性を確保することが困難である．

コンサルティングサービスのように人間の労働割合が大きい産業を労働集約型産業とよぶ．労働集約型産業では労働生産性が低いことが問題であり，労働集約型産業を対象とした支援手法の研究が幅広い分野で行われている[4]．セル生産システムにおけるオペレータの効率を向上させるため，実際の製造要件を満たしながら組み立て作業者の情報面と物理面の両方を支援するシステムを提案した研究[5]，労働集約型産業であるアニメーション制作の効率化に着目し，映像化による品質向上を目指した作品の電子化に関する研究[6]，などが存在する．

本研究においては，労働集約型産業であるコンサルティングサービスに着目し，コンサルタントの能力に依存して提供されるサービス品質が変動する現状に対し，コンサルタントを支援することにより，質の安定したサービスを提供可能とすることが重要であると考える．コンサルタントの業務の流れについて図1.1に示す．コンサルタントは，クライアント企業への訪問や対話，これまでのコミュニケーション内容の振り

返りなどにより，クライアント企業が抱える問題や状態を認識する．状態認識をおこなったのち，クライアント企業に対して適切であるコンサルティングサービスを判断し，サービスの提と実行に移る．本研究では，このようなコンサルタントのコンサルテーション過程における「状態認識」と「判断実行」のうち，クライアント企業の状態認識を支援対象とする．これは，クライアント企業の抱える問題を認識・予測できない限りコンサルティングサービスを提案することが困難であること，多様な情報から多種の問題発生の予測を考慮する必要がありコンサルタントの勘や経験に依存する部分であること，コンサルティングサービスの判断はコンサルティング企業によりサービス内容が異なり普遍的な判断基準を作成することが困難であることによる．

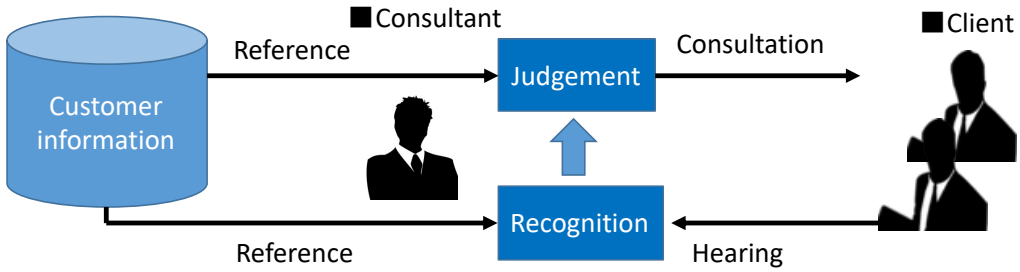



図 1.1: Workflow of consulting service

コンサルタントがクライアント企業の状態認識をする際，クライアント企業への訪問や対話，これまでのコミュニケーション内容の振り返り，クライアント企業情報などを参考とするが，本研究ではクライアント企業とのコミュニケーション内容を記したテキストデータに着目する．テキストデータを計算機によって解析し，問題発生と関連の強い内容の抽出，問題発生の予測をおこなうことで，コンサルタントのクライアント企業状態認識を支援することが可能となると考えられる．近年，テキストデータの解析手法として，テキストマイニング技術が注目されている．次節ではテキストマイニングについて説明する．

### 1.1.2 テキストマイニングの適用

#### テキストマイニングの特徴

コンピュータの処理能力の向上と安価なメモリによって大量のデータを対象にした分析が行えるようになってきた．その中でも従来行われてきた分析は，数値化・コード化されたデータを対象にその規則性を発見しようとするデータマイニングであった．しかし企業内には数値化，コード化されたデータよりも多くのテキストデータが存在している．このテキストデータはそこに何が書かれるか予め規定されていないため，従来見過ごしていた規則性を発見できる可能性があり，近年その活用ニーズが高まって

いる．このような市場のニーズとコンピュータの処理能力の向上の双方から注目されるようになった技術がテキストマイニングである[7]．Harset[8]は．テキストマイニングを「テキストデータを分析することにより，いままで誰も知らなかったような知識を発見すること」と定義し，テキストマイニングの概念に相当するものとして，探索的データ解析と文書分類の二つを挙げている．探索的データ解析とは，未知の情報を発見したり，現在答が知られていないような質問への回答を探すプロセスである．文書分類とは，文書の特定の内容をあらかじめ定義されたラベルの集合のいずれかに要約し，テキストデータ内の傾向やパターンを発見するプロセスである．また，喜田[9]は，「テキストマイニングは多様であり，一般的な定義がないといえる．つまり，それを実践する者にとって，関連領域，方法論の適用範囲をある程度自由に組み合わせ行っているというのが現状であると思われる」とし，関連領域分野として「1,自然言語処理あるいは計算機言語学，2.人工知能，エキスパートシステム，知識工学，3.認知科学および認知モデリング，4.計量言語学および計量文献学，5.言語学，社会学行動科学，6.記号論，テキスト論，カテゴリー論，意味論，7.内容分析あるいはテキスト分析」などを述べている．実践的分野においても受け入れられやすい技術となっており，経済学[10]，経営学[11]，教育学[12]，医学看護学[13]，哲学[14]，社会学[15]，文学[16]，化学[17]などの幅広い分野[18]で使用されている．またテキストマイニングにおいて主に対象とされるテキストデータはBlog Mining，Email Mining，WEB Mining，Social Media，Published Articles，Meeting Transcriptsなどに分類される[19]．

### ビジネス界におけるテキストマイニングの用途

1998年～2013年において「テキストマイニング」という言葉が掲載されている学術論文の中で，経済･経営学分野は合計で7%程度と比較的小さな割合を占めるに過ぎない．また新聞･雑誌検索結果をもとにテキストマイニングの対象データや用途が記事内容を調査した結果，コールセンター情報など（54%），営業日報(10%)，アンケート(4%)，電子メール(4%)，社内文書(3%)，ネット情報(21%)，特許情報(2%)，論文(2%)である．コールセンター情報や営業日報などの企業の内部データ，ネット情報などの外部データをもとにユーザー情報を解析することにより，需要開拓のヒントや，自社製品への不満点を分析するといった用途に活用しようとする様子がうかがわれる[20]．

テキストマイニングの利用が盛んになることに伴い，関連するソフトウェアやサービスが開発されている．このテキストマイニングツールの機能は以下の9種にまとめられる[21]．

- 単語や概念の抽出
  自然言語処理の前処理的な機能である．文章中の単語や概念を抽出して，リスト表示する．品詞や種類の特定をする．

- 単語や概念マップの生成
  単語や概念間の意味的関係を考慮して，マップの形でグラフィカルに表示する．文章全体の傾向を把握することに有効である．

- 類似文書検索
  与えられた文書に類似する文書を探す．

- 文書要約
  与えられた文書を要約する．文書の長さ指定や必要・不必要な単語リストの指定によりユーザーの目的に応じた要約を生成する．

- 文書分類
  あらかじめ用意したカテゴリに文書を分類する．

- 文書クラスタリング
  カテゴリを用意せずに文書をクラスタリングするう．予測しない分類を発見できる可能性があるが，生成されたクラスタの意味付けは分析者に依存する．

- シソーラスによる検索語展開
  シソーラスを用いて，検索語・同義語・関連語を追加する．

- 検索結果のグラフィカル表示
  検索結果をグラフィカルに表示し，注目する文書へのアクセスを容易にする．

これらの機能はテキスト分析を効率化することができるが，対象とするデータの選択・適当である分析手法の選択・分析結果からの解釈は分析者に依存する．

### 代表的研究

テキストマイニングを経済・経営学分野に応用した代表的な研究を紹介する．米国ではSecurities and Exchange Surveillance Commission規則により Management Discussion & Analysis of Financial Condition and Results of Operation の開示が求められるようになったことから，これらのテキストデータを科学的に解析し，企業評価に応用しようとする研究が積極的に行われるようになった．例として，経営者が株主に発送する手紙の文面と株価との対応関係を分析したClatworthy and Jonesの研究がある[22]．また，Yahoo! ファイナンスとRaging Bullに投稿されたメッセージを分析し，これらのメッセージの株価への影響が統計的に有意であることを検証したAntweler and Frankの研究がある[23]．さらに日本では，有価証券報告書記載のテキスト情報から企業の倒産予測を分析した荒川らの研究や[24]，金融経済月報を用いて金融市場における価格予想モデルを構築した和泉らの研究[25]，新聞記事を対象としマネジメントのリスク概念がどのようなコン

テクストで用いられているか解析した研究[26]がある．これらの研はの対象データは不特定多数の人々に読まれることを想定し決められたフォーマットに基づいて書かれているという点でテキストデータが均質であり，特定の内容について書かれているため対象問題と関係のあるものに絞られているというテキストであるという特徴をもつ．

上記の文章と対照に，不均質なテキストデータとしてコールセンターの文章や営業日報が挙げられる．これらの不均質なテキストデータには以下の特徴がある．

- 顧客との直接のコミュニケーション内容が記されていて，幅広い内容を含む．
- くだけた文章が多く書き手（話し手）による表記ぶれや誤字・脱字が存在し未知語や文法的な誤りが多いため，自然言語処理の観点から扱いにくい．
- 箇条書きやメモ書きが多く，テキストを構文的に解釈することが困難である．

以上の特徴により不均質なテキストデータを対象とする場合，自然言語処理の前処理・辞書整備や事前の特徴抽出が重要である．不均質なテキストデータを対象とした研究として，コールセンターの文章を文書分類の技術を用いて顧客からの電話をどの部門の担当者につなぐべきであるか判別した研究[27]，コールセンターの問い合わせ内容を事前に自動的に分類してから分析する研究[28]，コールセンターの文書から重要性の高い語句を抽出し語句間の関連を可視化する研究[29]，コールセンターの文章を対象とし概念抽出・統計的マイニング技術・分析結果の視覚化を可能とするインターフェースから構成されるテキストマイニングシステムを提案する研究[30]，営業日報を対象にキーとなる概念を抽出し因果関係を有する構造を抽出するという2段階のプロセスからなる情報抽出方法を提案し成功事例・機会損失事例を分析する研究[31]などが挙げられる．先行研究では，単一のメディアによる文章を対象としており，コールセンターだけでなく，業務日誌やメールなど，さまざまな記録を分析した研究はほとんど見当たらない．

本研究では，コンサルティング企業に蓄積されているクライアント企業との面談内容やコールセンターの相談内容など，多様な記録による多種の不均質なテキストデータを対象とする点に特徴がある．また，クライアント企業の状態認識において予測が必要であるを問題は，問題発生が確実に認識できる問題と，問題発生が認識できない場合のある不確実性を含む問題の，構造の異なる2種類の問題が存在する．これらの構造の異なる2種の問題予測を対象とするため，汎用的な問題に対して予測可能となる手法が必要である．

### コンサルティングサービス支援への適用

コンサルティングサービスにおけるクライアント企業の状態認識支援をおこなうため，テキストデータから問題発生を予測する手法が必要である．クライアント企業と

のコミュニケーションを記した，コンサルティング企業内に蓄積されているテキストデータテキストデータはコンサルタントとクライアント企業との面談内容やコールセンターの相談内容など，多様な記録によるテキストデータであるためデータが非均質である．また外部に公開されている対象問題に絞られた企業情報ではないので，企業についてより多様な情報が記されているという点で対象データに特徴がある．対象データに記されている相談内容は多岐に渡るため，テキストデータ内に出現する全ての語句を用いて解析をおこなうと無駄な変数が多く計算コストもかかる．よってあらかじめ対象問題に関連のある語句を抽出する必要がある．そこで本研究では状態認識のために分析を行う前に，変数として用いる語句を抽出する手法として対応分析を用いる．テキストデータの解析に対応分析を用いた研究として，災害支援者が支援に対して持つイメージを解析した研究[32]，バスケットボールクリニックの参加者の感想を分析することにより，参加者の主体性に関して分析した研究[33]などがある．これらの研究では対応分析の結果を分析に用いるが，対応分析は各サンプルの相関を可視化する技術であるので結果の解釈は人に依存する．本研究では対応分析の各サンプルの相関関係を捉えることのできる点に注目し，特徴的な語句を抽出する目的で対応分析を用いる．抽出された語句を用いて問題発生を予測するための判別式を作成する必要がある．まず予測対象となる問題について述べる．

コンサルタントがクライアント企業の抱える問題で察知しなければならない問題は多種存在するが，大きく2種に分類される．

- 確実である問題

  問題構造：クライアントが認知している問題・コンサルタントが後に認知する問題

  問題例：解約，企業業績

- 不確実性を含む問題

  問題構造：クライアントが企業内で発生を認知していない可能性がある問題

  問題例：横領，人事

確実である問題とは，クライアント企業が問題発生を認知しており，のちにコンサルタントが問題発生を認知するという点で正解データが確実である問題である．不確実性を含む問題は，横領問題などのクライアント企業が自社内で発生を認知できていない可能性のある問題を指し，正解データに不確実性を含む．この構造の異なる2種の問題を対象とすることにより，多様な問題が予測可能となることを目指す．問題発生の有無を予測するために判別手法を用いる．近年データマイニングの判別問題において機械学習を用いた手法が多く用いられている[34]．しかし機械学習は深層学習モデ

ルに代表されるように機械学習モデルが複雑なブラックボックスであるため解釈が困難であり[35]，XAI(Explainable AI)が求められている[36]．機械学習の解釈性について，局所的な説明，深層学習モデルの説明，説明の学習，説明法の見直し，説明の見直しに関する研究が多数存在し，数多くの説明法が提案されているが，どのような問題やデータにどの手法を用いるのが良いかという知見の蓄積は十分ではない．また，説明の信頼性についても評価基準が定まっていない[37]．本研究の対象とするコンサルティングサービスの支援においては，予測精度だけでなく，何が要因となり問題察知に至ったかという判別の要因となる語句や文章の解析が重要である．そこで，本研究では解釈可能な統計的手法である，DEA判別分析を用いる．この手法は母集団の分布について仮定を設ける必要がないため，本研究の対象とする非均質であり分布を判断することが困難なテキストデータを扱うことが期待できる．また，線形計画問題を解くだけで解を求めることができるので，大規模なデータにも対応できるなどの利点がある．

## 1.2 研究目的

本研究では人の能力によって変動するコンサルティングサービスの質を保つための支援システムを構築することを目的とする．構造の異なる問題を対象として問題予測手法を提案することにより，コンサルタントが察知すべき多様な問題に適用可能な支援システム構築を目指す．具体的には，解約問題と不正問題を対象とし，クライアント企業とのコミュニケーションを記したテキストデータを解析することにより，データに出現する語句からクライアント企業の問題発生を予測するための判別式を作成し，コンサルタントのクライアント企業に関する状態認識の支援を行う．以下に手順を示す．

- クライアント企業に蓄積されたテキストデータを問題発生の有無で分類する．
- 対応分析を用いて，問題発生と要因の強い語を抽出する．
- 抽出した語を用いて判別分析を行い，判別式を導出する．
- 判別分析から得られた式をもとに新たなクライアント企業の問題発生を予測する．

## 1.3 各章の位置づけ

本論文の構成は以下のとおりである．

第2章では，本論文の研究対象とテキストデータから問題発見を予測するための提案手法について説明する．

第3章では，解約問題を対象とし，実データの一部を用いて既存手法との比較により提案手法の有効性を検証する．

第4章では，解約問題を対象とし，実規模問題を解くための支援手法を提案し，拡張手法の有効性を検証するとともに実現場適用に向けてコンサルタントの予測結果との比較検証をおこなう.

第5章では，不確実性を含む不正問題を対象とし，提案手法の他問題への適用について検討する．実現場でのヒアリング結果と比較することにより，提案手法による実現場適用の可能性を評価する.

最後に，第6章で本論文の結論及び今後の展望を述べる.

各章の関係を図1.2に示す.

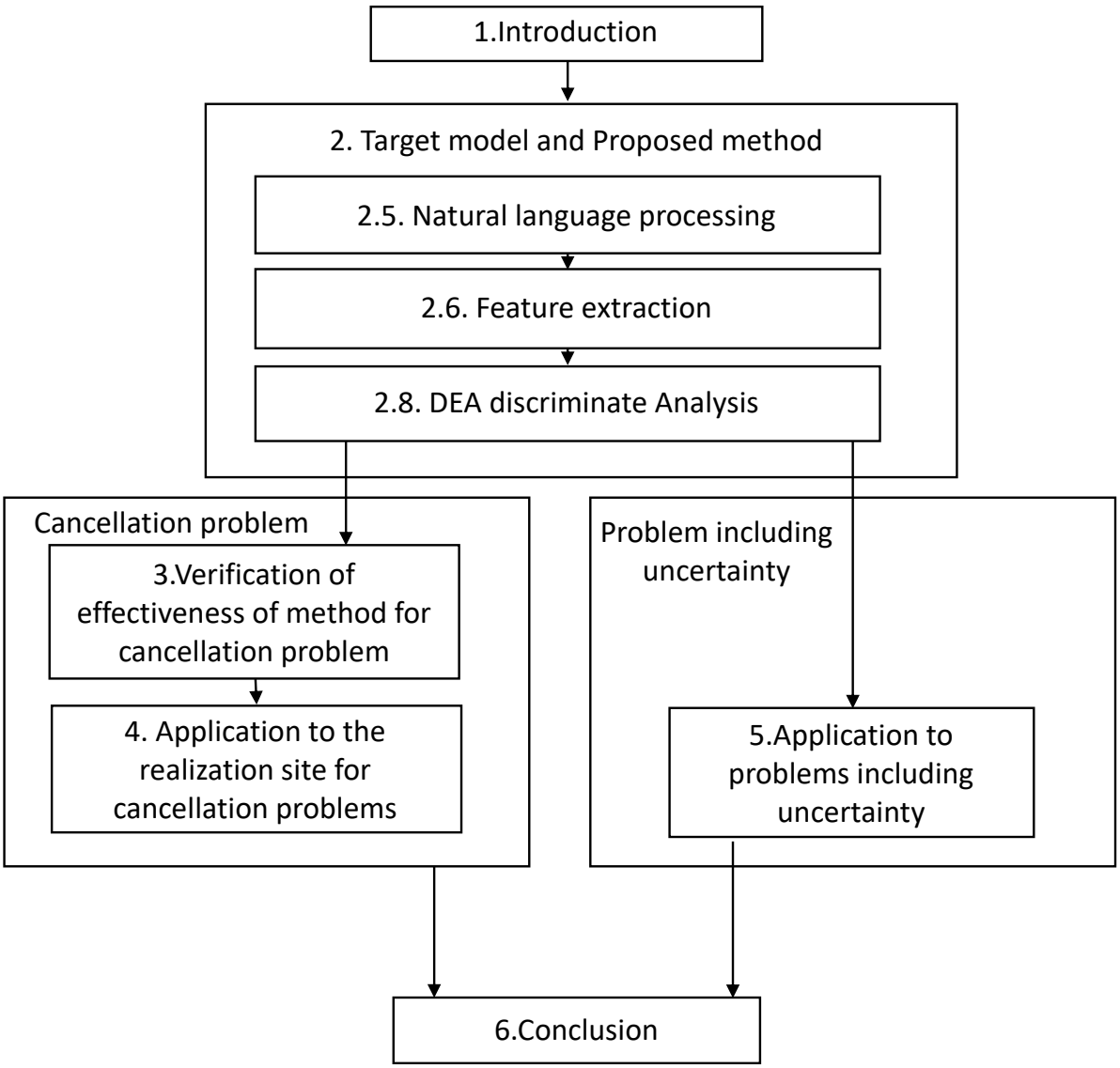



図 1.2: A relationship among chapter

# 第2章　テキストマイニングを用いた問題認識支援手法

## 2.1　緒言

本章では，第1章で述べた研究背景と目的を踏まえ，本研究の提案手法を説明する．2.2節では研究に用いるデータと状態認識の支援において対象とする構造の異なる2種類の問題について説明する．2.3節で提案手法の概要を説明する．2.4節で自然言語処理について，2.5節で特徴抽出について，2.6節で判別手法について，提案手法の各フェーズで用いられる手法を詳しく説明する．

## 2.2　研究対象

本研究ではコンサルティング企業に蓄積されているテキストデータを解析することにより，クライアント企業で起こりうる問題を予測する．コンサルティング企業において蓄積されているテキストの種類と内容を表2.1にまとめる．使用するテキストデータは，コンサルタントとクライアント企業の訪問時のやりとりの内容，メールでのクライアント企業からの質問内容，コールセンターでの対応内容など様々な方法によって記録されたデータである．また，不特定多数の人に読まれる想定で書かれるフォーマットの定まった文章でないという点で非均質なテキストデータである．

図2.1にテキストデータの例を示す．テキストデータには会社概要，訪問履歴，クライアント企業が現在抱えている問題，利用中のコンサルティングサービスなどについて詳細に記されている．よって，そのテキストの中には今後起こりうる問題を察知するきっかけとなる内容が記述されていると考えられる．現場のコンサルタントはこれらのテキスト情報などをもとに自身の経験や勘から問題察知をおこなうが，これは個人の能力に依存している．そこで本研究ではこれらのテキストデータを計算機で解析することにより，コンサルティングサービスの支援を行う．

研究の概念図を図2.2に示す．まずはテキストデータ内の記述を確認することにより，テキストデータを問題発生の有無で分類する．分類したグループ各々に対して提案手法を用いて解析することにより，問題発生の有無を判別する判別式を作成する．判別式の有効性を検証するため，新たなデータに対して作成した判別式を適用し予測を行う．

本研究では以下の2種の問題について提案手法を適用し，発生を予測する．

表 2.1: Contact method

| Text type | Context |
|---|---|
| FAX | サービスについて，申込受付，ひな形資料 |
| オンライン | オンラインの面談内容 |
| システム | サイトログイン数，アクセス数 |
| セミナー | セミナー参加について，アンケート |
| メール | 相談内容，サービスの診断結果 |
| 会員専用 | web版サービス内容 |
| 社員間 | 事務的内容，引継ぎ |
| 宅急便 | 送付内容 |
| 電話 | 相談内容，案内，不在時の記録 |
| 訪問 | 訪問時の記録 |
| 面談 | 会話中の内容 |
| 郵便 | 郵便手配，郵便着物内容 |
| 来社 | セミナー内容，相談内容 |

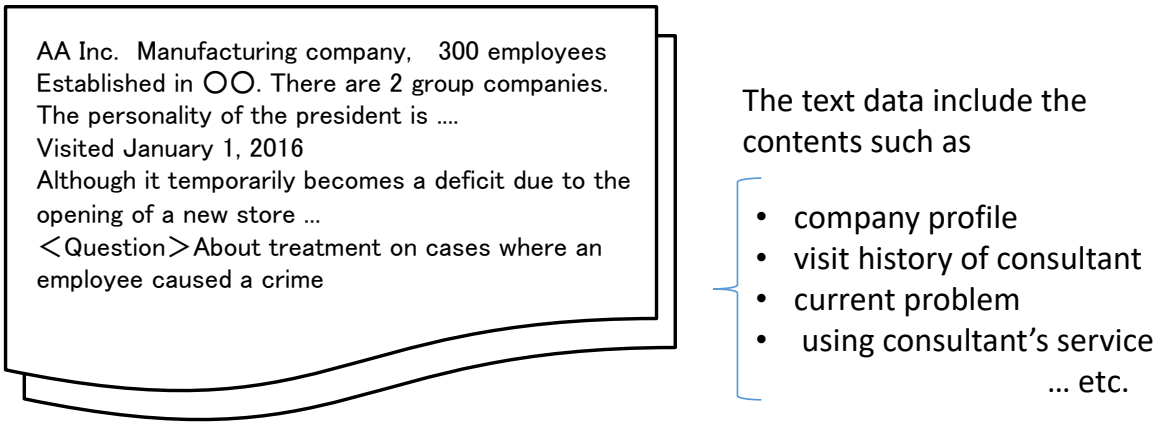



図 2.1: Example of target data

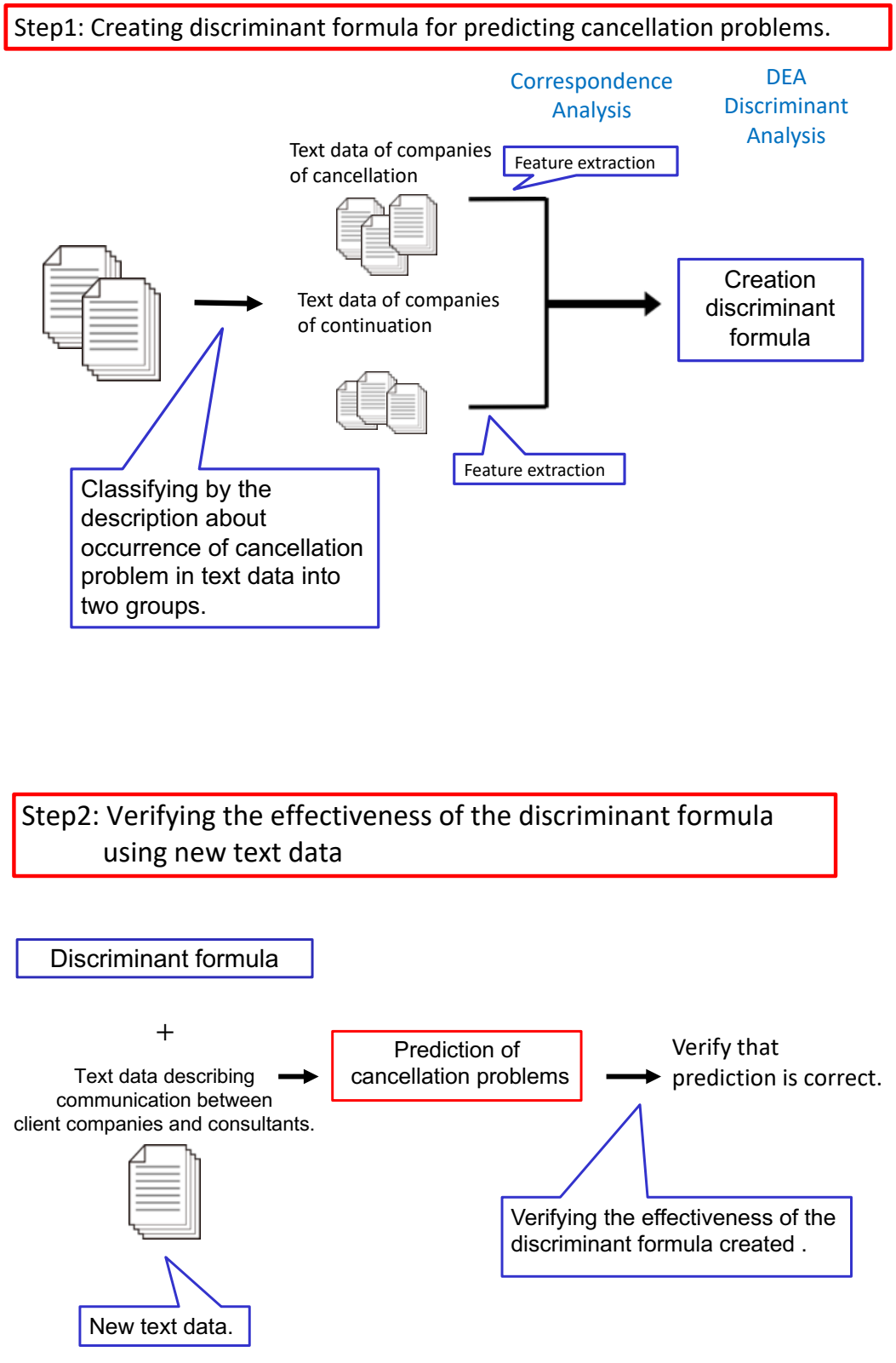



図 2.2: Overview of the research

1. 解約問題

2. 不確実性を含む不正問題

解約問題については，コンサルティングサービス契約済みの企業において解約を考えている企業を予測するための判別式を作成する．不正問題については特に横領に着目し，不正問題が発生する可能性のある企業を予測するための判別式を作成する．この2種の問題の構造について，確実な問題を図2.3に，不確性を含む問題を図2.4に示す．解約問題はある時点において，コンサルティングサービスを解約しているか継続中かが明確に分かるため，正解データが確実な問題である．一方，不正問題についてはある時点で不正問題が発生しているかどうかについて，コンサルタントに相談していないがクライアント企業内では問題が発生している場合，クライアント企業も気づいていないが実際は問題が発生している場合などが存在するので，テキストデータの内容から不正問題発生の有無は明確には分からない．よって判別式を作成する上での正解データに不確実性を含む問題となっている．これらの異なる構造である2種の問題を対象に問題発生を予測し提案手法の有効性を検証することにより，提案手法がコンサルタントが察知すべき多様な問題の発生予測に適用可能となることを目指す．

## 2.3 提案手法概要

テキストマイニングは，大量かつノイズを含むテキストデータを単語や文節で区切り，それらの出現の頻度や共同出現の相関，時系列などを解析することで有用な情報を取り出す分析方法である．テキストマイニングは解析を進めていく上での情報処理の流れに合わせて，以下の三つに分類することができる[38]．

1. 自然言語処理：自然言語で書かれた文書情報からその内容をあらわす概念を抽出

2. 分析・マイニング技術：抽出された概念を統計的に分析

3. 可視化技術：マイニング結果を人間が理解しやすい形に可視化し，対話的な分析を実現

本研究では自然言語処理で形態素解析，ノイズとなる語句の省略を行う．その後，形態素解析で得られた大量の語句の中からそれぞれのグループの中で要因の強い語句を抽出するため対応分析を用いる．表記ぶれを考慮するため類義語辞書を用いたデータの拡張を行い，それらの語句を変数に用いて判別し問題発生の可能性を予測する．このように，対応分析と判別分析を用いて解析することが本手法の提案である．図2.5に研究の流れを示す．

提案手法では，判別分析の前に対応分析と要因となる語の抽出をおこなう．これは，

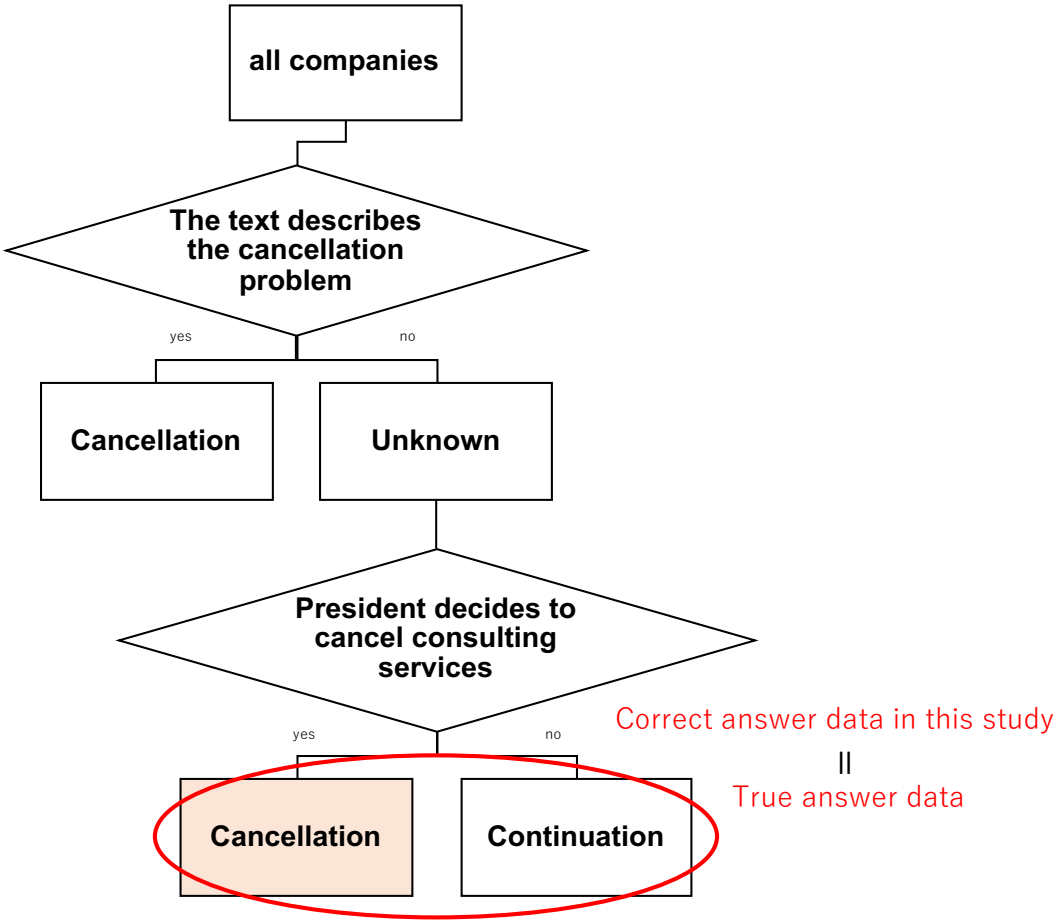



図 2.3: Termination problem

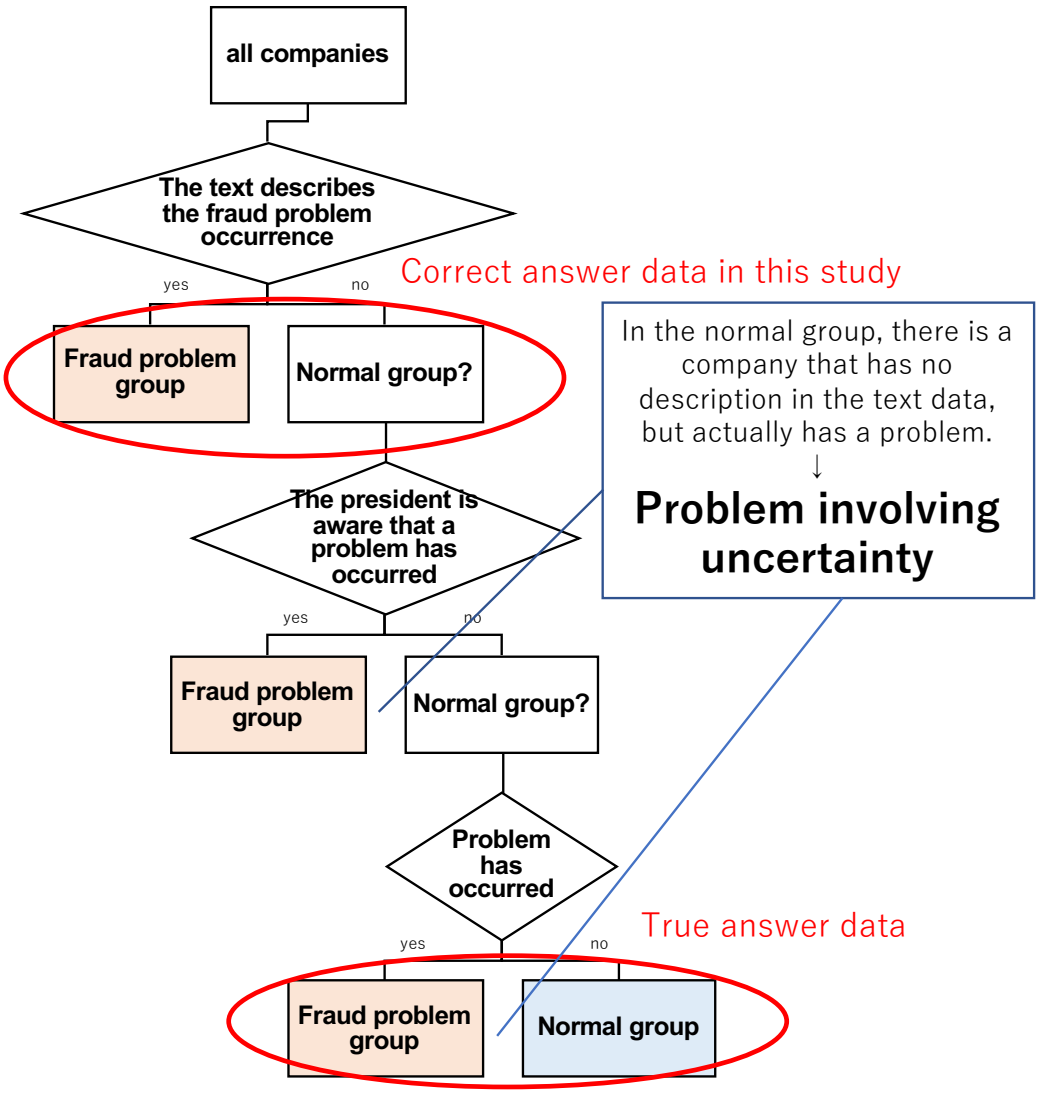



図 2.4: Fraud problem

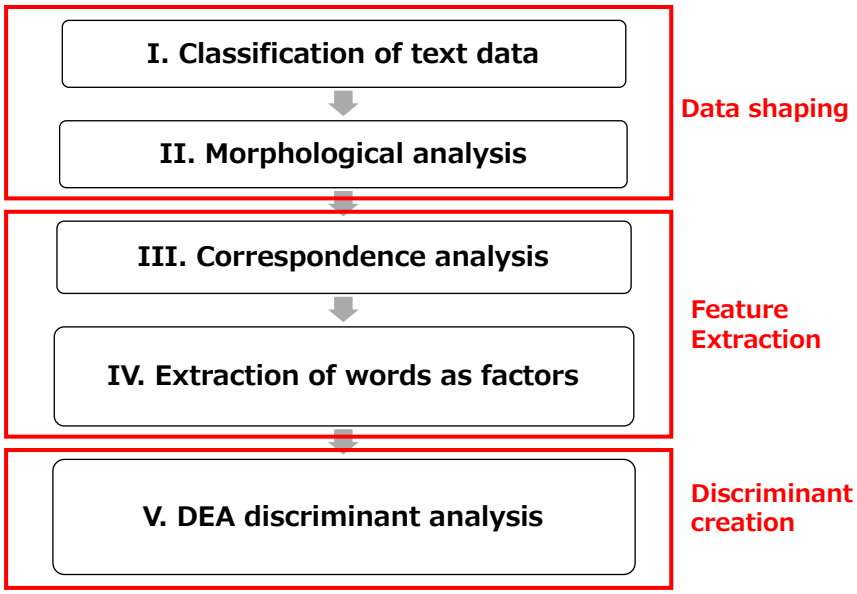



図 2.5: Flow of the research

テキストデータから得られるデータ量は膨大であり．判別分析で変数として用いる語を削減する必要があるためである．データが非均質である，企業ごとに文書量に差がある，グループ間の差を考慮する必要がある，などの課題点を解決するためにこれらの処理を提案する．また，判別するための手法としてDEA判別分析を用いている．これは，母集団の分布を設ける必要がないため実際の分布を把握することが困難なテキストデータをそのまま扱うことができることや，線形計画問題を解くだけで解を求めることができるので，大規模なデータにも対応できることなどの利点がある．これらの手法や処理を組み合わせた手法を，本稿では提案する．次章から提案手法について詳細に説明する．

## 2.4 自然言語処理

### 2.4.1 テキストデータの分類

本研究では，コンサルティング企業に蓄積された，クライアント企業とのコミュニケーションを記したテキストデータを用いて解析する．テキストデータは問題発生の時系列を考慮しながら，問題の発生の有無で分類する．本研究は問題発生の予測を行うため，問題に関する直接的な内容のテキストデータは削除する．また，問題発生時から一定の期間過去のテキストデータを用いる。これは問題発生から過去に遡りすぎると，テキストデータに問題発生の特徴は表れていないと考えられるためである．

### 2.4.2 形態素解析

形態素とはそれ以上細かくすると意味がなくなってしまう最小の文字列のことである．文を形態素に分解し，その品詞を特定することを形態素解析という．日本文は英文の様に空白で形態素を区切られていないため，形態素に分解する際に文の区切り方などで複数の形態素の抽出結果が得られる．そのため大量のテキストデータから形態素情報を取り出そうとする場合，コンピュータ上で形態素解析器を利用して一括処理するのが一般的である[39]．本研究では，工藤拓氏が開発したフリーの形態素解析器として公開されているMeCabを使用している[40]．MeCabは辞書とテキストデータに依存しない汎用的な設計であり，他の解析器より高速であるという特徴をもつ．

例として「りんごを買いました」を形態素解析すると

りんご　(名詞，一般)
を　　(助詞，格助詞，一般)
買い　　(動詞，自立，五段・ワ行促音便，連用形)
まし　　(助動詞，特殊・マス，連用形)
た　　(助動詞，特殊・タ，基本形)

のように出力される．
本研究で使用する語句の品詞を以下に示す．

- 名詞
- 形容動詞
- 動詞
- 形容詞
- ナイ形容詞
- 感動詞
- 副詞
- 副詞可能

助詞，助動詞，記号などの特殊文字，未定義語について一語では意味を成さない語なので分析対象から除外した．本研究では形態素解析によってテキストデータを形態素に分解することにより，各単語の出現回数を計算し，その値をもとに分析を行う．

### 2.4.3 ノイズとなる語句の省略

分析のノイズとなる以下の項目の語句は省略した.

1. コンサルティングサービスの名称
   例:適性診断サービス

2. 定型文
   例:FAX送信済み

3. 意味をなさない語
   例:年,月,件

4. 予測する問題を直接表す語
   例:解約,横領,不正

コンサルティングサービスの名称は,サービスを利用している企業だけでなくサービスを告知するメールのテキストデータ内にも多数出現するため,省略した.定型文は,コンサルタントが業務上で用いる文章であり対象問題発生と関わりがないと考えられるため省略した.意味をなさない語は抽出されたのちに解釈,議論不可能なため省略した.本研究では対象問題を予測することが目的であるため,問題を直接表す語句は省略した.このように明らかにノイズとなる語句を削除したのちに,形態素解析の結果を用いて対応分析をおこなう.

## 2.5 特徴抽出

### 2.5.1 対応分析

#### 対応分析の概要

対応分析はフランスの研究者,Jean-Paul-Ben-zecriにより1960年代に提唱された方法であり,データ表の行や列に含まれる情報を少数の成分に圧縮する手法である[41].

#### 対応分析のモデル

対応分析の対象データは以下に示すような二次元のデータ表が基本である[42].

本研究では,行項目(サンプル)に語句を列項目(カテゴリ)に企業名を入れる.サンプル,カテゴリのそれぞれに重みとなる変数を設定し,このデータ表を基にそれらの変数を算出することにより対応関係を可視化する.

##### 記号定義

表 2.2: Correspondence analysis

| | Word | AAA | BBB | $\cdots$ | KKK |
|---|---|---|---|---|---|
| Company name | | $a_1$ | $a_2$ | $\cdots$ | $a_K$ |
| A Co. | $b_1$ | $t_{11}$ | $t_{12}$ | $\cdots$ | $t_{1K}$ |
| B Co. | $b_2$ | $t_{21}$ | $t_{22}$ | $\cdots$ | $t_{2K}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| N Co. | $b_N$ | $t_{N1}$ | $t_{N2}$ | $\cdots$ | $t_{NK}$ |

対応分析に用いる記号を以下に定義する.

$i$ : 単語の番号 $(i = 1, 2, \cdots, K)$

$j$ : 企業の番号 $(j = 1, 2, \cdots, N)$

$t_{ij}$ : 単語 $i$ が企業 $j$ のテキストデータに出現する回数

$a_i$ : 単語 $j$ の係数となる重み (サンプルスコア)

$b_j$ : 企業 $j$ の係数となる重み (カテゴリスコア)

このとき, 基礎統計量は以下のように計算できる.

$$\text{平均}\bar{a} = \frac{\sum_{i=1}^{K} a_i}{K} \tag{2.1}$$

$$\text{分散}V_a = \frac{\sum_{i=1}^{K}\left\{(\sum_{j=1}^{N} - t_{ij})(a_i - \bar{a})^2\right\}}{KN - 1} \tag{2.2}$$

$$\text{共分散}V_{ab} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N} t_{ij}(a_i - \bar{a})^2(b_i - \bar{b})^2}{KN - 1} \tag{2.3}$$

$$\text{相関係数}R = \frac{V_{ab}}{\sqrt{V_a}\sqrt{V_b}} \tag{2.4}$$

対応分析は $\bar{a}, \bar{b} = 0$, $V_a, V_b = 1$ の条件のもとで式 (2.4) の相関係数 $R$ が最大となるようなサンプルスコア・カテゴリスコアを求めることが目標である. ここでラグランジュの乗数法を用いて式 (2.5) をたてることができる.

$$G = V_{ab} - \frac{\lambda}{2}(V_a - 1) - \frac{\mu}{2}(V_b - 1) \tag{2.5}$$

式 (2.5) をサンプルスコア・カテゴリスコアで偏微分し, 0 とおくと以下の連立方程式が求められる.

$$\begin{cases} \frac{\partial G}{\partial a_1} = 0 \\ \quad\vdots \\ \frac{\partial G}{\partial a_N} = 0 \end{cases} \qquad \begin{cases} \frac{\partial G}{\partial b_1} = 0 \\ \quad\vdots \\ \frac{\partial G}{\partial b_N} = 0 \end{cases} \tag{2.6}$$

式(2.6)と$\lambda = \mu$を用いて変形することにより，以下の固有値問題に帰着できる．

$$\begin{bmatrix} X_{11} & \cdots & X_{KK} \\ \vdots & \ddots & \vdots \\ X_{K1} & \cdots & X_{KK} \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix} = \lambda^2 \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix} \qquad (X_{11} \cdots X_{KK}\text{は任意の数}) \qquad (2.7)$$

式(2.7)を解くことにより固有値$\lambda^2$を求め，それぞれの固有値に対してサンプルスコア・カテゴリスコアの値を求める．ここで求めた固有値とは相関係数の2乗でありこれを次元と呼ぶ．以上の操作により多数あったサンプル・カテゴリが少数の成分に圧縮される．

### 2.5.2　対応分析の可視化

計算で得られた固有値を軸として，その固有値に対応するサンプルスコア，カテゴリスコアを散布図に布置することにより対応関係を可視化できる．図の原点から離れる

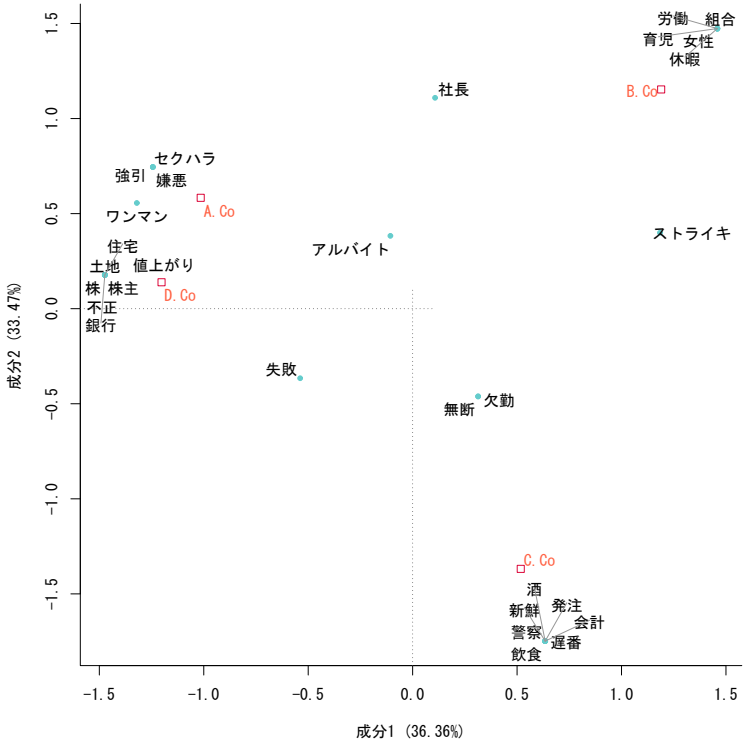



図 2.6: Sample of Correspondence Analysis

ほど，対象としたデータの中で特徴のある企業・語句が現れる．一方，原点付近には対象データの中で一般的な企業や語句があらわれるため，原点付近の語が対象グループ全体の特徴を強く表す．

図2.6に散布図の例を示す．図2.6は不正問題が起こった企業四社に対して分析を行った結果である．原点付近に“アルバイト”，“失敗”などが出現しているので，この四社

全体の特徴，つまり不正問題の発生が起こる企業には“アルバイト”や“失敗”という語句が出現する傾向があるとわかる．一方で原点から離れてC社のタグの近くに“飲食”や“酒”，“暴行”，“警察”などの語句があることから，“C社は飲食店を経営していて暴行事件が起こったことがある”などが読み取れる．これらの語句はC社のみに関係の強い語句であるため不正問題の要因となる語としては抽出しないようにする．そこで本研究ではテキストデータを問題発生の有無でグループに分けて，それぞれのグループに対して対応分析を行い，原点に近い語に注目し次の小節で示す方法で各グループの要因となる語を抽出する．

### 2.5.3 両方のグループに出現する要因となる語句の抽出法

本研究では各グループにおいて対応分析の原点に近い語をそのグループの要因となる語として抽出する．抽出方法を説明するため，例として図2.7を示す．左が不正問題発覚ありのグループの対応分析の結果，右が不正問題発覚なしグループの対応分析の結果を表している．不正問題発覚ありのグループに着目したとき，原点付近の語がグループ全体の要因を表す語なので“President”，“Part time job”が抽出される．しかし“President”は不正問題発覚なしグループにおいても原点付近に出現している．よって“President”は全テキストデータにおいて一般的に出現する語句であり，不正問題発覚ありグループの要因となる語として抽出されるべきでない．このように原点に近い語の中には，相手グループにおいても原点に近い語，つまり全テキストデータの中で一般的に使われている語が含まれている．このような語句を抽出しないようにするために，グループ間の差を考慮する抽出法を提案する．

計算に用いる記号を以下に示す．

記号定義

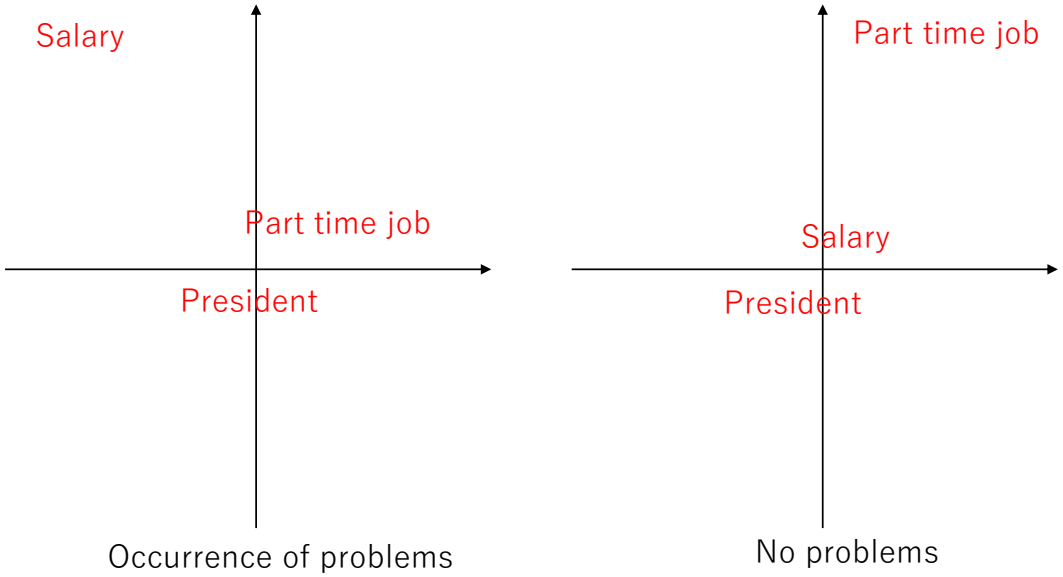



図 2.7: Sample of extracted words

$$
\begin{aligned}
N &: \text{最小出現回数} \\
i &: \text{単語数} \\
j &: \text{成分数} \\
W_i &: \text{テキストに出現する単語} \\
G_1, G_2 \in G &: \text{グループ集合} \\
D_G &: \text{グループ} G \text{における成分の総数} \\
x_{ijG} &: \text{グループ} G \text{における成分} j \text{の単語} W_i \text{の値} \\
d_{iG} &: \text{グループ} G \text{にける単語} W_i \text{の原点からの距離} \\
e_{iG} &: \text{更新された} d_{iG} \text{値} \\
C_{jG} &: \text{グループ} G \text{における成分} j \text{の寄与率}
\end{aligned}
\tag{2.8}
$$

対応分析の結果から全成分を考慮し，単語と原点の距離$d$を算出する．

$$
d_{iG} = \sqrt{\sum_{j=1}^{D_G} \{(x_{ijG} * C_{jG})^2\}} \tag{2.9}
$$

それぞれの単語に関して以下の操作によりグループ間の差を考慮して$d$値を$e$値に変

換することにより，上述した問題を解決できる．

$$\begin{cases} d_{iG_1} < d_{iG_2} \text{のとき} \\ e_{iG_1} = \frac{d_{iG_2}-d_{iG_1}}{d_{iG_2}} \\ e_{iG_2} = \infty \end{cases} \tag{2.10}$$

$$\begin{cases} d_{iG_2} \geq d_{iG_1} \text{のとき} \\ e_{iG_2} = \frac{d_{iG_2}-d_{iG_1}}{d_{iG_1}} \\ e_{iG_1} = \infty \end{cases} \tag{2.11}$$

各グループについて$e$値の大きい語句から一定数を抽出してDEA判別分析の変数に用いる．

### 2.5.4 外れ値の考慮

要因となる語は各グループの差によって抽出する．よって片方のグループで特定の企業のみで極端に使われている語句は，本来要因の強い語句でないが相手グループの特異値が影響して抽出されてしまうという問題がある．そこで，Smirnov-Grubbs 検定[43]を行い外れ値となる語句を削除する．Smirnov-Grubbs検定とは，一般的に使用される外れ値の検定方法である．以下に帰無仮説と対立仮説を示す．

- 帰無仮説：すべてのデータは同じ母集団に属する
- 対立仮説：データのうち平均値から最も外れたものは外れ値である

統計量は式（2.12）で，有意点は式（2.13）で求められる．

**記号定義**

$z$ ： 統計量
$\tau$ ： 有意点
$N$ ： データ数
$x_i$ ： 外れ値とみなすべきかどうか考えているデータ
$\bar{x}$ ： データの平均値
$t$ ： 自由度$N-2$の$t$分布
$\sigma$ ： 標準偏差

$$z = \frac{|x-\bar{x}|}{\sigma} \tag{2.12}$$

$$\tau = \frac{(N-1)t}{\sqrt{N(N-2)+nt^2}} \tag{2.13}$$

$t \geq \tau$の時，帰無仮説が棄却され$x_i$は外れ値として検出される．この式は再帰的に使う．つまり，最も外れた1標本のみを検定し外れ値と判定されたら，それを除外した$N-1$個の標本を使って2番目に外れた標本を検定し，外れ値が検出されなくなるまで繰り返す．

## 2.6 判別分析による不正予測

前節までの手法により抽出された語句を用いて，問題発生を予測するための判別をおこなう．問題発生の有無を予測するために判別手法を用いる．近年データマイニングの判別問題において機械学習を用いた手法が多く用いられている[34]．しかし機械学習は深層学習モデルに代表されるように機械学習モデルが複雑なブラックボックスであるため解釈が困難であり[35]，XAI(Explainable AI)が求められている[36]．機械学習の解釈性について，局所的な説明，深層学習モデルの説明，説明の学習，説明法の見直し，説明の見直しに関する研究が多数存在し，数多くの説明法が提案されているが，どのような問題やデータにどの手法を用いるのが良いかという知見の蓄積は十分ではない．また，説明の信頼性についても評価基準が定まっていない[37]．本研究の対象とするコンサルティングサービスの支援においては，予測精度だけでなく，何が要因となり問題察知に至ったかという判別の要因となる語句や文章の解析が重要であるため，可読性のある判別式作成手法に着目する．

### 2.6.1 判別分析の概要

判別分析とは，複数の要因から特徴づけられているデータが，いくつかの母集団に分類可能であることが明らかな場合に，どの母集団に分類されるかをその観測値から正しく分類・予測する手法である[44]．判別分析は大きく分けて統計的な手法と目標計画ベースの手法がある．さらに目標計画ベースの手法から発展したDEAによる手法がある．以下に三種の手法の長所・短所を示す[45]．

1. 統計的判別分析法

   s 長所:一般によく使われている誤差の正規分布を仮定せずに判別係数が求められる．

   短所:(1)母平均，母分散が既知である場合はあまりなく，標本平均，標本分散を用いるので，判別の信頼性が低くなる可能性がある．(2)グループ間にオーバーラップ(グループ間が重なり合った部分)が存在する場合，それを見つけるのが難しい．

2. 目標計画法にもとづく判別分析法

   長所:(1)統計的判別分析法で必要とされた標本平均，標本分散を使わなくて済む．(2)誤判別リスクを最小化できる．(3)判別係数の推定が$L_1$-ノルムのアルゴリズム

で可能になり，大規模なデータでも十分対応できる．
短所:線形判別関数を仮定しているために，オーバーラップが存在する場合に，その存在を見つけにくくなったり，判別の精度が低下する可能性がある．

3. DEA判別分析法
長所:(1)計算を2段階にすることによって，Stage 1でオーバーラップの存在を確認でき，stage 2でその対処を行える．(2)目標計画法にもとづく判別分析法の長所を引き継ぐことができる．
短所:DEA判別分析法では判別係数がゼロになることがある．

以上のようにDEA判別分析は母集団の分布について仮定を設ける必要がないので，実際の分布を把握することが困難なテキストデータをそのまま扱うことができる．また，標本サイズについても前提条件を必要としないので，グループ間に文書量の差があるという分析する上での問題をクリアできる．さらに線形計画問題を解くだけで答えを求めることができるので，大規模なデータにも対応できるという点で今後の応用を考える上でも利点がある．以上の理由からテキストデータを利用して不正予測を行う手法としてDEA判別分析を採用した．

### 2.6.2 DEA判別分析

DEA判別分析法は末吉ら[46]の研究によって提唱された判別モデルである．DEA判別分析は二段階で行われる判別分析法であり，第一段階で誤判別のデータとどちらに分類されるか不明なデータを検出する．第二段階では第一段階で検出されたデータに対して判別分析を行い分類することで，その判別の精度を高めている．以下にDEA判別分析のモデルを示す.

**記号定義**

1 定数

$i$ ： 企業を特徴づける要因$(i = 1, 2, \cdots, k)$
$j$ ： 企業$(j = 1, 2, \cdots, n)$
$z_{ij}$ ： 企業$j$における要因$i$の出現回数
$G_1, G_2$ ： グループ
$\eta$ ： オーバーラップ部分の幅

2 決定変数

$\lambda_i$ ： 判別係数

$d$ ： 判別境界

$S_{ij}^+, S_{ij}^-$ ： スラック変数

**Stage 1**

$$
\begin{aligned}
\min \quad & \textstyle\sum_{j\in G_1} S_{1j}^+ + \sum_{j\in G_2} S_{2j}^- \\
\text{s.t.} \quad & \textstyle\sum_{i=1}^k \lambda_i z_{ij} + S_{1j}^+ - S_{1j}^- = d + \eta \quad (j \in G_1) \\
& \textstyle\sum_{i=1}^k \lambda_i z_{ij} + S_{2j}^+ - S_{2j}^- = d \quad (j \in G_2) \\
& \textstyle\sum_{i=1}^k |\lambda_i| = 1 \\
& S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \geq 0 \qquad d : \text{制約なし}
\end{aligned}
\tag{2.14}
$$

式(2.14)の目的関数は誤判別を最小化するようになっている．判別境界に$\eta$の幅を持たせることで，グループ1，グループ2のどちらに判別されるか不明なデータを洗い出すことができる．Stage 1のイメージ図を図2.8に示す[24].

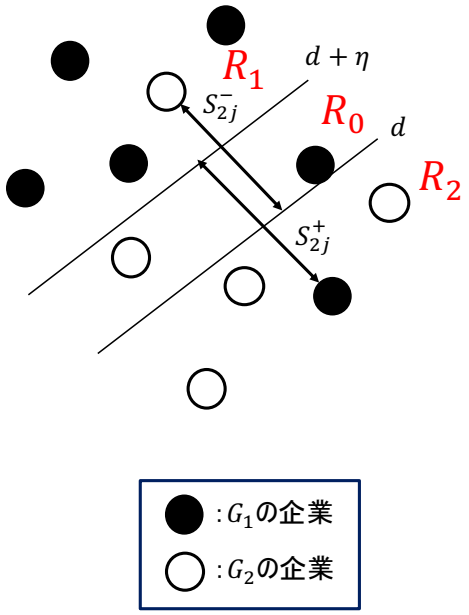



図 2.8: DEA Stage1

Stage 1で得られた最適解を$\lambda_i^*$と$d^*$としたとき$z_{ij}$は次の判別基準によって五種類に分

類される．

$$R_1 = \left\{ j \in G \middle| \sum_{i=1}^{k} \lambda_i^* z_{ij} \geq d^* + \eta \right\} \tag{2.15}$$

$$R_0 = \left\{ j \in G \middle| d^* + 1 > \sum_{i=1}^{k} \lambda_i^* z_{ij} > d^* \right\} \tag{2.16}$$

$$R_2 = \left\{ j \in F \middle| d^* \geq \sum_{i=1}^{k} \lambda_i^* z_{ij} \right\} \tag{2.17}$$

$$C_1 = \{ j \in R_1 | j \in G_1 \} \tag{2.18}$$

$$C_2 = \{ j \in R_2 | j \in G_2 \} \tag{2.19}$$

$C_1, C_2$は正しく分類されたことを示している．誤判別となったデータ集合は$G_1 \cap R_2, G_2 \cap R_1$であり，集合$R_0$はオーバーラップ領域にあるデータである．Stage 2ではこの誤判別となったデータとオーバーラップ領域に存在するデータへの対処を行う．$c$は$d^*$と$d^* + \eta$の間の新しい境界判別値である．

**Stage 2**

$$\min \quad \sum_{j \in G_1 \cap (R_0 \cup R_2)} S_{1j}^+ + \sum_{j \in G_2 \cap (R_0 \cup R_1)}$$

$$\begin{aligned}
\text{s.t.} \quad & \sum_{i=1}^{k} \lambda_i z_{ij} \geq d + \eta && (j \in C_1) \\
& \sum_{i=1}^{k} \lambda_i z_{ij} + S_{1j}^+ - S_{1j}^- = c && (j \in G_1 \cap (R_0 \cup R_2)) \\
& \sum_{i=1}^{k} \lambda_i z_{ij} + S_{2j}^+ - S_{2j}^- = c && (j \in G_2 \cap (R_0 \cup R_1)) \\
& \sum_{i=1}^{k} \lambda_i a_{ij} \leq d && (j \in C_2) \\
& \sum_{i=1}^{k} |\lambda_i| = 1 && \\
& d \leq c \leq d + \eta && \\
& S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \geq 0, && \lambda_i, c, d : \text{制約なし}
\end{aligned} \tag{2.20}$$

Stage 1で正しく判別されたデータについては制約式で制御して，Stage 1の結果を保障している．目的関数ではStage 1で正しく判別されたデータのスラック変数が除かれているため，Stage 1で誤判別されたデータがStage 2においても誤判別されたときのスラック変数の合計が最小化されている．また，判別境界値$c$は$d$と$d + \eta$の間をとるように設定されており，オーバーラップ領域に存在していたデータの判別が可能となる．Stage 2のイメージ図を図2.9に示す．

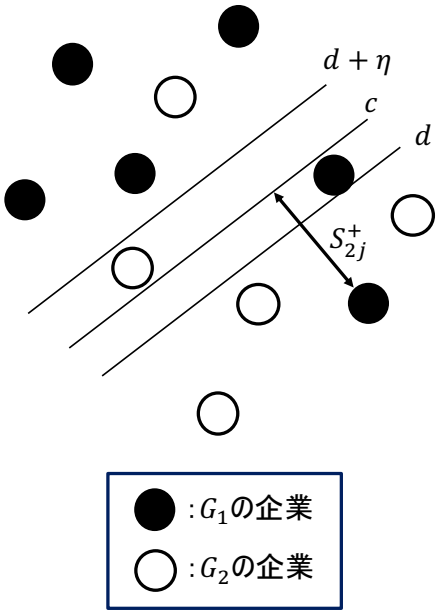



図 2.9: DEA Stage2

Stage 2の最適解を$c^*, \lambda_i^*$とすると次の基準によって判別される.

$$\sum_{i=1}^{k} \lambda_i^* z_{ij} \geq c^* \quad \text{のとき} j \text{番目のデータは} G_1 \text{に属するものと判別される.} \tag{2.21}$$

$$\sum_{i=1}^{k} \lambda_i^* z_{ij} < c^* \quad \text{のとき} j \text{番目のデータは} G_2 \text{に属するものと判別される.} \tag{2.22}$$

## 2.7 結言

本章では本研究で用いる提案手法について述べた.

次章では提案手法の有用性を検証するため, 解約問題に対して提案手法を適用し計算機実験を行う.

# 第3章　解約問題を対象とした手法の有効性検証

## 3.1　緒言

本章では解約問題を対象とし，提案手法を用いて実データを解析する．具体的には，クライアント企業のコンサルティングサービス解約問題を予測する．本章では，データの特徴の解析や提案手法の有効性を検証するため，実規模問題の一部のデータを用いて計算機実験をおこなう．3.2節では実験条件を述べ，3.3節では要因となる語の抽出における提案手法の有効性を確認する．3.4節では3.3節の結果を用いて判別式の作成，新たなデータに対して予測を行う．3.5節では，提案手法の有効性を検証するため，既存手法と比較する．3.6節では実験結果についての考察を述べる．最後に，3.7節で本章についてまとめる．

## 3.2　対象問題

提案手法を用いて，クライアント企業のコンサルティングサービス解約を予測する判別式を作成する．まず，テキストデータの分類について説明する．

### テキストデータの分類

実験対象とするコンサルティング企業では，各クライアント企業に対して一年に一回コンサルティングサービスについて契約更新または解約が行われる．本研究では提案手法を用いて訪問履歴などのテキストデータを解析することにより，クライアント企業が解約を考えていることを予測する．解約を早期に予測することができればクライアントの不満に気づき，より要望に沿ったサービスを提供し，サービスの継続利用につなげることができると考えられる．本章ではデータの特徴解析や提案手法の有効性を検証するため，実規模問題の一部のデータとして，契約期間が五年以上の企業を扱う．これは契約期間が短い企業は，契約の動機と提供可能なコンサルティングサービスとの乖離や企業相性など，予測できたとしても解決できない原因が多く存在するためである．また期間が短いと担当するコンサルタントの人数が少ないため，コンサルタント個人の能力が解約に影響する可能性がある．契約期間が一定年数の企業は，定期的に担当コンサルタントが変更されるためコンサルタント個人の能力による解約へ

の影響は低いと考えられる．以上の理由から本章では契約期間が五年以上の企業を対象データとして扱う．

解約発生企業は解約時から過去二年分，継続企業は現在から過去二年分のテキストデータを用いて解析する．これは，二年以上遡るとテキストデータに解約の兆候は表れていないと考えられるためである．

対象となるテキストデータを以下に示す．企業数は実規模データから一部を抽出した，解約グループ119社，継続グループ119社である．相談内容件数，文章量，総単語数，異なり語数のすべてにおいて，解約グループより継続グループが多いことから，継続グループの企業の方が活発にコミュニケーションをとれていることが確認できる．

テキストデータ

表 3.1: Experiment Condition

| | 解約グループ：$G_1$ | 継続グループ：$G_2$ |
|---|---|---|
| 企業数 | 119($e_1$〜$e_{20}$) | 119($n_1$〜$n_{20}$) |
| 段落(相談内容件数) | :50,768 | 74,291 |
| 文 | 65,230 | 89,481 |
| 総単語数 | 778,056 | 1,040,800 |
| 異なり語数 | 16,142 | 18,180 |

実験環境

- OS:Windows 8.1
- CPU:Intel(R) Core(TM) i5-4460S CPU @ 2.90GHz
- メモリ:8.0GB

## 3.3 要因となる語の抽出

本節では対応分析を用いて要因となる語の抽出を行う．

### 3.3.1 抽出語句

各グループにおいて要因が強い語句として抽出された上位１０語を表3.2，表3.3に示す．表のWord列が抽出された語句で，Strengthが要因の強さを表している．解約グループにおいて，“アプローチ”“巻き返し”などの解約問題に関係する語句が抽出されている．また，一見解約と関係のない“日報”が解約グループにおいて抽出されている．

表 3.2: Cancellation

| Word | Strength |
|---|---|
| 望む | 3.172 |
| 距離 | 3.142 |
| 日報 | 3.100 |
| 巻き返し | 2.917 |
| アプローチ | 2.760 |
| 悪化 | 2.692 |
| 当たり | 2.638 |
| 応援 | 2.534 |
| 早める | 2.320 |
| 横ばい | 2.307 |

表 3.3: Continuation

| Word | Strength |
|---|---|
| 年金 | 3.491 |
| 無理 | 2.962 |
| デメリット | 2.687 |
| 最悪 | 2.504 |
| 記事 | 2.475 |
| 性質 | 2.447 |
| 別途 | 2.398 |
| 倉庫 | 2.198 |
| 器具 | 2.155 |
| 委任 | 2.105 |

“日報”は従業員を管理するために行われるものである．相談内容に“日報”という単語が含まれるということは従業員の管理が上手くいっておらず，サービスに満足していない又は変更しなければならない状態にあると推測されることを，コンサルタントへのヒアリングによって明らかになった．このように，提案手法によって抽出された語句はコンサルタントの経験や勘からは要因として選ばれない語句が含まれており，コンサルタントと抽出された語句についてディスカッションすることにより新たな知見を得るきっかけとなることが期待できる．

### 3.3.2 外れ値の考慮

要因となる語は各グループの差によって抽出する．よって片方のグループで特定の企業のみで極端に使われている語句（対応分析の原点から極端に離れている語句）は，本来要因の強い語句でない（対応分析の原点付近に存在しない）が抽出されてしまうという問題がある．そこで，Smirnov-Grubbs 検定をおこない外れ値となる語句を実験対象から削除する．本実験で外れ値として削除された語句は解約グループのみであった．解約グループにおける対応分析の結果（原点からの距離）を表 3.4 に示す．外れ値として抽出された語句は“肩書”，“次回”，“設計”，“恐縮”，“集合”，“引去る”，“入札”，“ゴミ”，“参入”，“寝る”である．解約グループの中で原点からの距離が最も遠い 10 語が抽出されていた．さらに，解約グループにおける語句全体の原点からの距離の分布を図 3.1 に示す．図の縦軸は原点からの距離を示しており，横軸は語句を表している．赤点が外れ値として抽出された語句の値である．原点からの距離が大きく，特定の企業でのみ多く出現している単語が外れ値と判定されていることが分かる．また外れ値は他の語句から飛び値となっていることが分かる．

次に，継続グループで抽出された要因となる語について外れ値を考慮する前後の比較を表 3.5，表 3.6 に示す．解約グループの要因として抽出された語の上位 10 語を表し

表 3.4: Outliers

| Words | Strength |
|---|---|
| 電話 | 0.027 |
| 対応 | 0.028 |
| 今後 | 0.032 |
| 連絡 | 0.033 |
| 依頼 | 0.033 |
| 確認 | 0.034 |
| 不明 | 0.036 |
| 関係 | 0.037 |
| 下記 | 0.038 |
| 必要 | 0.038 |
| ⋮ | ⋮ |
| 月数 | 0.620 |
| 新社屋 | 0.637 |
| 右下 | 0.637 |
| 再開 | 0.637 |
| 遭う | 0.637 |
| 肩書 | 0.775 |
| 次回 | 0.775 |
| 設計 | 0.775 |
| 恐縮 | 0.775 |
| 集合 | 0.775 |
| 引去る | 0.775 |
| 入札 | 0.798 |
| ゴミ | 0.798 |
| 参入 | 0.798 |
| 寝る | 0.798 |

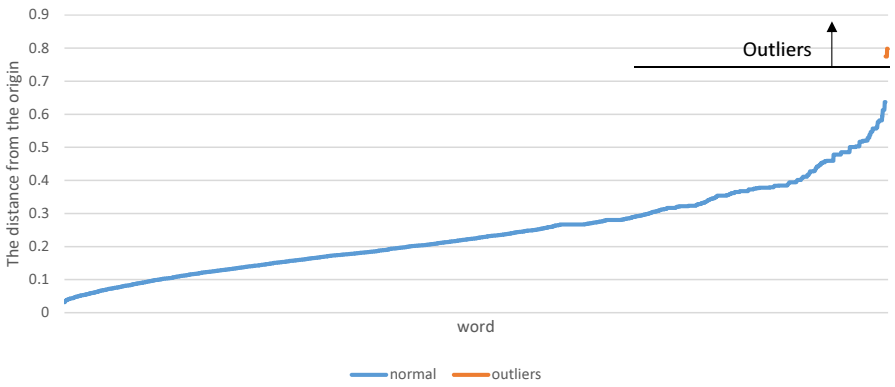



図 3.1: The direction from the origin (Cancellation group)

ている．外れ値を考慮しない場合は，外れ値として削除された“恐縮”,“設計”が抽出されている．外れ値を削除した場合は新たに“器具”,“委任”が追加されている．

表 3.5: Before deleting outliers

| Word | Strength |
|---|---|
| 恐縮 | 7.612 |
| 設計 | 5.339 |
| 年金 | 3.491 |
| 無理 | 2.962 |
| デメリット | 2.687 |
| 最悪 | 2.504 |
| 記事 | 2.475 |
| 性質 | 2.447 |
| 別途 | 2.398 |
| 倉庫 | 2.198 |

表 3.6: After deleting outliers

| Word | Strength |
|---|---|
| 年金 | 3.491 |
| 無理 | 2.962 |
| デメリット | 2.687 |
| 最悪 | 2.504 |
| 記事 | 2.475 |
| 性質 | 2.447 |
| 別途 | 2.398 |
| 倉庫 | 2.198 |
| 器具 | 2.155 |
| 委任 | 2.105 |

また，抽出された語句(上位10語)における対応分析の原点からの距離の総和を表3.7に示す．継続グループ内に外れ値がなかったため，解約グループの要因として抽出される語句に変化はなかった．一方継続グループの要因として抽出された語句の原点からの総和は，外れ値を考慮しないときに4.25であったの対して，外れ値を考慮したときの総和は3.87であった．外れ値を考慮することにより，原点からの距離の総和が小さくなっていることを確認した．このことから相手グループの外れ値の影響を受けることなく，よりグループの要因となる語句が抽出されていることが分かる．

表 3.7: Average of the direction from the origin

| | Cancellation | Continuation |
|---|---|---|
| No Outliers | 4.25 | 5.02 |
| Outliers | 4.25 | 3.18 |

## 3.4 DEA判別分析

3.3節で求めた語を変数に用いてDEA判別分析を行い，判別式を作成する．また，新たなデータに対して判別することにより，作成した判別式の有効性を検証する．本研究では機械学習法は用いていないが，説明のため判別式作成に用いたデータを学習データ，判別式を検証するために用いたデータを予測データとよぶ．

### 3.4.1 抽出語数変化と判別率

#### 実験条件

3.3節で抽出する語数を変化させて，判別式を作成する．実験条件を以下に示す．判別式作成用データは解約グループ継続グループともに60社ずつ，予測用データは解約グループ継続グループともに59社である．抽出語数を1〜250語に変化させて，抽出語数の変化と判別率の関係を検討する．

- 判別式を作成するために用いた企業（学習データ）
  - 企業数：120（解約グループ：60/継続グループ：60）
  - オーバーラップの幅：0.5
- 判別式を検証するために用いた企業（予測データ）
  - 企業数：118（解約グループ：59/継続グループ：59）
- 抽出語数：1〜250語

#### 実験結果

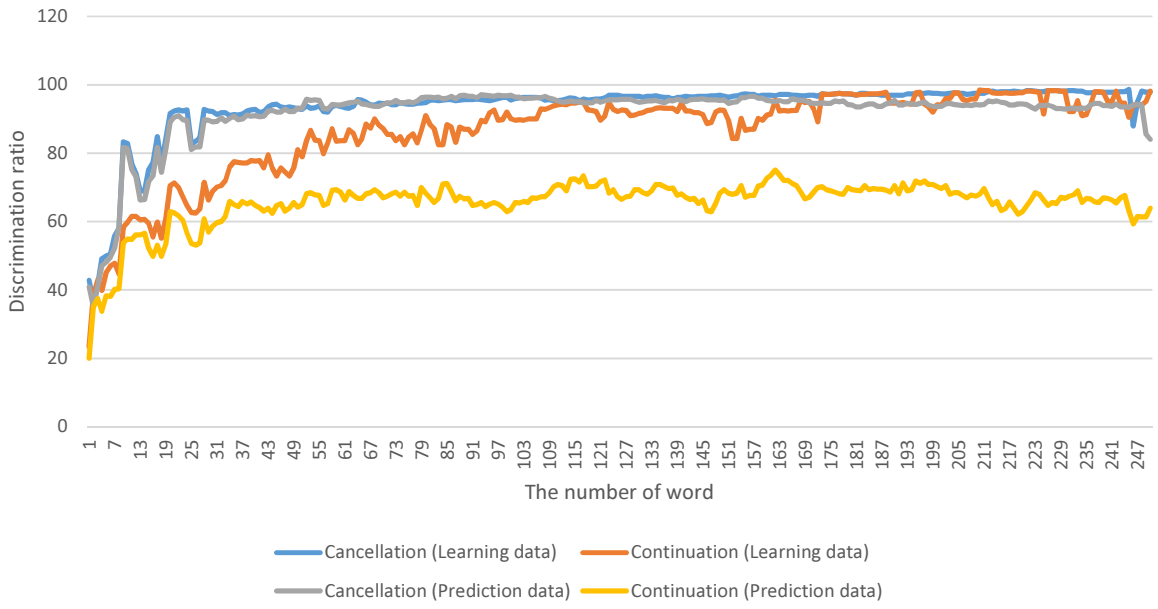



図 3.2: The number of extract word and discrimination rate

実験結果を図3.2に示す．図の縦軸は判別率を，横軸は抽出語数を表している．青線，赤線が判別式作成に用いたデータで，灰線，黄線が判別式を検証するために用いた新

たなデータの判別率を表している．学習データにおいて抽出語数20語で両グループともに判別率が90% をこえ，さらに抽出語数が多くなるにつれて100% に収束する．予測データにおいて，解約グループは学習データと同様に55語で80% を超えてそれ以上多い単語数で100% に近づく判別率である．一方継続グループは37語までは上昇しているものの，それ以上の抽出語数では50〜90% を振動している．解約グループの方が継続グループよりも判別率が高いことから，解約グループのテキストデータには解約を特徴付ける語句が含まれていることが確認できる．また両グループともに，ある一定以上判別率が高くなると抽出語数を増やしても判別率は振動する．よって，語句数が少なすぎると判別に必用な係数が足りない一方で，語句数は一定以上多くなっても対象問題の判別率は上昇しないことがわかった．

### 3.4.2 企業数変化と判別率

テキストデータの量と判別率の関係を調べるために，判別式作成に用いるテキストデータの企業数を変化させて判別式を作成する．実験条件を以下に示す.

- 判別式を作成するために用いた企業（学習データ）
  - 企業数：$g_1$ ($g_1$ = 20,40,60,80,100,120,140,160,180,200)
  - オーバーラップの幅：0.5
- 判別式を検証するために用いた企業（予測データ）
  - 企業数：$g_2 = 238 - g_1$ ($g_2$ = 218,198,178,158,138,118,98,78,58,38)
- 抽出語数：50語
- 繰り返し回数：10回

各グループにおいて，テキストデータを判別式作成に用いる企業数を10社ずつ増やして判別式を作成した．抽出語数は3.4.1小節の結果より50語に設定した．実験結果を図3.3，表3.8に示す.

図の縦軸は判別率，横軸は判別式を作成するために用いる企業の数を表している．学習データにおいて解約・継続グループともに常に判別率が90% 以上であるが，企業数が増えるにつれて緩やかに判別率が減少している．予測データにおいて，判別式作成に用いる企業数が増えるにつれて判別率が上昇している．これらのことから判別式作成のために用いる企業数が増えるにつれて，学習データにおける判別は難しくなるが，汎用的な判別式を作成することができ予測の精度は高まるといえる．

また，予測データにおいて3.4.1小節と同様に，解約グループに比べて継続グループ

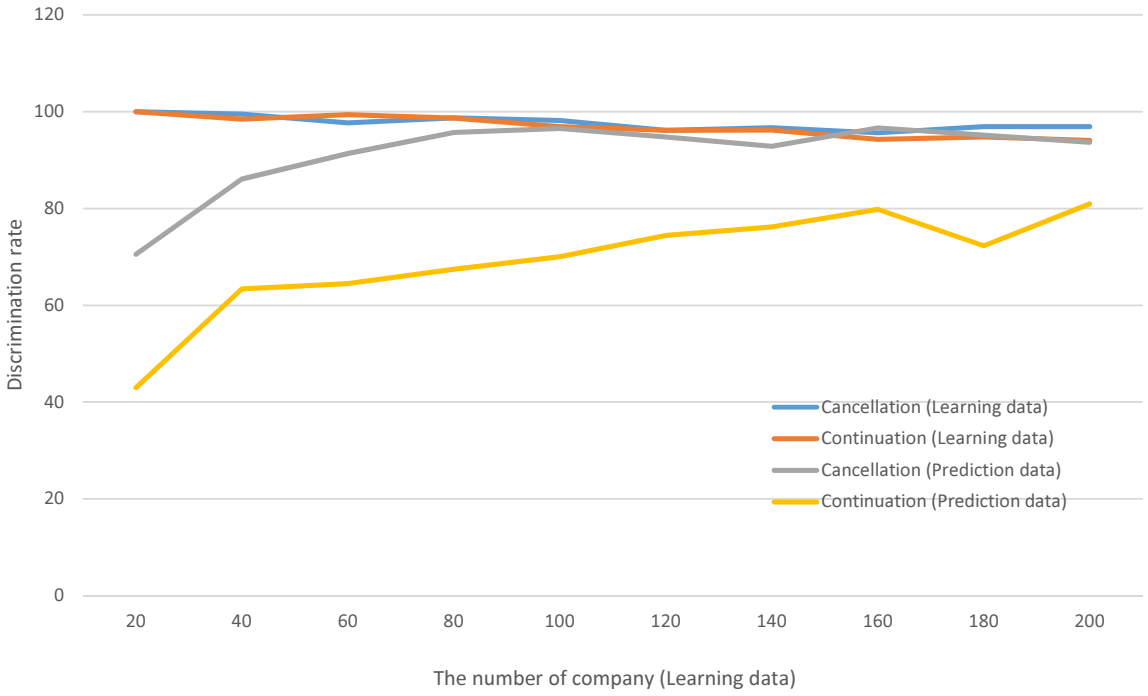



図 3.3: The number of company and discrimination rate

表 3.8: The number of company and discrimination rate

| $g_1/g_2$ | 20/218 | 40/198 | 60/178 | 80/158 | 100/138 |
|---|---|---|---|---|---|
| Cancellation (Learning data) | 100 | 99.5 | 97.66 | 98.75 | 98.2 |
| Continuation (Learning data) | 100 | 98.473 | 99.375 | 98.673 | 96.928 |
| Cancellation (Prediction data) | 70.528 | 86.060 | 91.348 | 95.696 | 96.521 |
| Continuation (Prediction data) | 42.975 | 63.374 | 64.499 | 67.460 | 70.048 |
| $g_1/g_2$ | 120/118 | 140/98 | 160/78 | 180/58 | 200/38 |
| Cancellation (Learning data) | 96.166 | 96.714 | 95.625 | 96.888 | 96.9 |
| Continuation (Learning data) | 96.124 | 96.201 | 94.292 | 94.778 | 94.078 |
| Cancellation (Prediction data) | 94.745 | 92.857 | 96.666 | 95.172 | 93.684 |
| Continuation (Prediction data) | 74.436 | 76.176 | 79.826 | 72.271 | 80.943 |

の判別率が低い．継続グループの判別率の低さの要因を検討するため，各グループのテキストデータの語句の数について解析した．ここで語彙の豊富さの指標としてタイプ･トークン比を用いた[47]．総語句数を$N$，異なり語句数を$V$とする時，タイプ･トークン比は式3.1で表される。タイプ･トークン比の値が高いほど，その文書は語彙が豊富であるといえる．

$$TTR = \frac{V}{N} \tag{3.1}$$

各グループのテキストデータの総語句数，異なり語句数，タイプ･トークン比，各企業の文書量の分散値を表3.9に示す．左列が解約グループ，右列が継続グループにおける全ての文書についての結果を表している．

表 3.9: Comparison of document quantity of cancellation group and continuation group

| | Cancellation | Continuation |
|---|---|---|
| Total word count | 778,056 | 1,040,800 |
| Number of words and phrases | 16,142 | 18,180 |
| Token ratio | 0.0207 | 0.0174 |
| Distribution of document volume of each company | 164.935 | 13978.51 |

総語句数，異なり語句数ともに解約グループよりも継続グループの方が多く，継続グループの方が文書量は多いとわかる．一方，タイプ･トークン比は解約グループの方が値が大きく語句が豊富であるとわかる．語彙の種類が多いとしても判別率は下がらず，むしろ豊富な語句の種類を含むテキストデータを用いて判別式を作成することにより，汎用的な判別式が作成できたと考えられる．ここで，各企業の文書量の分散値を比較すると，継続グループが12978.51解約グループが164.935となり，継続グループが解約グループに比べて高い値となっている．このことから継続グループにおいて企業ごとにテキストデータ量が偏っており，特定の企業に特定の言葉が多く使用されていることが分かる．

図3.4に各企業ごとの文書量を示す．縦軸が文書量で，横軸が文書量の昇順に並べた企業番号を表している．青点が解約グループであり，赤点が継続グループである．企業番号80までは両グループともに同じ分布をしているが，企業番号90以降になると継続グループは解約グループに比べて文書量の増加率が高くなっている．このことからも，継続グループでは解約グループより企業ごとに文書量に偏りがあることがわかる．この文書量の偏りが継続グループの判別率低下に影響していると考えられる．

## 3.5 比較実験

提案手法の有効性を確認するため，既存手法と比較し検討した．

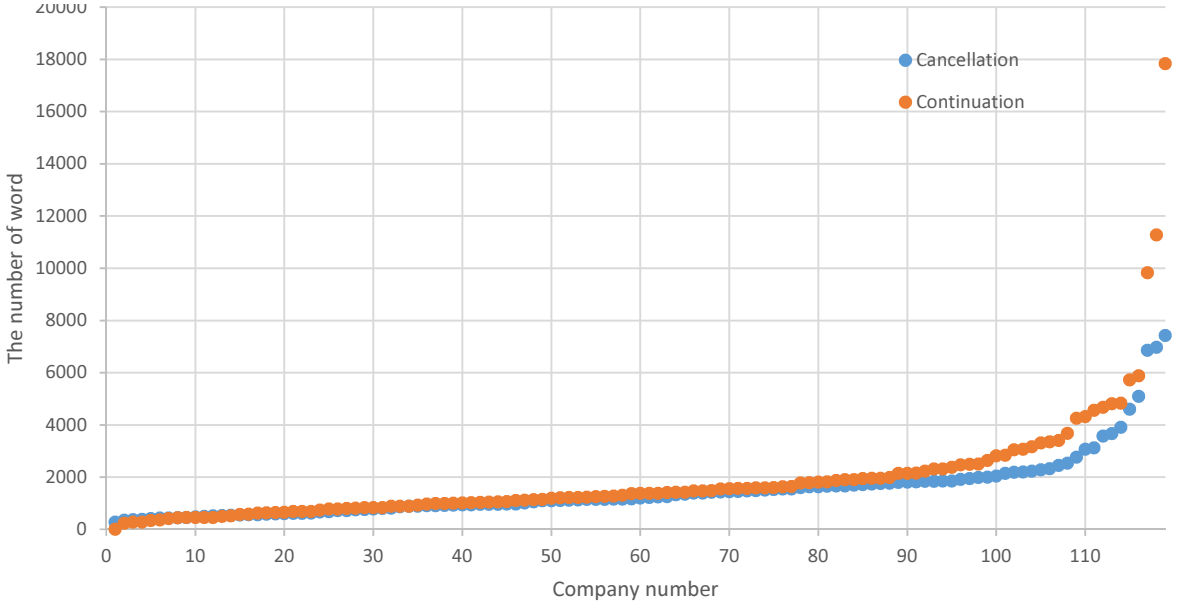



図 3.4: The number of word and company number

### 3.5.1 対応分析の有効性

提案手法では，DEA 判別分析で用いる変数を削減するために，対応分析を用いて各グループ内でテキストデータ全体に相関の強い語句を抽出し，グループ間の差を考慮することにより要因となる語の抽出をおこなっている．提案手法の有効性を確認するため，語句の重み付けで一般的な手法である $TF-IDF$ 法を用いた語句抽出方法と，抽出される語句と判別率について比較した．以下に $TF-IDF$ 法を式 (3.3) に示す．$tf$ 値はそれぞれの語の文書内での出現頻度を表すため，値が高いほど要因が強い．一方 $idf$ 値はそれぞれの語がいくつの文章内で共通して使われているかを表すため，値が高いと重要でなくなる．この二値を掛けたものが語の重みとなる．比較手法では，$TF-IDF$ 手法で重みの強いものから順に抽出する．

**記号定義**

$tf(t,d)$ ： 文章 $d$ 内のある単語 $t$ の $tf$ 値

$n_{t,d}$ ： ある単語 $t$ の文章 $d$ 内での出現回数

$\Sigma_{s\in d} n_{s,d}$ ： 文章 $d$ 内のすべての単語の出現回数の和

$$tf(t,d) = \frac{n_{t,d}}{\Sigma_{s\in d} n_{s,d}} \tag{3.2}$$

$$idf(t) = \log \frac{N}{df(t)} + 1 \tag{3.3}$$

$TF-IDF$ 法によって抽出された要因の強い上位10語を，解約グループを表3.10，継続グループを表3.11に示す．

表 3.10: Cancellation

| Word | tf-idf |
|---|---|
| 収支 | 0.986 |
| 流通 | 0.980 |
| 現地 | 0.977 |
| ビザ | 0.971 |
| リンク | 0.971 |
| 頻度 | 0.967 |
| 多用 | 0.966 |
| 個別 | 0.962 |
| 止め | 0.95 |
| モニタリング | 0.9480 |

表 3.11: Continuation

| Word | tf-idf |
|---|---|
| シール | 0.990 |
| 係長 | 0.978 |
| 恐縮 | 0.977 |
| 印紙 | 0.972 |
| 半日 | 0.969 |
| 支社 | 0.966 |
| 地位 | 0.958 |
| 税額 | 0.956 |
| 遡る | 0.954 |
| 経つ | 0.95 |

提案手法で抽出された上位10語の語句（表4.25，表4.26）と比較すると，全て異なる語が抽出されていた．また，比較手法の解約グループにおいて抽出された語の中には，一見して解約と結びつく語句は抽出されていなかった．

ここで，解約問題で最も要因の強い語として抽出された“望む”（提案手法），“収支”（比較手法）に着目して考察する．両語句の総出現回数，企業ごとの出現回数の分散値を表3.12に示す．さらに，各企業での出現回数の分布を図3.5に示す．図の縦軸は語の出現回数，横軸は企業番号を表している．

表 3.12: Comparison of the number of word

| | Total word count | Distribution |
|---|---|---|
| “望む” (Proposed method) | 62 | 0.319 |
| “収支” (Compared method) | 79 | 41.174 |

表3.12より，両語句の総出現回数は比較手法の方が多いが，提案手法が62で比較手法が79と差は小さい．一方，各企業における出現回数の分散値も比較手法の方が高いが，提案手法が0.319で比較手法が41.174であるため差が大きい．ここで，図3.5より出現回数の分布を解析すると，比較手法では1社で70回出現し，企業によって出現回数に偏りのある語句が抽出されていることがわかった．一方，提案手法で抽出された語句は，特定の企業でのみ多く出現しているのではなく，多数の企業で広く出現している語句であることがわかった．このことから，提案手法によって抽出された語句の方がグループ内の企業で広く使われており，グループの特徴を表す語句が抽出できているといえる．

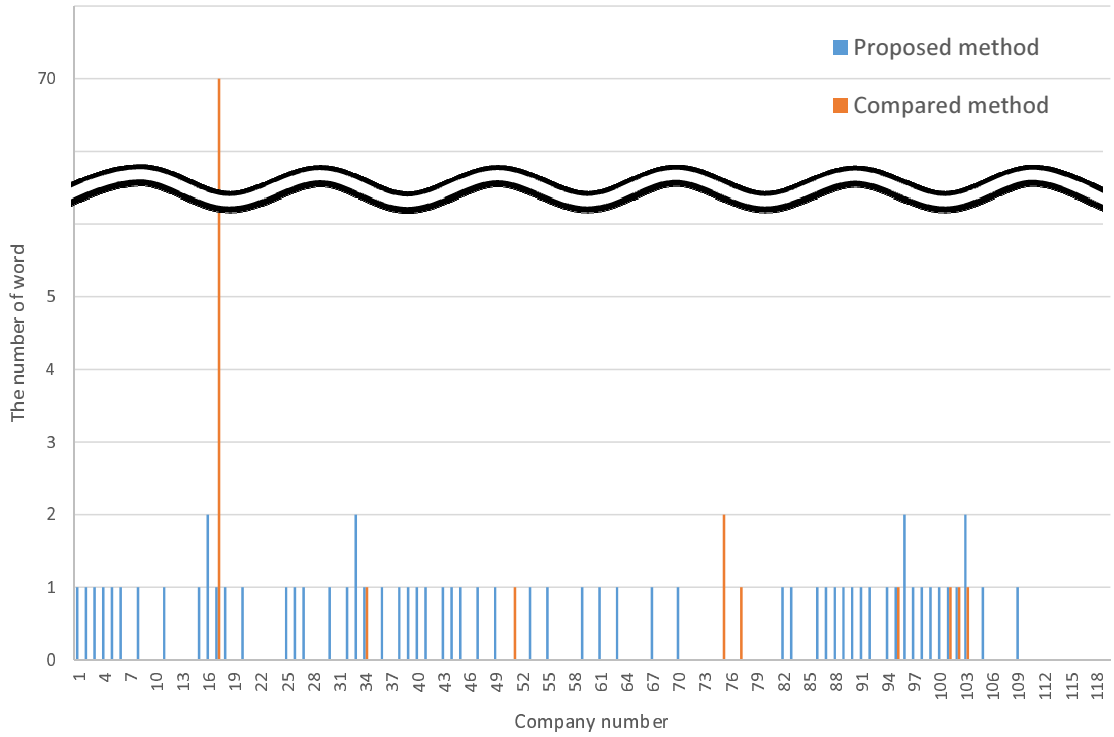



図 3.5: The number of word and company number

各手法によって抽出された語句を用いて判別率を比較する．抽出語数を1〜100語に変化させたときの判別率を図3.6，図3.7に示す．図3.6は学習データについての判別率，図3.7は予測データについての判別率を表している．図の縦軸は判別率，横軸は判別分析の変数に用いる抽出語数である．

学習データにおいて解約グループについては，提案手法・比較手法ともに抽出語数が一定数以上になると100% に収束している．継続グループについて，提案手法と比較手法に判別率の有意差があった．提案手法で語数が増えるにつれて判別率は100% に近づいているが，比較手法では抽出語数を増やしても判別率が振動している．また，予測データにおいても継続グループの判別率について有意差があった．比較手法では上記で確認したように，抽出語数について企業ごとの出現回数に偏りがあるので判別の精度が下がっていると考えられる．継続グループは企業ごとに文書量の差が大きいため(図3.4)，解約グループよりも判別率に影響を受けたと考えられる．

### 3.5.2 DEA判別分析の有効性

提案手法では，誤判別リスクを最小化できるDEA判別分析を用いている．有効性を検証するため，統計判別分析の判別率と比較する．本研究では，判別式の結果から問

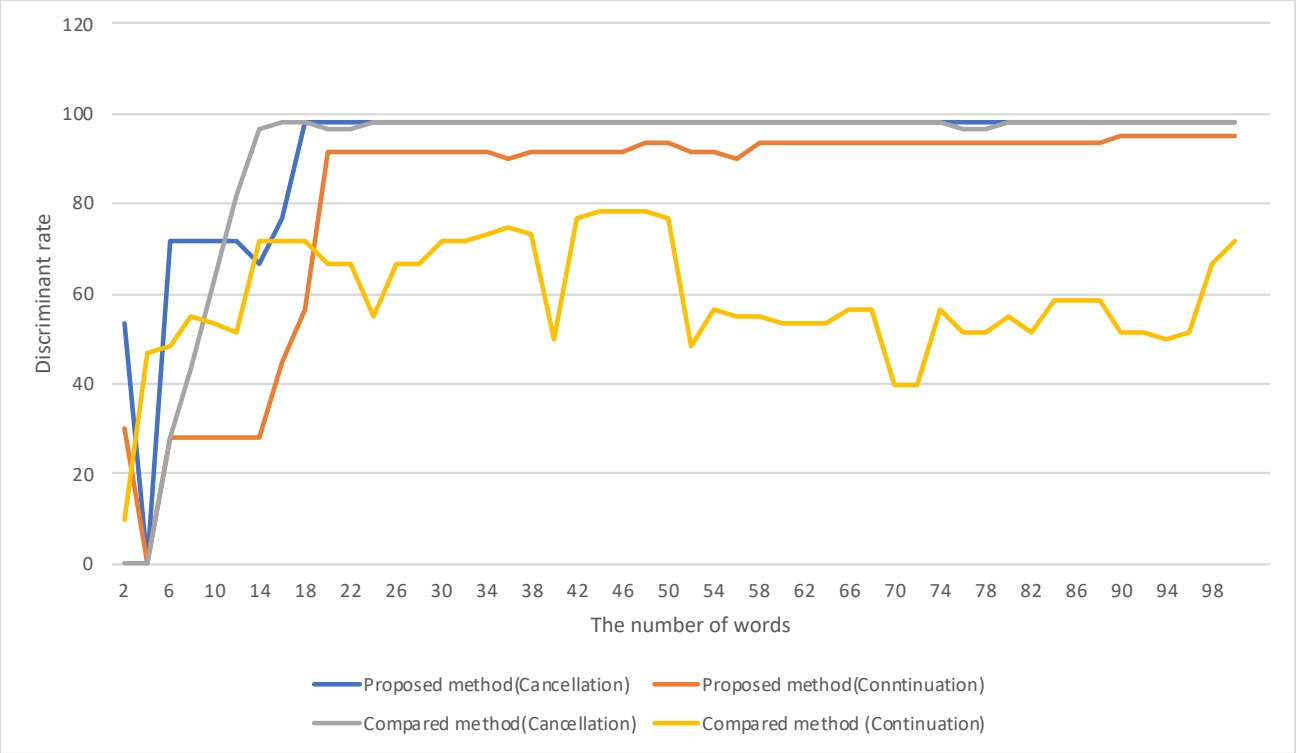



図 3.6: Learning data

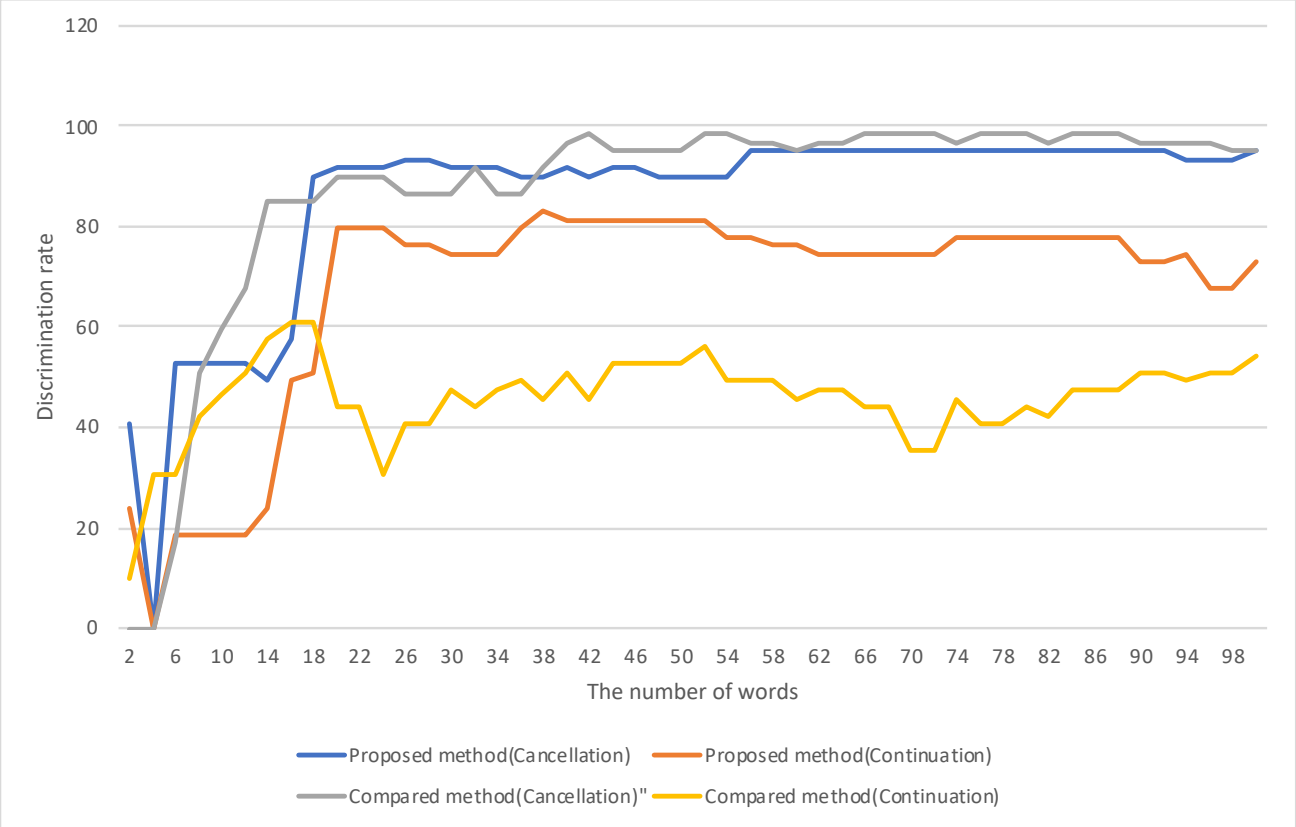



図 3.7: Prediction data

題の発生要因となる語句について考察を可能とするために，線形な判別式を作成することを目的としている．よって統計的判別手法の中で，線形判別分析を比較手法とする．各手法において実験した結果を表3.13に示す．表は試行回数の平均値を表している。学習データでは解約グループにおいて既存手法100%，提案手法98.3%であり，予測データでは解約グループにおいて既存手法89.85%，提案手法68.50%，継続グループにおいて既存手法72.50%，提案手法54.50%であり，学習データ予測データともに提案手法のほうが判別率が高く提案手法の有効性が確認できた．

提案手法のほうが判別率は高く，提案手法の有効性が確認できた．

表 3.13: Comparison of document quantity of cancellation group and continuation group

| | Cancellation | Continuation |
|---|---|---|
| Proposed method (Learning data) | 100 | 98.3 |
| Compared method (Learning data) | 98.3 | 98.3 |
| Proposed method (Prediction data) | 89.85 | 72.50 |
| Compared method (Prediction method) | 68.50 | 54.50 |

## 3.6 考察

各計算機実験の結果から，以下の特性が分かった．

- 要因となる語の抽出
  抽出された語句の中には，コンサルタントが見て対象問題と関係があると分かる語が含まれていおり，対象グループの要因となる語が実際に抽出されていることを確認した．さらにコンサルタントが気づくことのできなかった語句についても，計算機実験から得られた抽出語をもとに議論することによって，新たな関連性を推測することできた．

- DEA判別分析による解約予測
  抽出語数の変化と判別率の関係を検討するための計算機実験により，判別に用いる語句数は少なすぎると判別できないが，増やしすぎてもノイズとなる語句が抽出され判別率が上昇しないことがわかった．また，企業数の変化と判別率について検討すると，判別式作成のために用いる企業数が多いほど汎用的な判別式が作成できていたことがわかった．また，判別率は解約グループに比べ，継続グループの方が判別率が低いことが分かった．これは，各企業の文書量の偏りに起因すると考えられる．

- 提案手法の有効性
  要因となる語の抽出において$TF-IDF$法と比較することにより，提案手法では比較手法に比べてグループ内の企業全体で広く使用されている語句が抽出されていることを確認した．抽出した語を用いて判別率を比較した結果，継続グループにおいて学習データ・予測データともに提案手法の方が判別の精度が良いことがわかった．また統計的判別分析と提案手法を比較することにより，提案手法で用いているDEA判別分析の有効性を確認した．

## 3.7 結言

本章では，2章で述べた本研究の提案手法を基に計算機実験を行い，解約問題を対象に判別結果の考察をおこなった．具体的には，3.2節で実験条件を述べ，3.3節で提案手法によって抽出される語句の考察を行い，3.4節で提案手法によって判別式の作成，作成した判別式の有効性を検証した．3.5節では，提案手法の有効性を確認するために，既存手法と比較を行った．

次章では解約問題を対象とし，実現場適用に向けて実規模問題を解析する．

# 第4章　実規模問題を対象とした解約予測の支援手法

## 4.1　緒言

本章では解約問題を対象とし，実現場適用に向けて実規模問題を解析し，拡張手法の有効性を検証する．4.2節では実規模問題の特徴と拡張した支援手法について説明する．4.3節では対象となる実規模問題について述べ，4.4節で対象とするデータ範囲を選定するための解析をおこなう．4.5節では，実現場適用に向けて拡張した各手法の説明と有効性の検証について述べる．4.6節では，拡張された手法を用いて実規模問題を解析し，解約問題の予測をおこなう．4.7節では，コンサルタントの解約予測結果と比較検証することにより，提案手法によるコンサルタント支援の可能性を検証する．4.8節では実験結果についての考察を述べる．最後に，4.9節で本章についてまとめる．

## 4.2　実現場適用に向けた支援手法

本章では解約問題を対象とし，実現場適用に向けて実規模問題を解析し，拡張手法の有効性を検証する．

### 4.2.1　実規模問題の特徴

実規模問題を対象とする際，以下の特徴が存在する．

1. 解約問題が発生する企業は全企業の中でも一部であるため，問題発生ありグループとなしグループの企業数に差がある．提案手法では特徴的な語句を抽出する際，両グループに出現する語句を対象として特徴抽出をおこなっているため，グループのサンプル数に差がある場合，企業数の多いグループの特徴を捉えることが困難である．

2. 対象企業数が多いため，要因として抽出された語句がテキストデータに一語も出現しない企業が存在し，判別不能な企業が発生する．

3. 問題が発生する企業は全企業の中でも一部であるため，問題発生ありグループとなしグループの企業数に差がある．企業数の差により，企業数の多い問題発生がないグループの判別に有利な判別式が作成されてしまう．

4. 判別式作成用データと予測用データのテキスト量に差がある．解約問題において，判別式作成用データは対象とする月以外の11カ月分の更新月を対象とする企業データであるのに対して，予測用データは対象月のみの企業群であるため，予測用データが判別式作成用データに比べて少ない．判別式に用いられている語句が1語も存在しない企業が，予測用データ内に存在し判別不能な企業が出現する．

それぞれの問題を解決するため以下の手法拡張をおこなった．

1. 片方のグループにのみ出現する語句も対象とすることにより，企業数の多いグループについて，より幅広い特徴を捉えることを可能とする．さらに，IDF値を用いることにより，グループ内でより汎用的な語句を抽出する．

2. テキストデータに要因となる語句が出現しないということも判別に利用されるべきであるが，単語の出現回数が0回の場合判別式作成に寄与しない．そこで標準化を用いて各企業の単語の出現回数を正規化することにより，出現回数が0回に負の値を与え,「ある単語が出現しない」ことも判別において考慮可能とする．

3. DEA判別分析の目的関数に重みを与えることにより，グループ間のサンプル数の差を考慮する．

4. 予測用データ内で判別式に用いられる語句を探索する際に，判別式に用いられる語句の類義語も考慮することにより，予測用データと判別式に用いられる語句の重なりを大きくする．

各手法を拡張した際の本研究の流れを図4.1に示す．特徴抽出，標準化，DEA判別分析の目的関数，類義語辞書を用いたデータ拡張のステップが拡張点である．拡張した各ステップについて詳細に説明する．

### 4.2.2 特徴抽出法

実現場においてコンサルタントが察知しなければいけない問題，つまり本研究が予測対象とする問題は，発生確率が低いまたは発覚しづらい問題である．そのため，問題発生ありと問題発生なしの各グループの企業数に差がある．グループの企業数に差がある場合，両グループに出現する語句のみを用いた特徴抽出をおこなうと企業数の多いグループの特徴を捉えることが困難である．片方のグループにのみ出現する語句も対象とすることにより，企業数の多いグループについて，より幅広い特徴を捉えることを可能とする．さらに，IDF値を用いることにより，グループ内でより汎用的な語句を抽出する手法の拡張をおこなう．
提案手法では両グループに出現する語句をグループ間の差を考慮することにより，要

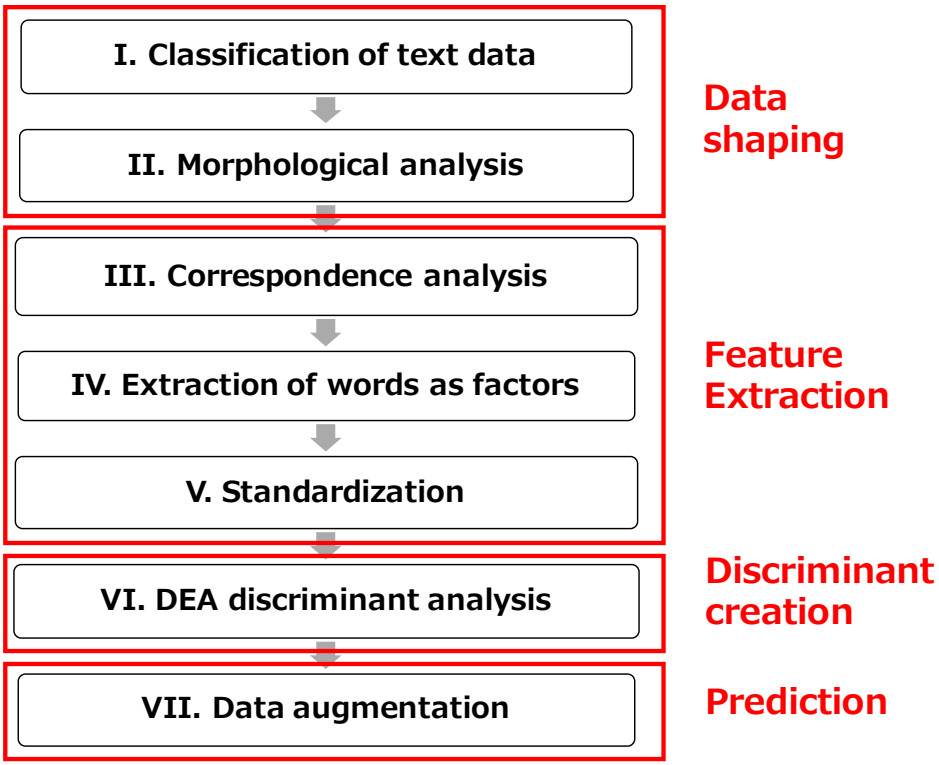



図 4.1: Flow of the proposed method

因となる語句を抽出した．片方のグループにのみ出現する語句についても考慮するため，以下のルールを用いて語句を抽出する．

1. 両グループに出現する語句の中から，最も要因の強い語句を選択する．
2. 選択された語句よりも原点からの距離の小さい片方のグループにのみ出現する語句を，原点から距離の小さい語句から順に抽出する．設定した抽出語数に達すると終了．
3. 1で選択した語句を抽出する．
4. 設定した抽出語数未満であれば1へ．抽出語数を満たすと終了．

### 4.2.3 標準化

本研究で対象とするテキストデータは多様なメディアによる非均質なデータである．そのため内容が幅広く使用される語句も様々である．問題予測において，テキストデータ内にある要因となる語句が出現しないということも判別に利用されるべきであるが，単語の出現回数が0回の場合判別式作成に寄与しない．そこで標準化を用いて各企業の単語の出現回数を正規化することにより，出現回数が0回に負の値を与え，「ある単語が出現しない」ことも判別において考慮可能とする．

統計学において標準化とは，与えられたデータを平均が0，分散が1のデータに変換する操作のことである．任意の正規分布に従うデータ$x$を標準正規分布に従うデータに変換するために用いられる場合が多い．データ$x$の各データを標準化して得られる標準化変数と呼ばれる値はそれぞれが標準正規分布に従う．標準化の式を式4.1に示す．ここで，$\mu$は平均，$\sigma$は標準偏差である．

$$z_i = \frac{x_i - \mu}{\sigma} \tag{4.1}$$

本研究では，標準化により語句の分布尺度を統一する．また出現回数が0回の語句は判別分析において判別に影響を与えないが，標準化により出現回数が0回の語句にも負の値が与えられることにより,「ある語句が出現しない」ことも判別分析において考慮可能となる．

### 4.2.4 判別分析の目的関数拡張

実現場においてコンサルタントが察知しなければいけない問題，つまり本研究が予測対象とする問題は，発生確率が低いまたは発覚困難な問題である．そのため，問題発生ありと問題発生なしの各グループの企業数に差がある．企業数の差により，企業数の多いグループに有利な判別とならないようにするため，DEA判別分析の目的関数を式4.2とし，各目的関数に重みづけをおこないグループ間の差を考慮する．ここで$N$は全企業数，$n_1$はグループ1の企業数，$n_2$はグループ2の企業数をあらわす．

$$\min \quad \frac{n_2}{N}\sum_{j\in G_1} S_{1j}^{+} + \frac{n_1}{N}\sum_{j\in G_2} S_{2j}^{-} \tag{4.2}$$

### 4.2.5 類義語辞書を用いたデータ拡張

判別式作成用データと予測用データのテキスト量に差がある．解約問題において，判別式作成用データは対象とする月以外の11カ月分の更新月を対象とする企業データであるのに対して，予測用データは対象月のみの企業群であるため，予測用データが判別式作成用データに比べて少ない．判別式に用いられている語句が1語も存在しない企業が，予測用データ内に存在し判別不能な企業が出現する．予測用データ内で判別式に用いられる語句を探索する際に，判別式に用いられる語句の表記ぶれを考慮することにより，予測用データと判別式に用いられる語句の重なりを大きくする．

テキストデータでは同じ意味を表していても，異なる語句が用いられる場合がある．そのような表記ぶれを考慮するため，類義語を用いたテキストデータの拡張を行う．類義語辞書には日本語WordNetを用いる．日本語WordNetはプリンストン大学で開発

されたPrinceton WordNetとヨーロッパのEuroWordNet協会が推進するGlobal WordNet Gridに着想を得て開発された[48]．WordNetとは語を類義関係のセット(synset)でグループ化している点に特徴があり，一つのsynsetが一つの概念に対応している．また，各synsetは上位下位関係などの多様な関係で結ばれいている．世界的にも各国言語のワードネットが作成され，言語処理研究や言語学研究などに広く利用されている．日本語WordNetでは，Princeton WordNetのsynsetに対応して日本語が付与されている．日本語WordNetに収録されているsynset数，単語数，語義数を表4.1に示す．日本語WordNetを用いることにより，類義語を考慮したデータ拡張をおこなう．

表 4.1: Japanese WordNet

| | |
|---|---|
| Synset | 57,238 |
| Words | 93,834 |
| The pair of synset and words | 158,058 |
| Definition statement | 135,692 |
| Example sentence | 48,276 |

本節では，データ拡張の必要性と拡張方法について述べる．判別式に用いられた語句が予測対象データにおいて1語も出現しない場合，予測が不能となってしまう．本研究の対象とするテキストデータは前述の通り，多様なメディアにテキストであり非均質であり，同様の意味の語句にも表記のばらつきがある．そこで，予測対象データにおいて類義語も考慮しデータ拡張をおこなうことにより，判別式で用いられる語句との重なりを大きくする．データ拡張と各データの関係を図4.2に示す．

類義語辞書を用いたデータ拡張方法について，図4.3に例を示す．例では「月」という単語についての，概念，上位語，下位語，類義語を表している.「月」には複数の概念が存在し，それぞれの概念に単語，上位語，下位語が紐づけられている．同じ概念に紐づけられた語句を類義語とみなす．本研究では，意味解析をおこなっていないため，各語句がテキスト内でどの概念で用いられているか判別できない．そのため各単語のすべての概念において紐づけられている語句を類義語であるとみなしデータ拡張をおこなう．

## 4.3 対象問題

前章と同様に，クライアント企業がコンサルティングサービスを継続するか解約するかを予測する，解約問題を対象とする．本章では実現場適用に向けて，ある月にコ

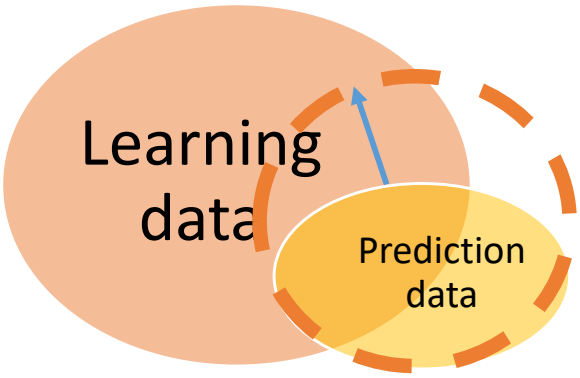



図 4.2: Data augmentation

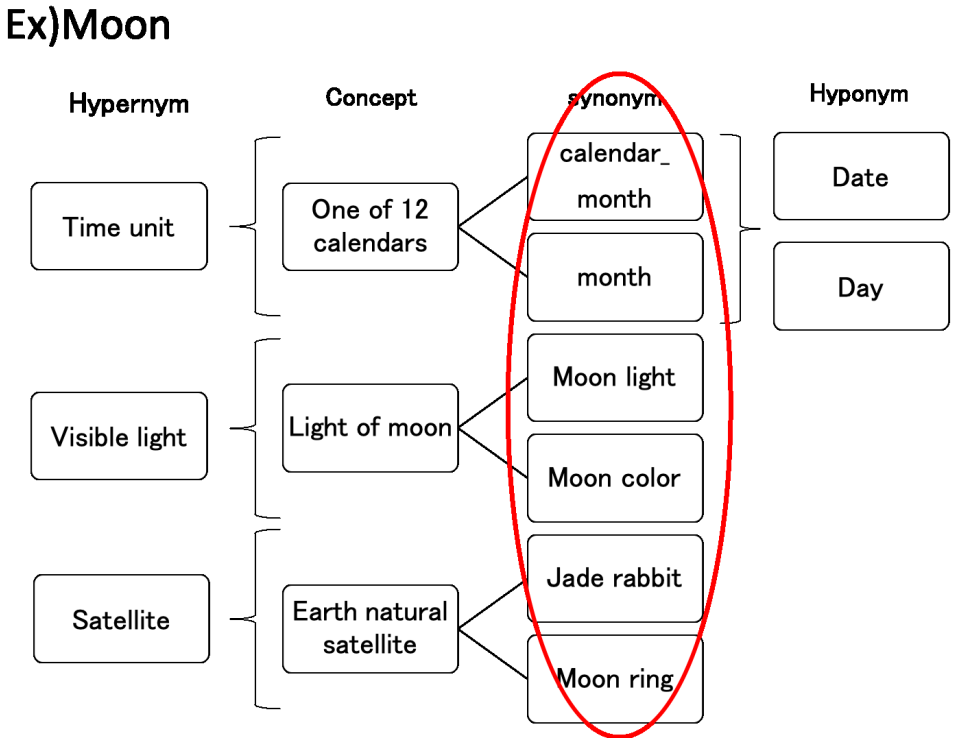



図 4.3: Example of WordNet

ンサルティングサービス契約の更新月を迎える企業を予測対象とし，その他の月にコンサルティングサービス契約の更新月を迎える企業のデータを用いて判別式作成をおこなう．対象データについて以下にまとめる．

実験環境について，以下にまとめる．本章では実規模問題を対象とし，解約グループ315社，継続グループ4048社である．

表 4.2: Experiment Condition

| | 解約グループ：$G_1$ | 継続グループ：$G_2$ |
|---|---|---|
| 企業数 | 315 | 4048 |
| 段落(相談内容件数) | 252,031 | 1,590,739 |
| 文 | 307,627 | 5,910,520 |
| 総単語数 | 745,542 | 4,697,331 |
| 異なり語数 | 11,441 | 59,501 |

**実験環境**

- OS:Windows 8.1
- CPU:Intel(R) Core(TM) i5-4460S CPU @ 2.90GHz
- メモリ:8.0GB

## 4.4 テキストデータ範囲の選定

判別式作成のための作成用データと予測用データに用いるテキストデータは図4.4, 4.5のデータ範囲を対象とする．作成用データは対象とする月以外の月に更新月を迎えるクライアント企業を対象とし，更新月までの直近3カ月のデータを除いたテキストを対象としている．予測用データも同様に対象とする月に更新を迎えるクライアント企業のテキストデータの中で，更新月までの直近3カ月のデータを除いたテキストを対象としている．これは，更新月に近いテキストデータにはコンサルティングサービスの継続や解約に関する直接的な内容が含まれ，削除する必要があるためである．ここで，図中で*pastyear*とされている変数について検証する．テキストデータが少なすぎると判別式作成が困難であり，テキストデータが多すぎると計算時間が膨大になる，また過去のテキストデータまで遡りすぎると対象問題と関連の薄い内容が含まれノイズとなってしまう危険性がある．そのため，過去何カ月分のデータを対象とするかは検証の必要がある．

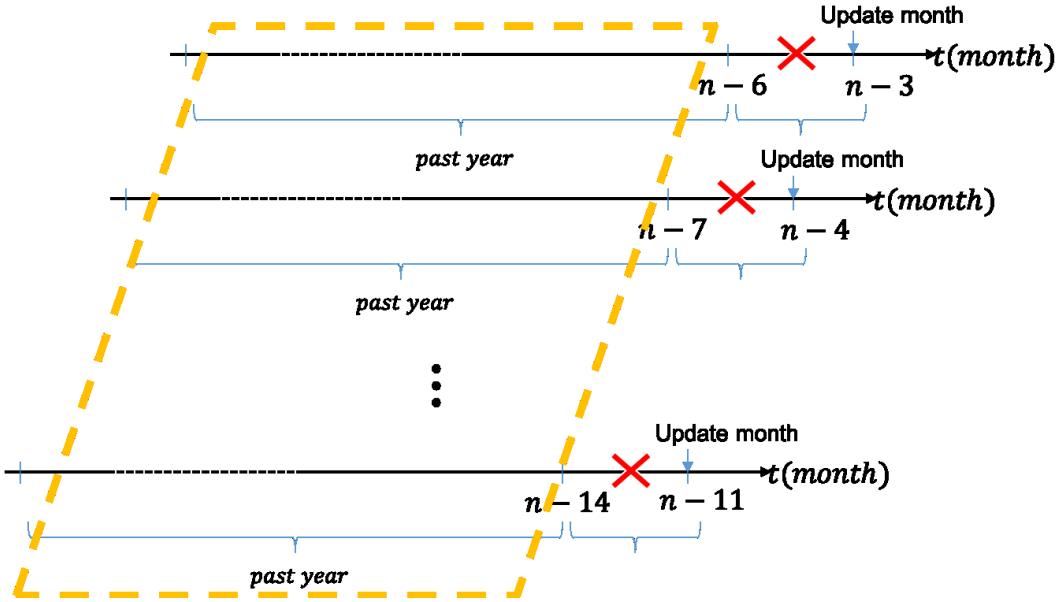



図 4.4: Learning data

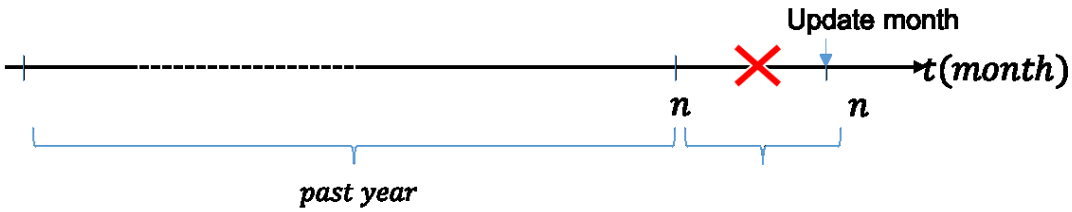



図 4.5: Prediction data

### 4.4.1 実験条件

過去何カ月の履歴データを対象とするべきか検証するために以下の3条件について，予測結果の比較を行い検証する．対象月から何カ月前までを対象のテキストデータとするかで条件を変更する．ただし全条件において対象月直近の3カ月のテキストデータは前述の通り削除する．

- CaseA : 6カ月
- CaseB : 12カ月
- CaseC : 24カ月

表4.3に各条件における企業数を示す.

表 4.3: Experiment Condition of each case

| | Learning data | | Prediction data | |
|---|---|---|---|---|
| | Cancellation | Continuation | Cancellation | Continuation |
| Case A | 554 | 2923 | 71 | 372 |
| Case B | 877 | 4511 | 79 | 418 |
| Case C | 878 | 4514 | 82 | 425 |

### 4.4.2 実験結果

Case Aの予測結果を表4.4，Case Bの予測結果を表4.5，Case Cの予測結果を表4.6に示す.

表 4.4: Case A

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Learning data（Cancellation) | 92.1 | 92.8 | 94.7 | 94.2 | 94.5 |
| Learning data(Continuation) | 86.4 | 88.9 | 88.9 | 90.4 | 91.8 |
| Prediction data（Cancellation) | 68.4 | 70.8 | 72.3 | 75.6 | 73.2 |
| Prediction data(Continuation) | 60.3 | 62.1 | 65.9 | 66.2 | 67.6 |

#### 判別式作成用データの判別率

表 4.5: Case B

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Learning data（Cancellation) | 91.8 | 92.1 | 92.9 | 94.6 | 94.9 |
| Learning data(Continuation) | 82.1 | 87.8 | 88.2 | 88.9 | 91.4 |
| Prediction data（Cancellation) | 70.4 | 74.9 | 76.8 | 78.2 | 75.1 |
| Prediction data(Continuation) | 68.1 | 67.3 | 68.9 | 73.5 | 74.6 |

表 4.6: Case C

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Learning data（Cancellation) | 90.2 | 91.4 | 93.7 | 94.8 | 95.2 |
| Learning data(Continuation) | 86.2 | 87.5 | 89.3 | 89.8 | 90.3 |
| Prediction data（Cancellation) | 65.2 | 66.3 | 67.1 | 70.7 | 72.9 |
| Prediction data(Continuation) | 69.4 | 68.1 | 72.3 | 75.5 | 75.2 |

判別式作成用データ (Learning data) の判別率について，抽出語数が少ない場合 CaseA が解約 : 92.1%，継続 : 86.4%で他条件に比べて最も良い値であった．抽出語数を増加すると Case A， Case B， Case C 間の差は減少している．これは Case B，Case C においてデータ数が増加しているため判別するためには，Case A よりも語数が必要であるためであると考えられる．抽出語数 500 語の場合，Case A，Case B，Case C のすべての条件において 90% を超えているため，全条件において作成用データを分類するための判別式は作成できていることが確認できる．

予測用データの判別率

予測用データ (Prediction data) の判別率について，Case A について継続グループの判別率が全抽出語数において 70%を超えず，他条件に比べ低い値となっている．これは，データ量が少ないため幅広い内容を含む継続グループの特徴をとらえきれていないためであると考えられる．Case B について解約グループの判別率は抽出語数 500 の時に 72.9%　で最良となっているが，他条件に比べて低い値となっている．これは，各企業はコンサルティングサービスの契約について 1 年に 1 回更新月を持つが，Case C では過去 2 年分のテキストデータを対象とするため，解約グループの中に前年契約を更新した際のテキスト内容が含まれてしまっているためであると考えられる．また，ここで Case C において抽出された語句について解析する．Case C で解約グループにおける要因の強い上位 10 語の出現回数を図 4.6 に示す．横軸は抽出された語句，縦軸は出現回数

をあらわす．また棒グラフの赤色は一年以上前のテキストデータに出現された回数，青色は更新月から一年以内に出現した語句を示す.「ベスト」「遵守」など1年以上前のテキストデータに多く出現している語句が抽出されている．解約グループは，解約が近づくとコンタクト回数が減少しテキスト量が減るため1年以上前のテキストデータの影響が大きくなると考えられる．よって，対象とするテキストデータはCase Bが良いと考えられる．

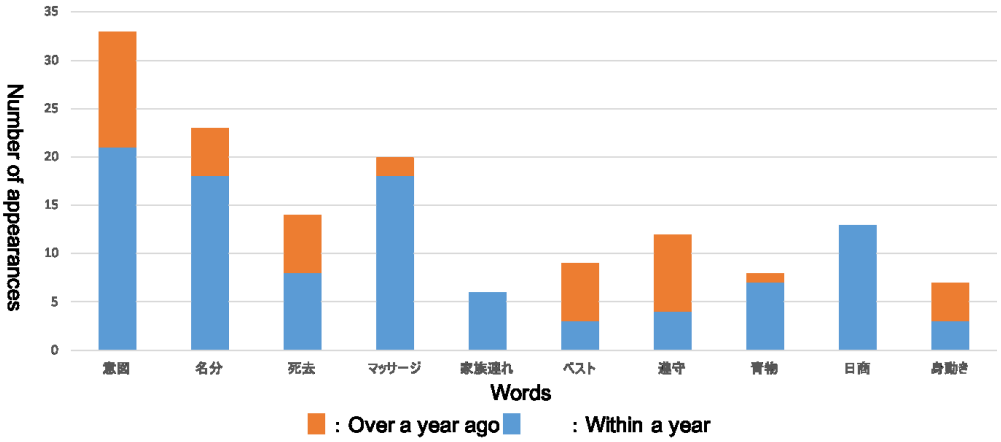



図 4.6: Extract words

## 4.5 実現場適用に向けた手法拡張の有効性検証

拡張した各手法の有効性を検証するため，既存手法と比較実験をおこなった．

### 4.5.1 特徴抽出法の有効性

#### 実験条件

ある月の解約予測を対象問題とし，実規模データを予測する．実験条件を以下に示す．作成用データの解約グループは315社，継続グループが4048社で，予測用データは解約グループ110社，継続グループ451社である．

表 4.7: Experiment Condition

| Learning data | | Prediction data | |
|---|---|---|---|
| Cancellation | Continuation | Cancellation | Continuation |
| 315 | 4048 | 110 | 451 |

- 判別境界の幅：0.5

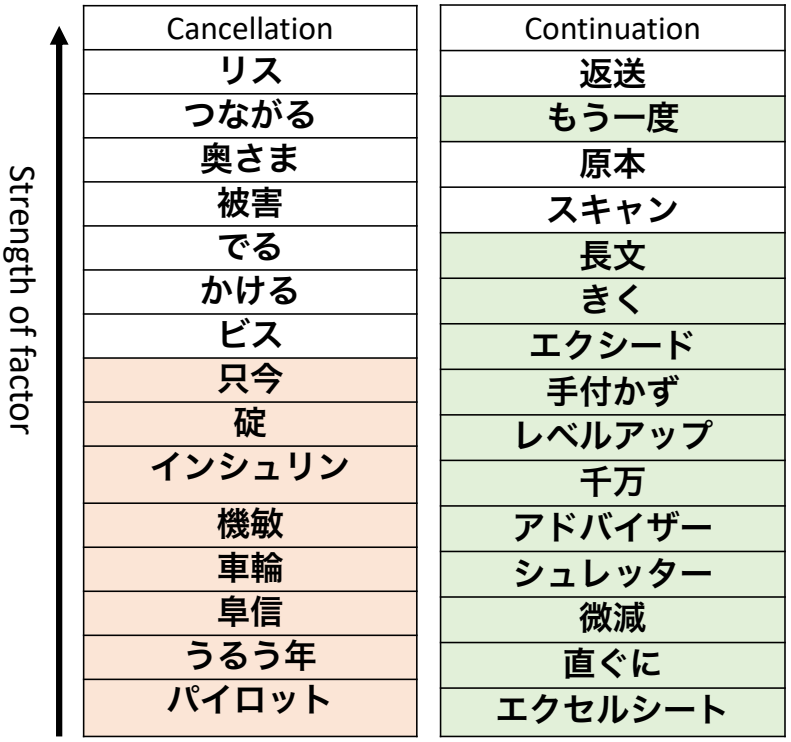
- 抽出語数：100〜500語

抽出語句

各グループにおいて要因の強い上位15語を表4.8に示す．縦が要因の強さを表しており，左が解約グループ，右が継続グループにおいて要因の強い語句として抽出された．白色の単語が両グループに出現する語句で，色付きが片方のグループのみに出現する語句である．両グループともに，片方のみに出現する語句が多く抽出されていることが確認できた．

解約グループで抽出された上位15語の原点からの距離，IDF値，更新値を図4.9に示す．更新値とは距離とIDF値を掛けた値である．図4.9において，両グループに出現する語句はグループ間の差を考慮するため，原点からの距離が小さい順にはならないが，片方のグループのみに出現する語句は原点からの距離が小さい順に抽出される．本研究では，対応分析の原点から近い距離に存在する語句に注目し容易となる語句を抽出するが，原点付近には原点付近に存在する企業でのみ少数回出現する，特定の企業にのみ出現する語句も付置される．

表 4.8: Extract words




対応分析の成分1と成分2を軸としたときの各単語の原点からの距離を図4.7に示

表 4.9: Words extracted based on the updated value

| | 原点からの距離 | IDF値 | 更新値 |
|---|---|---|---|
| リス | 0.002 | 2.115 | 0.005 |
| つながる | 0.003 | 1.782 | 0.006 |
| 奥さま | 0.005 | 4.143 | 0.022 |
| 被害 | 0.004 | 3.450 | 0.015 |
| でる | 0.002 | 1.764 | 0.004 |
| かける | 0.003 | 1.034 | 0.003 |
| ビス | 0.006 | 4.654 | 0.029 |
| 只今 | 0.007 | 4.654 | 0.031 |
| 碇 | 0.007 | 5.753 | 0.038 |
| インシュリン | 0.007 | 5.753 | 0.038 |
| 機敏 | 0.007 | 5.753 | 0.038 |
| 車輪 | 0.007 | 5.753 | 0.038 |
| 阜信 | 0.007 | 5.753 | 0.038 |
| うるう年 | 0.007 | 5.753 | 0.039 |
| パイロット | 0.007 | 5.753 | 0.039 |
| クリーンエンシュー | 0.007 | 5.753 | 0.040 |

す．縦軸は成分1，横軸は成分2である．青が単語，赤が企業を表している．実際原点付近に企業が存在していることが確認できる．

この企業に特有の単語も抽出されてしまう．そこでより汎用的な語句を抽出するた

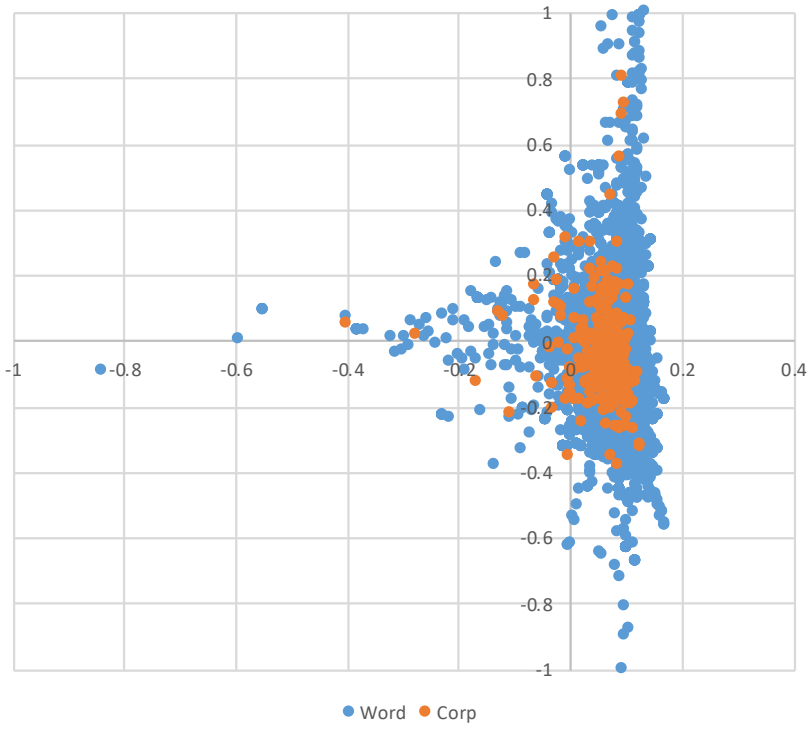



図 4.7: Distance from origin of Correspondence analysis

め逆文書頻度であるIDF値を用いて，原点からの距離を更新する．図4.9における両グループで出現する語句の上位3語である「リス」「つながる」「奥様」と片方のグループのみに出現する語句の上位3語である「只今」「碇」「インシュリン」の企業における出現分布の図を図4.8，図4.9に示す．IDF値の高い両方のグループに出現する語句は，幅広い企業で複数回出現しているため，要因の強い語句として抽出されるべきである．片方のグループのみに出現する語句は,「只今」に関して両グループに出現する語句よりは少ないが，複数の企業に複数回出現しているため汎用的な語句であるといえる．「碇」「インシュリン」に関しては，1社で1回の出現する語句なので抽出されるべきでない．これらの語句を削減するためIDF値を用いて算出される更新値をもとに抽出する．

更新値をもとに要因の強さを並び替えた語句を図4.10に示す．図4.10において，両グループ出現する語句に変化はなかった．また「只今」についても順位に変動はなかった.「只今」を除く片方のグループのみに出現する上位3語の出現分布の図を図4.10に示す．IDF適用により，汎用的な語句である「只今」は順位が変わらず抽出され，IDF適用前の「碇」「インシュリン」などの特定の企業でのみ出現している語句よりも，多

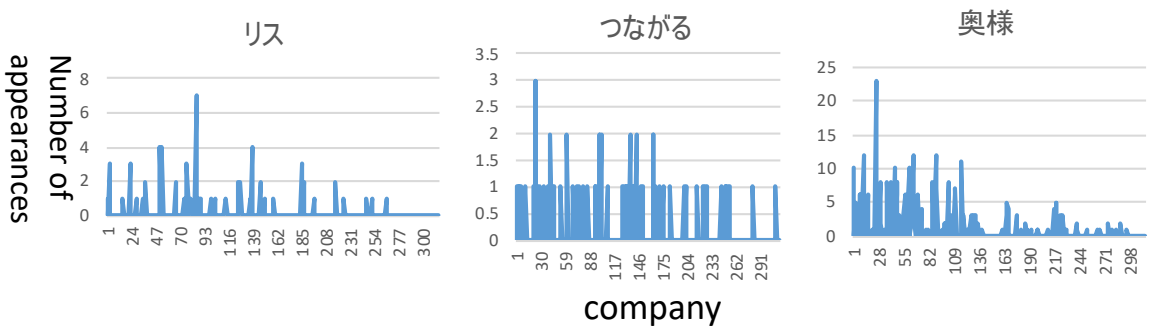



図 4.8: Number of appearances(Both group)

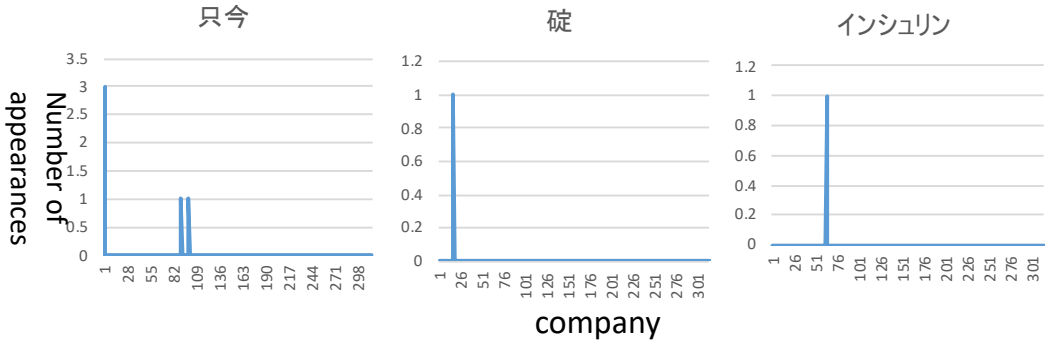



図 4.9: Number of appearances(One group)

数の企業で複数回出現している「掌房」「調理」が抽出されていることを確認した

表 4.10: Extract words

| | 原点からの距離 | IDF値 | 更新値 |
|---|---|---|---|
| リス | 0.002 | 2.115 | 0.005 |
| つながる | 0.003 | 1.782 | 0.006 |
| 奥さま | 0.005 | 4.143 | 0.022 |
| 被害 | 0.004 | 3.450 | 0.015 |
| でる | 0.002 | 1.764 | 0.004 |
| かける | 0.003 | 1.034 | 0.003 |
| ビス | 0.006 | 4.654 | 0.029 |
| 只今 | 0.007 | 4.654 | 0.031 |
| 當房 | 0.008 | 4.654 | 0.035 |
| 調理 | 0.008 | 4.654 | 0.036 |
| さし | 0.007 | 5.059 | 0.036 |
| 訪時 | 0.007 | 5.059 | 0.037 |
| 一纏 | 0.008 | 4.654 | 0.037 |
| しり | 0.007 | 5.059 | 0.037 |
| まぶしい | 0.007 | 5.059 | 0.038 |
| 碇 | 0.007 | 5.753 | 0.038 |

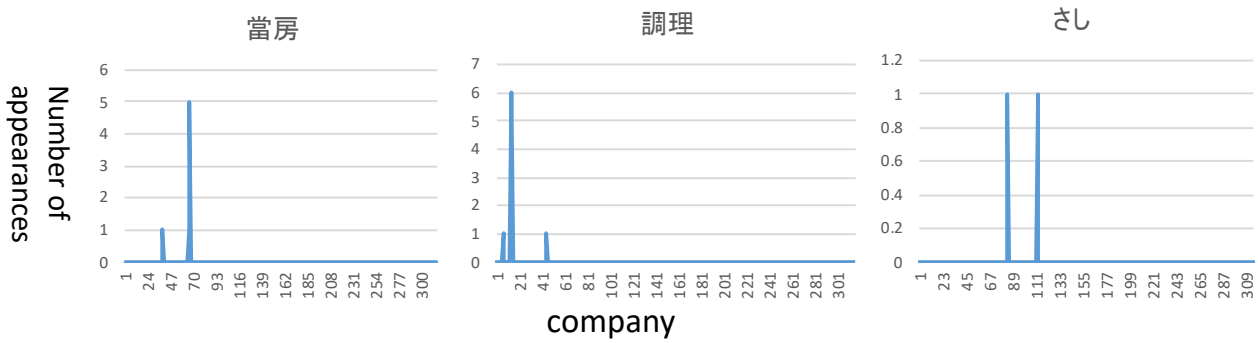



図 4.10: Number of appearances

各手法によって抽出された語句の被覆企業数を表に示す．全抽出語数において，特徴抽出手法の拡張により被覆企業数が増加した．以上より，特徴抽出の拡張手法の有効性が確認された．

表 4.11: Number of coating companies

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Existing method | 1923 | 2184 | 2561 | 2519 | 2537 |
| Improved method | 2449 | 2482 | 2812 | 2913 | 3034 |

### 4.5.2 標準化の有効性

#### 実験条件

ある月の解約予測を対象問題とし，実規模データを予測する．標準化適用の有無で予測結果を比較し，標準化の有効性を検証する．実験条件を以下に示す.

表 4.12: Experiment Condition

| Learning data | | Prediction data | |
|---|---|---|---|
| Cancellation | Continuation | Cancellation | Continuation |
| 315 | 4048 | 110 | 451 |

- 判別境界の幅0.5

- 抽出語数100〜500語

#### 実験結果

まず作成用データの予測結果について，検討する．標準化なしの判別率と判別不能企業数を表4.13，標準化ありの判別率と判別不能企業数を表4.14に示す．判別不能企業とは，対象企業のテキストデータ内において判別式作成で用いられている語句の値が0であり，解約･継続の判別ができない企業である．また，判別率は判別可能な企業における判別の正答率をあらわす．標準化なしの場合，抽出語数500語の時に解約グループ96.8%，継続グループ97.6%が最良値となっている．標準化ありでは，抽出語数300，400語の時に解約グループ98.6%，継続グループ99.9%となり，標準化なしの場合より良い判別結果となった．また判別不能企業数について，標準化なしの場合両グループについて半数以上が判別不能企業となっているが，標準化ありの場合判別不能企業数は0となった．これは出現回数が0回の単語に負の値が付与されることにより,「特定の語句

が出現しない」ということが判別において考慮可能となったためである．

表 4.13: No standardization(Learning data)

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discriminant rate（Cancellation) | 98.3 | 98.4 | 98.3 | 98.4 | 96.8 |
| Discriminant rate(Continuation) | 86.3 | 94.3 | 94.3 | 94.3 | 97.6 |
| Undecideble company（Cancellation) | 256 | 254 | 255 | 253 | 253 |
| Undecideble company(Continuation) | 3833 | 3890 | 3890 | 3890 | 3715 |

表 4.14: Standardization(Learning data)

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discriminant rate（Cancellation) | 98.4 | 97.8 | 98.4 | 98.6 | 98.6 |
| Discriminant rate(Continuation) | 98.2 | 99.9 | 99.2 | 99.9 | 99.9 |
| Undecideble company（Cancellation) | 0 | 0 | 0 | 0 | 0 |
| Undecideble company(Continuation) | 0 | 0 | 0 | 0 | 0 |

検証用データの予測結果と判別結果について，検討する．標準化なしの判別率と判別不能企業数を表4.15，標準化ありの判別率と判別不能企業数を表4.16に示す．標準化なしの場合，抽出語数500語の時に解約グループ72.0%，継続グループ66.7%が最良値となっている．標準化ありでは，抽出語数500語の時に解約グループ75.0%，継続グループ67.8%となり，標準化なしの場合より良い判別結果となった．また判別不能企業数について，標準化ありの場合は標準化なしに比べて減少している．しかし両グループともに半数以上が判別不能企業であり．以上より，標準化は出現回数が0回の単語に負の値を与えることにより，作成用データにおいて判別負の企業数を0とし，予測用データにおいても判別不能企業数が減少したため，標準化の有効性が確認された．

しかし予測用データにおいては標準化ありの場合でも半数以上が判別不能企業であることから，さらなる手法の拡張が必要である．次節では，判別不能企業数を削減する対策手法として，データ拡張を提案する．

表 4.15: No standardization(Prediction data)

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discriminant rate（Cancellation) | 68.1 | 55.2 | 77.3 | 62.7 | 72.0 |
| Discriminant rate(Continuation) | 43.0 | 63.1 | 37.1 | 68.2 | 66.7 |
| Undecideble company（Cancellation) | 63 | 52 | 66 | 51 | 60 |
| Undecideble company(Continuation) | 337 | 245 | 362 | 226 | 271 |

表 4.16: Standardization(Prediction data)

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discriminant rate（Cancellation) | 69.4 | 56.4 | 77.0 | 63,1 | 69.2 |
| Discriminant rate(Continuation) | 42.8 | 63.7 | 40.4 | 66.9 | 67.1 |
| Undecideble company（Cancellation) | 61 | 48 | 59 | 45 | 44 |
| Undecideble company(Continuation) | 332 | 236 | 347 | 206 | 241 |

### 4.5.3 判別分析の目的関数拡張の有効性

目的関数拡張の有効性を検証するため以下の3条件で比較検証を行う.

- CaseA : データの一部を用いた同じ企業数
- CaseB : グループ間に企業数の差がある実規模データ（既存手法）
- CaseC : グループ間に企業数の差がある実規模データ（改良手法）

#### 実験条件

各条件の実験条件を表**??**に示す. CaseB, CaseCでは実現場における解約率を参考にグループ間の差を設定した. 検証のため, 全条件において予測対象データの企業数は一定である.

- 抽出語数 : 300
- 判別境界の幅 : 0.5
- 試行回数 : 5回

表 4.17: Experiment Condition

| | Learning data | | Prediction data | |
|---|---|---|---|---|
| | Cancellation | Continuation | Cancellation | Continuation |
| Case A | 300 | 300 | 90 | 90 |
| Case B | 300 | 2000 | 90 | 90 |
| Case C | 300 | 2000 | 90 | 90 |

**実験結果**

最良値である試行時のCaseA，CaseBにおける判別スコアが-1.5〜1.5の範囲の各企業の判別スコアを図4.11，図4.12に示す．縦軸は判別スコアを表し，判別スコアが判別境界より高いと解約と判別され，低いと継続と判別される．横軸は企業番号を表し，左が解約グループ，右が継続グループの企業である．黄色の横線は判別境界を表し，実線と点線の間がオーバーラップ領域である．これは，判別が2段階に分かれているDEA判別分析においてstage1で判別の困難である企業を抽出するための領域である．CaseAでは，CaseBに比べてstage1からうまく判別できていることが確認できる．CaseAが一部のデータを対象とし総企業数が少ないため，判別しやすいと考えられる．CaseBでは企業数の少ないグループである解約グループについて，stage1でオーバーラップ領域の判別スコアである企業が多数存在する．企業数の多いグループに偏った判別式が作成されていることが確認できる．

全条件の判別率の平均を表4.18に示す．作成用データについて，CaseAが解約グループ99.0% 継続グループ97.0%となり全条件の中で最も良い値である．CaseAが一部のデータを対象とし総企業数が少ないため，判別しやすいと考えられる．CaseBにおいて，解約グループ88.7% 継続グルー91.4%とであり企業数の継続グループに判別式が偏っているのに対して，CaseCにおいて，解約グループ94.8%継続グループ86.6%と解約グループのほうが高い判別率となっている．予測用データにおいてCaseBで解約グループ61.2% 継続グループ86.2%であり，企業数の多い継続グループに偏った判別式が作成されてしまい解約グループがうまく判別できていない．CaseCにおいて，解約グループ75.0%継続グループ82.0%であり，解約グループのほうが高い判別率でありグループ間の判別率の差が小さいことから，企業数の大きさに影響しない判別式が作成できていることが確認できる．さらに目的関数の重みに変数を加えることにより，解約・継続の両グループにおいてバランスの良い判別式を作成できることが示唆される．


図 4.11: CaseA

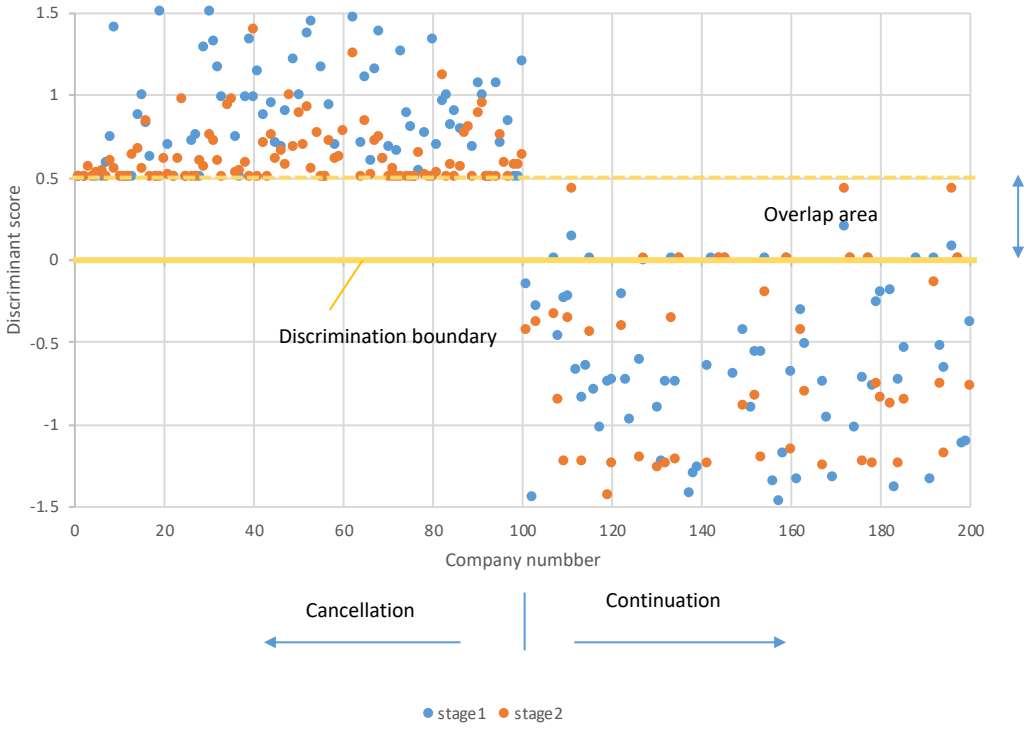



図 4.12: CaseB

表 4.18: Discrimination rate

| | Learning data | | Prediction data | |
|---|---|---|---|---|
| | Cancellation | Continuation | Cancellation | Continuation |
| Case A | 99.0 | 97.0 | 78.8 | 89.6 |
| Case B | 88.7 | 91.4 | 61.2 | 86.2 |
| Case C | 94.8 | 86.6 | 75.0 | 82.0 |

### 4.5.4 類義語辞書を用いたデータ拡張の有効性

#### 計算機実験

以下の3条件で比較することにより，データ拡張の有効性を検証する．

- CaseA : データ拡張無し
- CaseB : 各語句の類義語も同一の語とみなすことによるデータ拡張
- CaseC : 各語句の上位語を対象とすることにより表記ぶれを考慮する

Case A はデータ拡張を用いない判別をする．CaseB では，判別式で用いられる語に対して類義語も対象とすることによりデータ拡張をおこなう．CaseC では，判別式で持ちいられる語の上位語も判別の対象語句とすることにより，表記ぶれを考慮する．

#### 語句の拡張

解約グループで抽出された要因の強い上位5語とその語に対する類義語と上位語の例を表4.19，表4.20に示す.「リス」については，危機管理サービスの名称の一部であるが，略称であるため形態素解析がうまくいかず「リス」で区切られたため類義語，上位語ともに存在しない．

表 4.19: CaseB(Synonym)

| Words | Synonym |
|---|---|
| リス | - |
| フレーズ | イディオム，慣用句 |
| 痩せる | 痩せ細る，やせこける，スリムアップする，減量する |
| 只今 | やがて，ちょうど，ついさっき，近いうち，今にも，まもなく |
| 滞納 | 不払い，不履行，デフォルト |

表 4.20: CaseC(Hypernym)

| Words | Hypernym |
|---|---|
| リス | - |
| フレーズ | 構文，パッセージ |
| 痩せる | 変わる，転じる，なる |
| 只今 | - |
| 滞納 | しくじり，失敗，損金 |

データ拡張量を表4.21に示す．両条件においてデータ拡張ができている．特徴抽出において汎用的な語句を抽出できていることが示唆される．各条件の予測データの判

表 4.21: Data augmentation

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Case B | 342 | 602 | 1047 | 1892 | 2363 |
| Case C | 214 | 576 | 947 | 1695 | 1948 |

別率と判別不能企業数を表4.22，表4.23，表4.24に示す．

表 4.22: Case A(No data augmentation)

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discriminant rate（Cancellation) | 68.1 | 55.2 | 77.3 | 62.7 | 72.0 |
| Discriminant rate(Continuation) | 43.0 | 63.1 | 37.1 | 68.2 | 66.7 |
| Undecideble company（Cancellation) | 63 | 52 | 66 | 51 | 60 |
| Undecideble company(Continuation) | 337 | 245 | 362 | 226 | 271 |

CaseAでは解約グループで72.0%継続グループで66.7%，CaseBでは解約グループで75.0%継続グループで67.8%，CaseCでは解約グループで69.0%継続グループで67.1% で抽出語数500語で最良となり，全条件の中でCaseBの判別率が最良値であった．全条件において抽出語数が増えるにつれて判別不能企業数が減少している．また，データ拡張をおこなったCaseB，CaseCにおいて判別率が最良である抽出語数500の時に，CaseBではCaseAと比較して解約グループ12社継続グループ38社判別不能企業数が減少し，CaseCではCaseAと比較して解約グループ5社継続グループ21社判別不能企業数が減少している．これらよりデータ拡張の有効性が確認された．またCaseBとCaseCで比較

表 4.23: CaseB(Synonym)

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discriminant rate（Cancellation) | 70.0 | 60.9 | 78.5 | 61.7 | 75.0 |
| Discriminant rate(Continuation) | 43.0 | 62.7 | 43.2 | 70.9 | 67.8 |
| Undecideble company（Cancellation) | 60 | 48 | 52 | 42 | 48 |
| Undecideble company(Continuation) | 333 | 239 | 360 | 221 | 233 |

表 4.24: CaseC(Hypernym)

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discriminant rate（Cancellation) | 68.1 | 55.2 | 78.2 | 62.9 | 69.0 |
| Discriminant rate(Continuation) | 43.0 | 63.3 | 39.3 | 67.3 | 67.1 |
| Undecideble company（Cancellation) | 63 | 52 | 61 | 48 | 55 |
| Undecideble company(Continuation) | 337 | 241 | 352 | 211 | 250 |

すると，判別率，判別不能企業数ともにCaseBの方が良い値である．上位語を用いて表記ブレを考慮するより，類義語を用いたデータ拡張が有効であることが確認された．CaseCでは上位語にまとめることにより，意味の異なる語句も同じ語句としてまとめられてしまい，特徴を捉えづらくなってしまったため判別率が下がったと考えられる．

## 4.6 実規模問題を対象とした解約予測

実規模問題を対象に解約予測をおこない，有効性を検証する．抽出された語句と，判別結果について検討する．

### 4.6.1 要因となる語の抽出

各グループにおいて要因が強い語句として抽出された上位10語を表4.25，表4.26に示す．表のWord列が抽出された語句で，Strengthが要因の強さを表している．解約グループにおいて，“滞納”“不本意”などの解約問題に関係する語句が抽出されている．また，一見解約と関係のない“リス”が解約グループにおいて抽出されている．“リス”は管理サービス名の略称であり，解約グループでは管理サービスに関する内容が含まれているということは，管理サービスが導入されていないまたは管理サービスがうまくいっていず，サービスに満足していない状態にあると推測されることを，ディスカッションによって発見した．このように，提案手法によって抽出された語句はコンサルタ

表 4.25: Cancellation

| Word | Strength |
|---|---|
| リス | 3.172 |
| フレーズ | 3.142 |
| やせる | 3.100 |
| 滞納 | 2.917 |
| 被害 | 2.760 |
| 奥様 | 2.692 |
| 只今 | 2.638 |
| 些事 | 2.534 |
| これだけ | 2.320 |
| 不本意 | 2.307 |

表 4.26: Continuation

| Word | Strength |
|---|---|
| 返送 | 3.491 |
| もう一度 | 2.962 |
| 原本 | 2.687 |
| スキャン | 2.504 |
| 長文 | 2.475 |
| きく | 2.447 |
| レベルアップ | 2.398 |
| 直ぐに | 2.198 |
| 各日 | 2.155 |
| 整頓 | 2.105 |

ントの経験や勘からは要因として選ばれない語句が含まれており，ディスカッションすることにより新たな知見を得るきっかけとなることが期待できることが実規模問題においても確認された．

### 4.6.2 判別率

実現場適用に向けて拡張された手法を用いて，ある月を対象として予測した結果を示す．表4.27に拡張前の判別率と判別不能企業率，表4.28に手法拡張後の判別率と判別不能企業数を示す．

表 4.27: Existing method

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discrimination rate（Cancellation) | 68.1 | 55.2 | 77.3 | 62.7 | 72.0 |
| Discrimination rate(Continuation) | 43.0 | 63.1 | 63.1 | 67.1 | 66.7 |
| Unidentifiable company rate（Cancellation) | 54.5 | 47.3 | 60.0 | 46.4 | 54.5 |
| Unidentifiable company rate(Continuation) | 74.7 | 54.3 | 80.3 | 50.1 | 60.1 |

抽出語数が少ないと判別がうまくいっていず，判別不能な企業率も高い．拡張前は判別不能な企業が全企業の半数以上であったのに対して，手法拡張後は判別不能な企業率が最も良い時に両グループで10% 以下にとどまっている．最も判別率が良いのは手法拡張後の抽出語数400語の時の解約グループ71.7%，継続グループ70.4% であり両グループともに70%　以上の判別率となった．実現場適用に向けて次節では，コンサルタントの予測結果と比較する．

表 4.28: Proposal method

| | Number of extract words | | | | |
|---|---|---|---|---|---|
| | 100 | 200 | 300 | 400 | 500 |
| Discrimination rate（Cancellation) | 68.0 | 62.4 | 65.3 | 71.7 | 72.2 |
| Discrimination rate(Continuation) | 52.4 | 58.7 | 62.2 | 70.4 | 68.9 |
| Unidentifiable company rate（Cancellation) | 20.9 | 16.4 | 10.9 | 10.9 | 7.3 |
| Unidentifiable company rate(Continuation) | 13.5 | 11.8 | 11.5 | 10.0 | 8.2 |

## 4.7 コンサルタントの予測結果との比較

実現場適用に向けて，前節で最良解の予測結果とコンサルタントの解約予測結果を比較する．図 4.13 は解約グループの比較，図 4.14 は継続グループの比較結果を示す．縦軸がコンサルタントの解約予測の値であり，横軸が企業数を表す．また，棒グラフの青色が計算機実験において解約と判別された企業で，赤色が継続グループと判別された企業である．例えば図 4.13 において，実際に解約している企業の中でコンサルタントが 100% 継続すると予測した企業数は 4 社であり，そのうち計算機で解約と予測した企業は 1 社，継続と予測した企業は 3 社であると読み取ることができる．

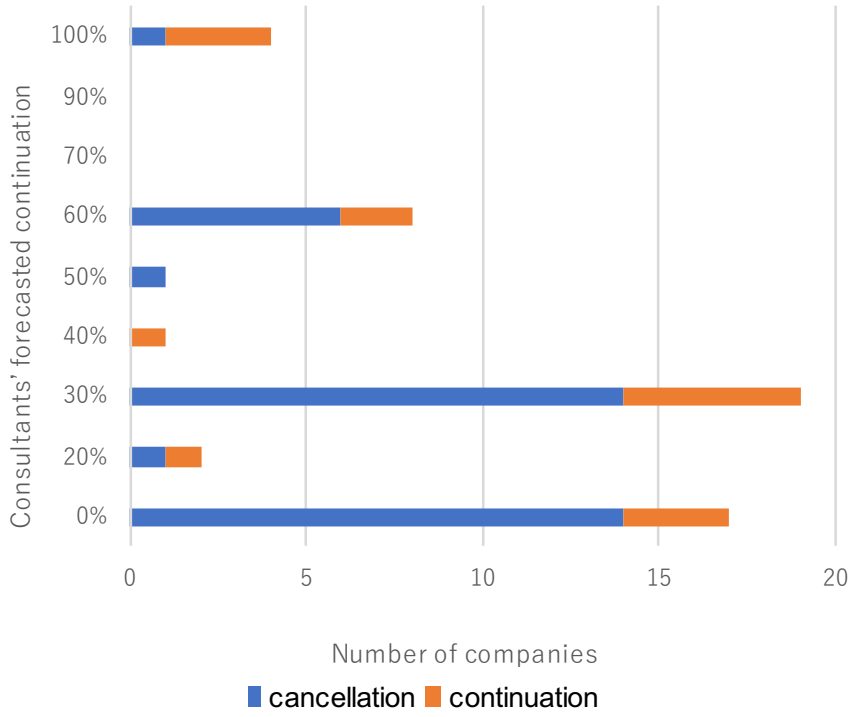



図 4.13: Cancellation

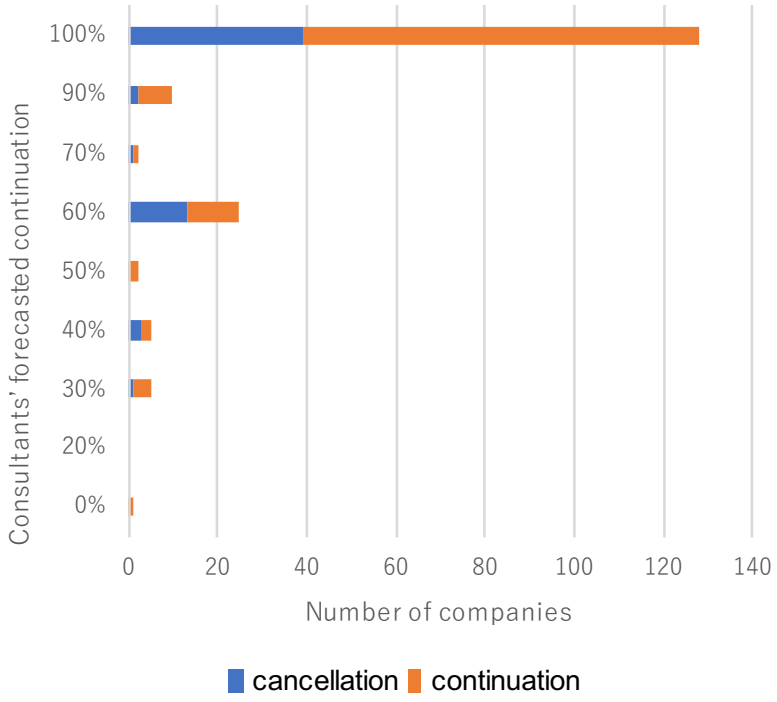



図 4.14: Continuation

図 4.13 において，コンサルタントの更新見込みが低い企業について，提案手法でも判別ができていた．コンサルタントが更新見込み 100% と見込んでいた企業に対して計算機の予測結果も 4 社中 3 社が継続すると予測されていて，これらの企業は解約を予測することが難しい企業であったと推測される．コンサルタントが 50%，60% 継続すると予測していた企業については計算機実験の予測では 9 社中 7 社が解約すると予測できていることから，コンサルタントの更新見込みが 50% 近くの企業について計算機の予測結果のほうが良い判別結果となっている．図 4.14 において，コンサルタントの更新見込み予測が 100% である企業に対して，計算機実験では半数以上継続と予測したが，コンサルタントの予測結果より低い値となった．これは継続グループの企業は幅広く，解約グループよりも特徴を捉えづらいためであると考えられる．コンサルタントが更新見込み率 50% 以下と予測していた企業に対して，計算機で継続を予測できていた企業が多数存在した．以上のことから，コンサルタントが解約予測をおこなう際に捉えいてる特徴を提案手法においても考慮できていることが示唆された．また，コンサルタントが自身のない更新見込み 50% 付近の企業について，提案手法による解約予測支援が可能であることが示唆された．

## 4.8 考察

各計算機実験の結果から，以下の特性が分かった．

- テキストデータ範囲の選定
  対象とするテキストデータは過去9カ月のテキストが最良であることを確認した．過去の多くの履歴を用いすぎると問題と関連のない語句がノイズとなってしまい，履歴データが少なすぎると汎用的な判別式が作成されないためである．

- 実現場適用に向けた手法拡張の有効性

  - 実規模問題の有する特徴を挙げ，各問題に対して手法を拡張した．特徴抽出法の拡張では，対象とする語句を広げる一方でIDF値を用いることにより，汎用的に用いられる語句の中から要因となる語句を抽出することを可能とした．
  - 類義語辞書を用いたデータ拡張では，2種のデータ拡張手法を比較することにより，上位語より類義語を用いるデータ拡張が妥当であると確認され，データ拡張前と比較し判別不能企業数が減少することを確認した．
  - 標準化により，判別に寄与することのできなかった出現回数が0回の語句に負の値を与えることにより,「単語が出現しない」ことも判別において考慮可能となった．
  - 判別分析の目的関数拡張により，グループ間のデータ数の差に依存しない判別式を作成することが可能となった．

- 実規模問題を対象とした解約予測
  抽出された語句の中には，コンサルタントが見て対象問題と関係があると分かる語が含まれていおり，対象グループの要因となる語が実際に抽出されていることを確認した．さらにコンサルタントが気づくことのできなかった語句についても，計算機実験から得られた抽出語をもとに議論することによって，新たな関連性を推測することできた．手法の拡張により判別率の向上，判別不能企業数の減少が確認された．

- コンサルタントの予測結果との比較
  コンサルタントが解約予測をおこなう際に捉えいてる特徴を提案手法においても考慮できていることが示唆された．また，コンサルタントが自身のない更新見込み50% 付近の企業について，提案手法による解約予測支援が可能であることが示唆された．

## 4.9 結言

本章では解約問題を対象とし，実現場適用に向けて実規模問題を解析し，拡張手法の有効性を検証した．4.2節では実規模問題の特徴と拡張した支援手法について説明した．4.3節では実験条件を述べ，4.4節で対象とするデータ範囲を選定するための解析をおこなった．4.5節では，実現場適用に向けて拡張した各手法の有効性を検証した．4.6節では，拡張された手法を用いて実規模問題を解析し，解約問題の予測をおこなった．4.7節では，コンサルタントの解約予測結果と比較検証することにより，提案手法によるコンサルタント支援の可能性を検証した．4.8節では実験結果についての考察を述べた．次章では結果に不確実性を含む問題である不正問題に提案手法を適用し，その有効性を確認する．

# 第5章　不確実性を含む問題への適用

## 5.1　緒言

本章では，汎用的な問題に支援適用可能なシステム構築のため，不確実性を含む問題を対象に問題の発生予測をおこなう支援手法を提案する．また不確実性を含む問題を対象として，提案手法を用いて実データを解析することにより有効性を検証する．具体的には，クライアント企業の不正問題を対象とする．5.2節では不確実性を含む問題の特徴と拡張手法について述べる．5.3節で実験対象について述べ，5.4節では，クライアント企業に関する業種や地域が判別率に及ぼす影響について検証する．5.5節で提案手法によって抽出された要因となる語句について考察し，判別式の作成，新たなデータに対して予測を行う．さらに，ロジスティック曲線を用いて不正問題を発生確率で表し，コンサルタントの注目すべき企業を特定する．5.6節では，不確実性を含む正解データに対して検証するためクライアント企業に実際にヒアリングを行った結果と提案手法による判別を比較する．5.7節では実験結果についての考察を述べる．最後に，5.8節で本章についてまとめる．

## 5.2　不確実性を含む問題を対象とした支援手法

本章では，汎用的な問題に支援適用可能なシステム構築のため，不確実性を含む問題を対象に問題の発生予測をおこなう支援手法を提案する．不確実性を含む問題をの特徴を述べ，拡張手法について説明する．

### 5.2.1　不確実性を含む問題

解約問題はある時点において，コンサルティングサービスを解約しているか継続中かが明確に分かるため，正解データが確実な問題である．一方，不正問題についてはある時点で不正問題が発生しているかどうかについて，コンサルタントに相談していないがクライアント企業内では問題が発生している場合，クライアント企業も気づいていないが実際は問題が発生している場合などが存在するので，テキストデータの内容から不正問題発生の有無は明確には分からない．よって判別式を作成する上での正解データに不確実性を含む問題となっている．コンサルタントが支援すべき企業は，正解データにおいて不正問題発生無しと分類されているが実際には不正問題が発生し

ている可能性の高い企業である．コンサルタントが支援できる企業数は限られているため，誤判別データの中から実際には不正問題が発生している可能性の高い企業を特定する必要性がある．本研究では，2 値分類から問題発生の確率を算出するために手法を拡張する．

### 5.2.2 ロジスティック曲線を用いた支援手法

ロジスティック曲線は2分類のデータを直線の代わりにシグモイド関数で表す手法である[49]．判別分析で取得した判別スコアから，シグモイド関数を用いることにより問題発生の確率を算出することが可能となる．これにより，正解データに不確実性を含む問題に対して，問題発生ありと判別された企業の中からコンサルタントが注意を払うべき企業を特定することが可能となる．式(5.1)に判別スコアから確率を計算式を示す．

$$p = \frac{1}{1 + (\frac{1-\pi_1}{\pi_1}) * \exp(-z)} \tag{5.1}$$

$p$は問題発生確率，$\pi_1$は問題発生の事前確率，$z$は判別スコアを表す．

## 5.3 対象問題

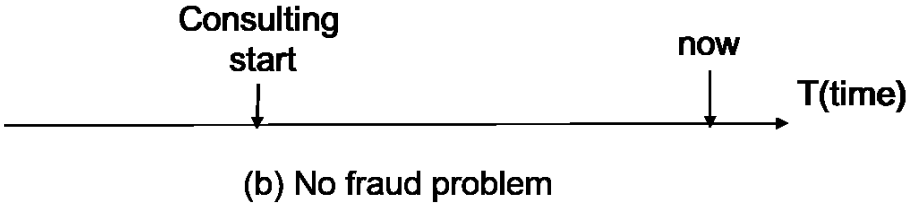
提案手法を用いて，クライアント企業の不正問題発生を予測する判別式を作成する．まずはテキストデータの分類について説明する．

#### テキストの分類

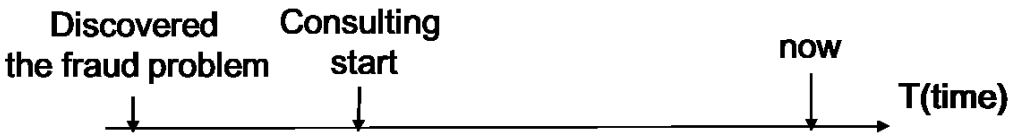
不正問題発覚の時系列とグループ分類の関係を図5.1，図5.2に示す．

図5.1は，不正問題なしのグループに分類される企業の時系列を示す．コンサルティングが始まってから不正問題発覚の記述がない場合は不正問題なしのグループに分類される．また不正問題発覚の記述があったとしても，発覚がコンサルティングの開始以前であれば不正問題なしのグループに分類する．これは過去にその企業で不正問題が起こっていたとしても，不正問題発覚がコンサルティング開始以前のことであるので，コンサルティング開始時点では改善されているはずであり，テキストデータに不正問題発生の特徴は表れないと考えられるためである．

図5.2は不正問題ありのグループに分類される企業の時系列を示す．コンサルティング期間中に不正問題が発覚した場合，不正問題ありのグループに分類される．このとき不正問題発覚以降のテキストデータはすべて分析対象から除外する．これは，本研究が不正問題発生の予測を目的とするためである．同様の理由により，コンサルティ

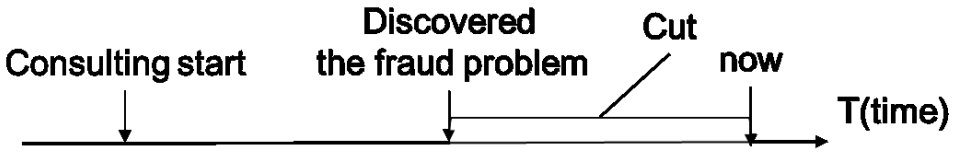



(b) No fraud problem

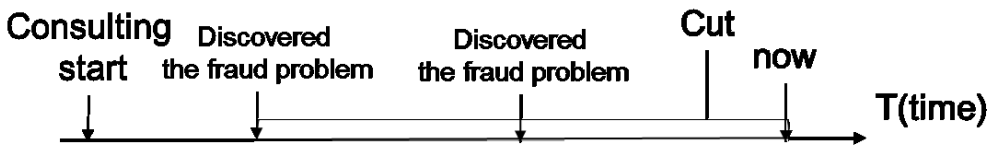



(a)The fraud problem discovered before consulting start

図 5.1: Normal group




(b)The fraud problem discovered before consulting start




(a)The number of fraud problems discovered
before consulting start

図 5.2: Fraud problem group

ング開始から複数回不正問題が発覚した場合も，一回目の不正問題発覚以降のテキストデータは削除する．

テキストデータより得られた実験条件を以下に示す．

表 5.1: Experiment Condition

| | 不正問題グループ：$G_1$ | 不正問題発覚グループ：$G_2$ |
|---|---|---|
| 企業数 | 20($e_1$〜$e_{20}$) | 20($n_1$〜$n_{20}$) |
| 段落(相談内容件数) | :3,731 | 3,946 |
| 文 | 11,978 | 14,550 |
| 総単語数 | 236,880 | 269,980 |
| 異なり語数 | 85,92 | 9,074 |

**実験環境**

- OS:Windows 8.1
- CPU:Intel(R) Core(TM) i5-4460S CPU @ 2.90GHz
- メモリ:8.0GB

## 5.4 業種と地域の影響

不正問題と企業の業種は関連が強い[50]．そこで本節では，クライアント企業の業種や地域が判別に及ぼす影響について検証する．対象データの企業の業種について，表5.2に示す．業種はTSR業種コードブックを参考に分類した[51]．

対象データの会社が存在する地域について，表5.3に示す．共同研究先企業のコンサルティング担当地域振り分けに基づいて，企業を5つの地域に分類した．

クライアント企業の業種や地域の偏りによって判別式に及ぼす影響を確かめるため，判別式作成に用いる不正問題発覚なしのテキストデータを以下の条件でそろえて，判別式の作成，新たなデータに対して検証を行った．

**実験条件**

1. 業種に偏りを持たせる
   (a) 製造業のみ
   (b) 卸売業のみ

表 5.2: A type of industry

| | A. Agriculture | B. Fishery | C. Mining | D. Construction industry |
|---|---|---|---|---|
| $G_1$ | 0 | 0 | 0 | 5 |
| $G_2$ | 0 | 0 | 18 | 21 |
| | E. Manufacturing | F. Electric Industry | G. Information and Communication Industry | H. Transport Industry |
| $G_1$ | 3 | 0 | 1 | 1 |
| $G_2$ | 1 | 0 | 1 | 3 |
| | I. Wholesale | J. Finance | K. Real Estate | L. Academic Research |
| $G_1$ | 18 | 0 | 3 | 0 |
| $G_2$ | 24 | 1 | 2 | 5 |
| | M. Accommodation | N. Entertainment | O. Education | P.Medical |
| $G_1$ | 2 | 1 | 0 | 1 |
| $G_2$ | 2 | 1 | 2 | 0 |
| | Q.compound service | R. service industry | S. Public affairs | Total |
| $G_1$ | 0 | 4 | 0 | 40 |
| $G_2$ | 1 | 8 | 0 | 90 |

表 5.3: Region

| Region | a | b | c | d | e | Total |
|---|---|---|---|---|---|---|
| $G_1$ | 8 | 10 | 7 | 11 | 4 | 40 |
| $G_2$ | 27 | 26 | 8 | 20 | 9 | 90 |

2. 地域に偏りを持たせる

3. ランダム(5回)

4. 学習データと予測データの業種・地域をそろえる

条件1では，判別式の作成のために用いる企業を 1 種の業種に絞って判別式を作成した．$1-(a)$では共同研究企業の取引先の多い業種である製造業，$1-(b)$では業種別で不正問題発生率が最も高い業種である卸売業について解析する．条件2では，最も取引先の多い地域である$a$地域の企業のみを用いて判別式を作成する．条件3は，ランダムに抽出した企業を用いた判別結果の平均値を計算する．条件4はクライアント企業の業種や地域の差がグループに無いように，業種・地域数をそろえて判別式を作成した．各条件で抽出して作成した判別式の判別率を表5.4に示す.

学習データにおいて，ランダム抽出の条件より業種や地域に偏りを持たせたときのほうが判別率が良い．業種や地域によって使用される語句に特徴があるため，うまく判別されていると考えられる．一方予測データにおいては，ランダムのときのほうが判別率が良い．業種に偏りを持たせて作成した判別式は限られた語で作成されているため，汎用性が低いといえる．

各グループの業種と地域をそろえた時の判別率が最も良い．これは業種や地域の影

表 5.4: Discrimination rate

| 抽出条件 | Manufacturing industry | Wholesale trade | Average | Region | Discrimination rate when aligning industries / areas |
|---|---|---|---|---|---|
| Learning data(% ) | 97.5 | 90 | 93.75 | 95 | 97.37 |
| Prediction data(% ) | 45 | 52.5 | 48.75 | 47.5 | 60 |

| Extraction condition | 1 | 2 | 3 | 4 | 5 | Average |
|---|---|---|---|---|---|---|
| Learning data(% ) | 90 | 87.5 | 92.5 | 92.5 | 95 | 91.5 |
| Prediction data(% ) | 57.5 | 52.5 | 52.5 | 50 | 45 | 51.5 |

響を受けずに，不正問題発生の有無で判別するための判別式を作成できたためであると考えられる．グループの業種と地域をそろえたときの予測データにおける各グループの判別率を表5.5に示す．本実験においても不正問題発覚ありグループよりも不正問題発覚なしグループの方が判別率が低い．

表 5.5: Discrimination rate when aligning industries / areas

| | Prediction data($G_1$) | Prediction data($G_2$) | Average of Prediction data |
|---|---|---|---|
| Discrimination rate(%) | 80 | 40 | 60 |

## 5.5 不正問題を対象とした問題予測

### 5.5.1 要因となる語の抽出

計算機実験によって得られた,要因となる語句として抽出された上位10語を不正問題発覚ありのグループは表5.6に，不正問題発覚なしのグループは表5.7に示す.

不正問題発覚ありのグループにおいて抽出された語句の中に“残業”,“給与”,“払う”という語句が抽出されている．横領が発生した場合，返済にあたって横領者が企業に残業代や未払い給与を請求することが多い．そのため，不正問題が発生しした企業がコンサルタントに残業や給与についての制度や規定を相談したため，不正問題発覚ありグループのテキストデータにこれらの語句が出現していたと考えられる．また不正問題発覚なしのグループにおいて“労務”が抽出されている．労務規定や就業規則をしっかりと作成しており，就業者の管理ができているため不正問題が発生していないと考えられる．このように両グループにおいて対象問題と関連性を推察できる語句

表 5.6: Fraud problem

| Word | Strength |
|---|---|
| 理由 | 2.167 |
| 控除 | 1.223 |
| 本人 | 1.1589 |
| 教育 | 0.986 |
| 計画 | 0.923 |
| 利用 | 0.920 |
| 残業 | 0.913 |
| 判断 | 0.865 |
| 給与 | 0.844 |
| 払う | 0.712 |

表 5.7: No fraud problem

| Word | Strength |
|---|---|
| 保険 | 2.872 |
| 完了 | 2.375 |
| 追記 | 2.185 |
| 部分 | 1.709 |
| 依頼 | 1.679 |
| 税理士 | 1.629 |
| 売上 | 1.598 |
| 当社 | 1.582 |
| 回収 | 1.489 |
| 労務 | 1.342 |

が抽出されていた.

### 5.5.2 DEA判別分析

5.5.1節で求めた語を変数に用いてDEA判別分析を行い，判別式を作成する．また，新たなデータに対して判別することにより，作成した判別式の有効性を検証する．本研究では学習法は用いていないが，説明のため判別式作成に用いたデータを学習データ，判別式を検証するために用いたデータを予測データとよぶ.

#### 抽出語数変化と判別率

5.5.2節で抽出する語数を変化させて，判別式を作成する．実験条件を以下に示す.

表 5.8: Experiment Condition

| | 不正問題発覚あり | 不正問題発覚なし |
|---|---|---|
| 作成用データ | 20 | 20 |
| 予測用データ | 20 | 20 |

- 抽出語数：1〜100語

- 繰り返し回数：10回

実験結果を図5.3に示す．図の縦軸は判別率を，横軸は抽出語数を表している．青線，赤線が判別式作成に用いたデータで，灰線，黄線が判別式を検証するために用いた新たなデータの判別率を表している．学習データにおいて抽出語数7語で両グループとも

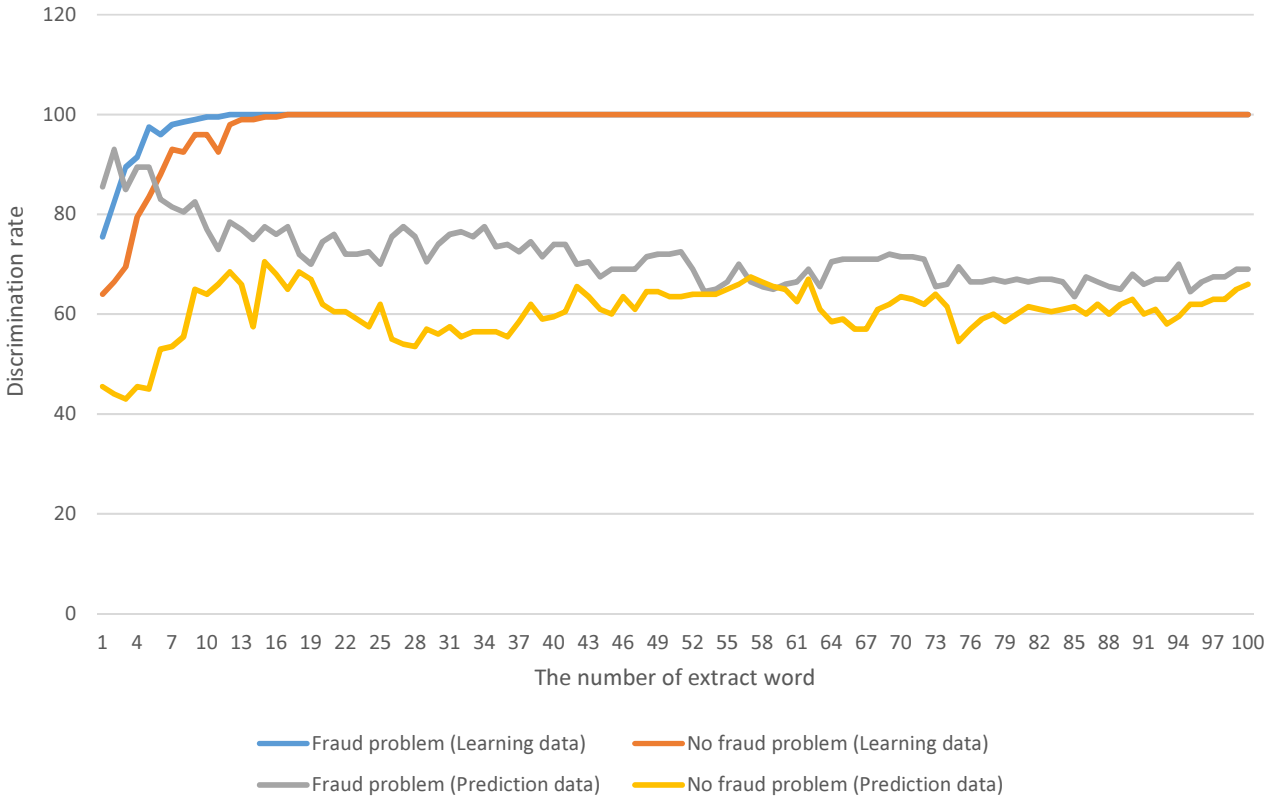



図 5.3: The number of extract word and discrimination rate

に判別率が90% をこえ，不正問題発覚ありグループは12語，不正問題発覚なしグループは17語で判別率が100% に達し，それ以上の語句数では常に100% であった．不正問題発覚ありのグループの方が少ない語数で100% に収束することから，不正問題発覚ありのグループのテキストデータ内には不正問題の特徴を表す語句が含まれていると考えられる．予測データにおいて，不正問題発覚ありグループの方が不正問題発覚なしグループよりも判別の精度は良かった．いずれのグループも3.4節でおこなった解約問題を対象とした判別率よりも低い結果となった．これは不正問題発覚なしのグループの中には発覚はしていないが，実際は不正問題が発生している企業が存在する可能性があるという，本対象問題の不確実性を含む問題構造によるものと考えられる．

### 企業数の変化と判別率

テキストデータの量と判別率の関係を調べるために，判別式作成に用いるテキストデータの企業数を変化させて判別式を作成する．実験条件を以下に示す.

- 判別式を作成するために用いた企業（学習データ）
  - 企業数：$g_1$ ($g_1$ = 5,10,15,20,25,30,35)
  - オーバーラップの幅：0.5
- 判別式を検証するために用いた企業（予測データ）
  - 企業数：$g_2 = 238 - g_1$ ($g_2$ = 35,30,25,20,15,10,5)
- 抽出語数：20語
- 繰り返し回数：10回

各グループにおいて，テキストデータを判別式作成に用いる企業数を10社ずつ増やして判別式を作成した．抽出語数は3.4.1小節の結果より20語に設定した．実験結果を図5.4，表5.9に示す.

表 5.9: The number of company and discrimination rate

| $g_1/g_2$ | 5/35 | 10/30 | 15/25 | 20/20 | 25/15 | 30/10 | 35/5 |
|---|---|---|---|---|---|---|---|
| Fraud problem (Learning data) | 97.4 | 97.7 | 99.6 | 100 | 99.52 | 100 | 99.457 |
| No fraud problem (Learning data) | 64.8 | 95.4 | 100 | 99.55 | 100 | 100 | 99.971 |
| Fraud problem (Prediction data) | 73.86817227 | 77.1954023 | 79.56 | 79.1 | 84.466 | 80.1 | 87.6 |
| No fraud problem (Prediction data) | 53.057 | 61 | 63.92 | 68.35 | 70.4 | 69.2 | 71.2 |

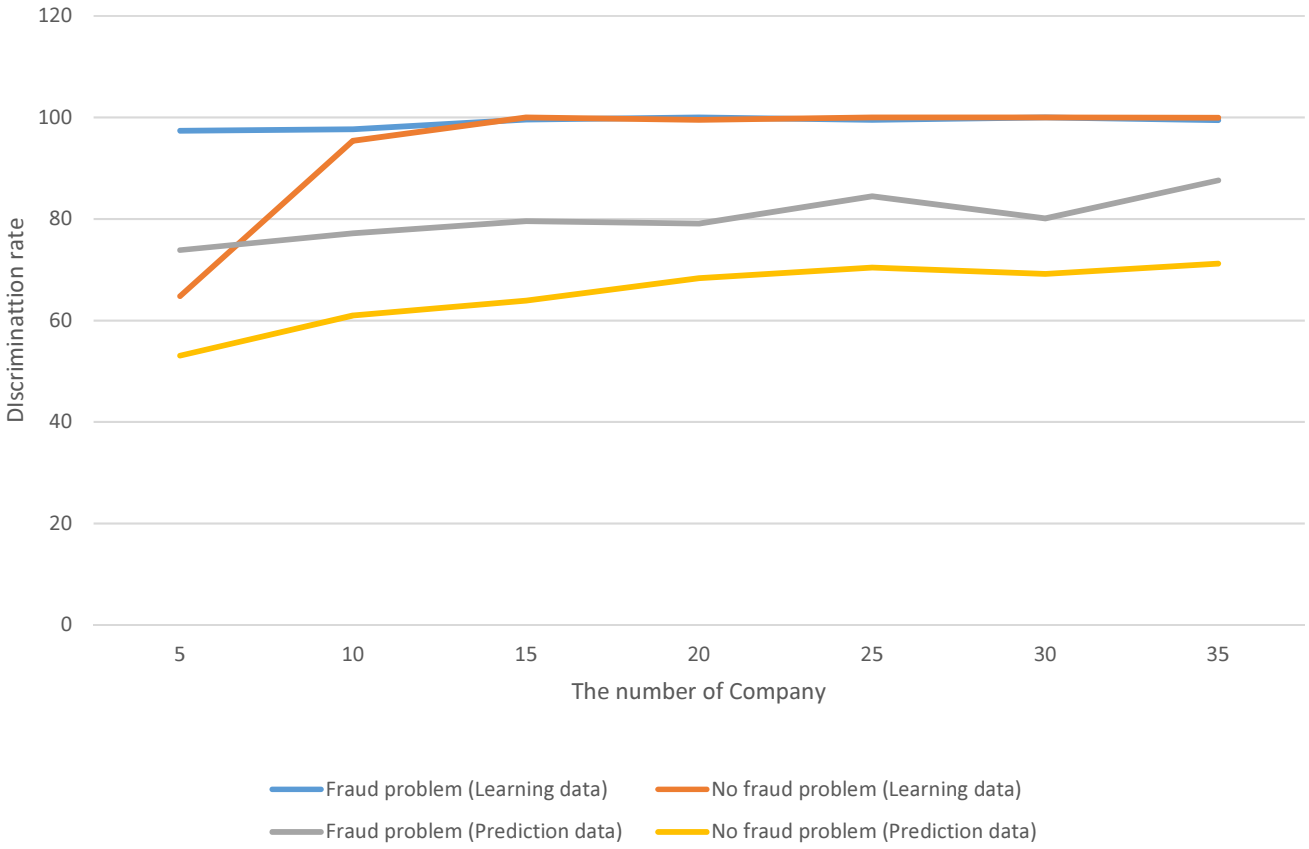



図 5.4: The number of company and discrimination rate

図の縦軸は判別率，横軸は判別式を作成するために用いる企業の数を表している．学習データにおいて学習データの企業数が10社を超えると両グループともに常に判別率が90% 以上であり，100% に収束している．予測データにおいて，判別式作成に用いる企業数が増えるにつれて判別率が上昇している．このことから学習データ量が増えるとさらに精度の良い判別式作成が可能であると考えられる．また本計算機実験においても予測データにおいて，不正問題発覚ありグループよりも不正問題発覚なしグループの方が判別率が低い結果となった．これは5.5.2で考察した理由と同じであると考えられる．

### 5.5.3 ロジスティック曲線を用いた不正問題発生確率

不正問題は，テキストデータ内に不正問題についての記述がなくても実際には不正問題が発生している企業が存在するという点で正解データに不確実性を含む問題である．そこで正解データを用いた判別率の評価のみでは，提案手法の有効性の検証が不十分であると考えられる．判別スコアをロジスティック曲線を用いて，不正問題発生の確率に変換する．キストデータより得られた実験条件を以下に示す.

表 5.10: Experiment Condition

| | 不正問題発覚あり | 不正問題発覚なし |
|---|---|---|
| 作成用データ | 20 | 70 |
| 予測用データ | 20 | 20 |

抽出語数100

オーバーラップの幅 0.5

試行回数 5

判別率の平均を以下に示す．判別率の平均は作成用データについて，不正問題あり

表 5.11: Experiment Condition

| | Fraud problem | No fraud problem |
|---|---|---|
| Learning data | 99.6 | 95.4 |
| Prediction data | 77.2 | 70.4 |

グループで99.6%，不正問題なしグループで95.4%，予測用データについて，不正問題ありグループで77.2%，不正問題なしグループで70.4%であった．

5試行のうち最良解となったときの予測用データにおける判別スコアを図5.5に示す．縦

軸が判別スコア，横軸が企業を表していて，1〜18が不正問題ありグループ19〜34が不正問題なしグループを表す．また，判別コアが判別境界より高いと不正問題ありと判別され，判別境界の値より低いときに不正問題なしグループと判別される．

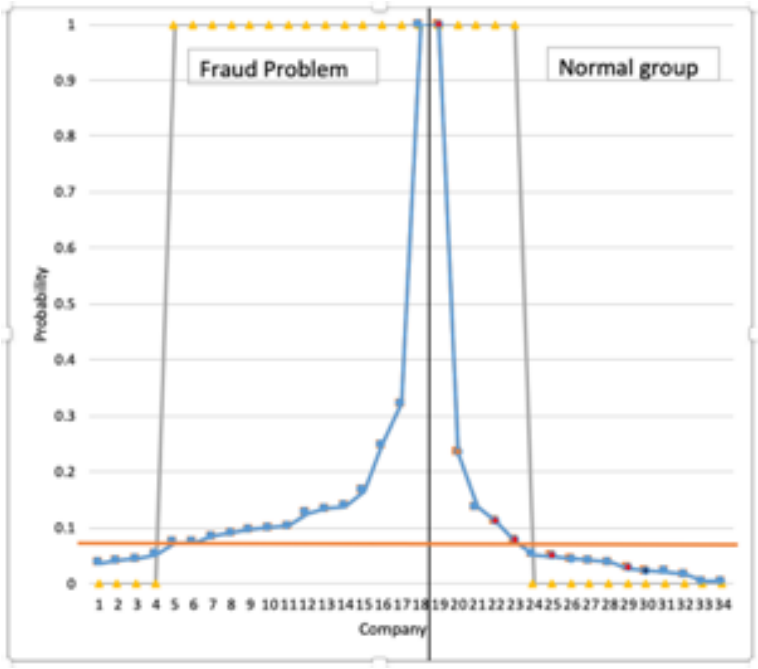



図 5.5: Discriminant Score

この判別結果に対して，ロジスティック曲線を用いて確率に変換した結果を図5.6に示す．縦軸がロジスティック曲線により算出された確率，横軸が企業番号を示す．事前確率が1%，5%，10%，15%，20%，50%，75%，90%の時の確率の結果を表している．事前確率が高くなると曲線は膨らみ，全企業において発生確率が上昇する．本研究において，企業番号が21〜40のテキストデータに不正問題発生の記述がなかった企業の中で，不正問題発生確率が高い企業に注目するべきであると示すことによりコンサルタント支援が可能となる．

## 5.6 コンサルタントの予測結果との比較

不正問題は，テキストデータ内に不正問題についての記述がなくても実際には不正問題が発生している企業が存在するという点で正解データに不確実性を含む問題である．そこで正解データを用いた判別率の評価のみでは，提案手法の有効性の検証が不十分であると考えられる．そこで，テキストデータ内に不正問題の記述はなかったが，計算機実験において不正問題発生ありと判別された企業について，コンサルタントによって不正問題発生の可能性の検討及びクライアント企業へのヒアリングをおこなっ

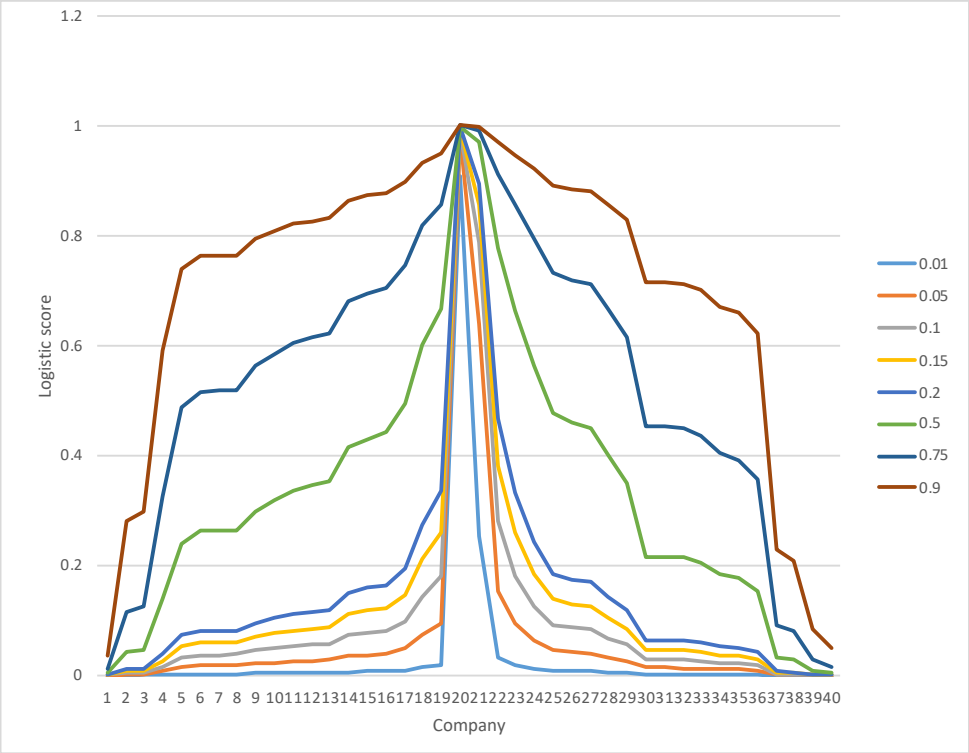



図 5.6: Probability of fraud problem

た．コンサルタントの予測結果及びコンサルタントによるクライアント企業へのヒアリング結果と提案手法の予測結果を比較した．

### 5.6.1 要因とされる抽出語句の比較

コンサルタントによる不正問題発生の可能性において，発生する可能性があると判断した理由についての記述と計算機実験において抽出された語句を比較する．表5.12はコンサルタントの不正問題発生可能性の判断理由を表す．表5.13は計算機実験において，不正問題発生に要因のある語として抽出された上位10語を表す．下線部の語句である”Profit””Part time job” がどちらにも出現している．コンサルタントが不正問題発生の要因として考える特徴を計算機においても抽出できていることが示唆された．

### 5.6.2 ロジスティック曲線を用いた予測結果の比較

ロジスティック曲線を用いて算出した不正問題発生確率とコンサルタントの予測及びクライアント企業へのヒアリング結果を比較する．図5.7に比較結果を示す．縦軸が計算機実験委における不正問題発生確率，横軸が企業ラベルを表す．縦線がグループの境界をあらわしていて，不正問題発生ありの企業が企業ラベル1〜18，不正問題なしの企業が企業ラベル18〜35までである．横軸は班別境界を表していて，赤いエリアが

表 5.12: Reason for the consultant's judgment

| |
|---|
| No fraud has been reported yet. Cash management and profit management are poor, so you should not notice even if they are fraudulent. |
| In a family business, accounting is for relatives only, and since there is a human relationship, it feels that there is no injustice, but I think there is an environment that allows injustice. |
| Although each employee works separately, there is no management system and employee education is only on-the-job training. |
| Taxi business handles cash and may be fraudulent |
| It seems that the risk is high because there are many foreign (Part time job) employment |

表 5.13: Extracted word

| |
|---|
| Paid |
| Part time job |
| Allowance |
| Number |
| Main rule |
| Trial calculation |
| Profit |
| Postcript |
| Settlement |
| Applicable |

判別が正しいエリア，青と黄色のエリアが後判別のエリアをあらわすが，不確実性を含む問題であることを考慮した場合不正問題発生無し企業の中にも実際は不正問題が発生している企業が存在するので，黄色のエリアに属する企業がコンサルタントが注意すべき企業となる．各企業の赤丸がコンサルタントが不正問題の発生の可能性があると判断した企業，赤色のひし形がクライアントへのヒアリングの結果不正問題発生が発覚した企業，青色のひし形がヒアリングの結果からも不正問題が発覚しなかった企業をあらわす．

図5.7より各企業が不正問題ありなしの2値分類ではなく，発生の確率で表されていることが確認できる．不正問題なしのグループにおいて，計算機で100% 不正問題ありと判別された企業について，コンサルタントも不正問題発生の可能性があると判断した．また，不正問題なしのグループにおいて計算機が不正問題ありと判別した企業の中で，ヒアリングの結果実際に不正問題発生が発覚した企業が存在した．一方，計算機が不正問題発覚なしと判断した企業についてはヒアリングの結果からも不正問題は

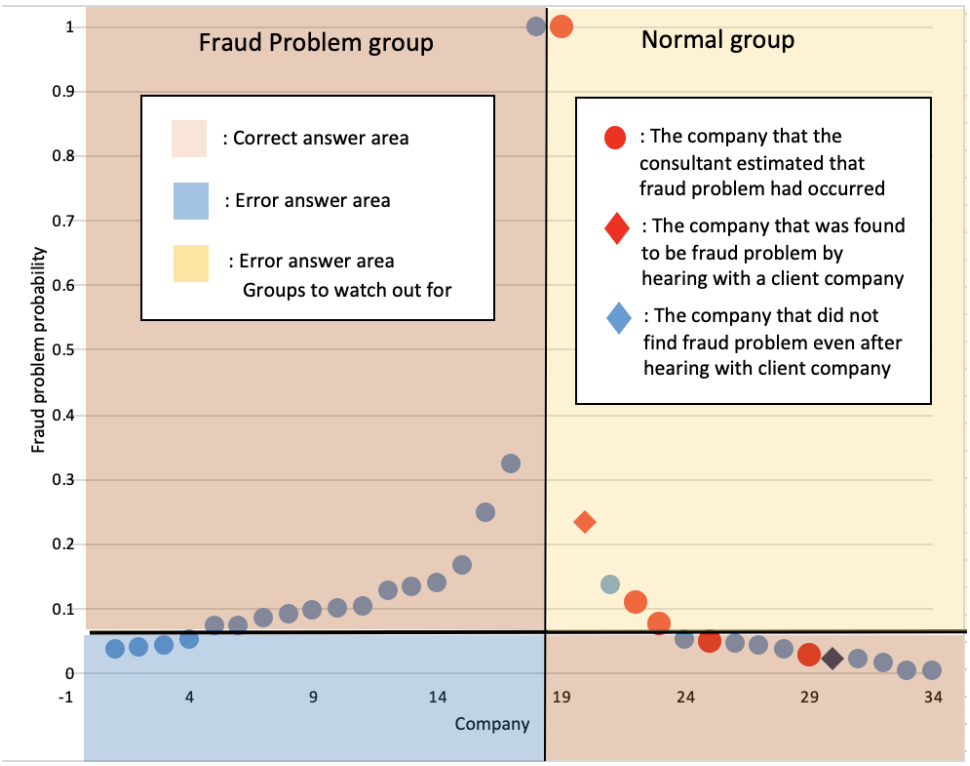



図 5.7: Probability of fraud problem

発覚しなかった．コンサルタントに相談していずテキストデータ内に不正問題の記述がない企業の中にも，実際は不正問題が発生している企業やコンサルタントが不正問題発生の可能性があると判断する企業は存在した．計算機の不正問題発生確率とヒアリングの結果が一致していることから，提案手法の支援可能性が示唆された．

## 5.7 考察

各計算機実験の結果から，以下の特性が分かった．

- クライアント企業の業種と地域による判別式への影響
  クライアント企業の業種と地域を変化させて判別式を作成することによって，業種や地域の判別式への影響を確認した．業種を絞って作成した判別式は特定の語による判別なので汎用性が低いことがわかった．また，グループ間で業種や地域をそろえて判別式を作成することにより，不正問題の有無以外の要因で判別されることなく，不正問題を判別するための式が作成されていることを確認した．これらの結果より，不正問題についてクライアント企業の業種や地域は判別式に影響することが分かった．

- 要因となる語の抽出
  抽出された語句の中に，不正問題と関連があると推察できる語句が抽出されてい

ることが分かった．これらの抽出された語句から，語句と不正問題の関係を考察することにより，適当なサービスの提案，新しいコンサルティングサービスの創出などの支援につながると考えられる．

- DEA 判別分析による不正問題予測
  抽出語数の変化と判別率の関係を検討するための計算機実験により，学習データにおいて不正問題発覚ありのグループの方が不正問題発覚なしグループより少ない抽出語句で判別できていたことから，不正問題が発覚している企業のテキストデータの中には不正問題の特徴を現す語句が存在していると考えられる．また，企業数の変化と判別率について検討すると，判別式作成のために用いる企業数が多いほど汎用的な判別式が作成できていたことがわかった．両実験において予測データの判別率は，不正問題発覚ありグループよりも不正問題発覚なしグループの方が判別率が低かった．これは，正問題発覚なしのグループの中には発覚はしていないが，実際は不正問題が発生している企業が存在する可能性があるという，本対象問題の不確実性を含む問題構造によるものと考えられる．

- ロジスティック曲線を用いた不正問題発生確率
  不確実性を含む問題構造を検討するため，ロジスティック曲線を用いて不正問題発生確率を算出した．発生確率を提示することにより，コンサルタントが注目すべき企業を特定する支援となる．

- 実現場へのヒアリングと提案手法による判別結果の比較
  不確実性を含む問題構造を検討するため，ロジスティック曲線を用いて算出した不正問題発生確率とコンサルタントの予測及びクライアント企業へのヒアリングをおこなった．比較結果から，コンサルタントに相談していなくても不正問題が発生していた企業は存在した．計算機の不正問題発生確率とヒアリングの結果が一致していることから，提案手法の支援可能性が示唆された．

## 5.8 結言

本章では不確実性を含む問題を対象として，提案手法を用いて実データを解析した．クライアント企業の不正問題を対象とし，5.2節では不確実性を含む問題の特徴と拡張手法について述べた．5.3節で実験条件を述べ，5.4節では，クライアント企業に関する業種や地域が判別率に及ぼす影響について検証した．5.5節で提案手法によって抽出された要因となる語句について考察し，判別式の作成，新たなデータに対して予測を行う．さらに，ロジスティック曲線を用いて不正問題を発生確率で表し，コンサルタントの注目すべき企業を特定する．5.6節では，不確実性を含む正解データに対して検

証するためクライアント企業に実際にヒアリングを行った結果と提案手法による判別を比較した．5.7節では実験結果についての考察を述べた．

# 第6章　結論と今後の展望

本論文ではコンサルティングサービスの支援システム構築を目指し，2種の異なる問題を対象にクライアント企業の状態認識のための，テキストマイニングを用いた問題発生予測手法を提案した．

## 6.1　結論

本論文の結論を各章以下のようにまとめる．

第1章では，本研究の背景と目的を明らかにした．中小企業の活性化が重要視される中で，中小企業が自社内では対処することが困難である問題の解決を支援する，中小企業向けコンサルティングの支援体制の重要性も高まっている．コンサルティング企業は幅広い経営相談に対応することができるが，相談内容は幅広く専門性が高いためサービス提案や問題察知はコンサルタントの経験や勘に依存し，安定したサービスを提供できないという問題がある．よって，コンサルティング企業の専門性に依存しない，安定したサービスを提供するための支援システムが望まれる．本研究では支援システム構築として，クライアント企業の相談内容を記したテキストデータを解析することにより，今後発生する問題を察知するための判別式を作成することを目指す．本研究で用いるテキストマイニングの特徴，関連研究を挙げ，本研究が対象とするテキストデータや対象問題の特徴について述べた．

第2章では,本研究で扱う対象問題の説明と,問題察知に用いる判別式作成の提案手法についての説明を行った.提案手法では対応分析とDEA判別分析を組み合わせることにより，テキストデータから得られる膨大な数の語句から問題察知に関わりの強い語を抽出し，DEA判別分析の変数を削減して解析する．また，本研究では対象問題として解約問題と不正問題について予測する．解約問題は現在コンサルティング契約は更新中か解約か明確に分かるという点で，不確実性を含まない問題である．一方，不正問題は現在発覚していなかったとしても実際には不正問題が発生している可能性があるという点で不確実性を含む問題である．これらの構造の異なる問題について提案手法を適用し，有効性を検証する．

第3章では，解約問題を対象として実規模問題の一部のデータを用いて提案手法の有効性の検証を行った．既存手法であるIF-IDF値，線形判別手法と提案手法を比較することにより提案手法の有効性を検証した．

第4章では，実現場適用に向けて，実規模問題を対象とした解約予測をおこなった．実規模問題の持つ特徴に対応するため，特徴抽出法の拡張，データ拡張，標準化の導入，判別分析の目的関数の拡張をおこない有効性を検証した．さらに，コンサルタントの予測結果と比較し，本支援システムの実現場適用の可能性を検討した．

第5章では，不確実性を含む問題を対象として，問題発生を予測した．不確実性を考慮するためロジスティック曲線を用いて，発生確率を算出した．さらに，コンサルタントの予測結果・実現場でのヒアリング結果と比較することにより，支援手法の実現場適用への可能性を検討した．　以上より，本論文の成果は以下のようにまとめられる．

- コンサルティングサービスの支援システム構築を目指し，テキストデータから2種の異なる問題構造に対する問題発生を予測する手法を提案した．

- 特徴抽出により，コンサルタントが見て対象問題と関係があると分かる語が含まれており，対象グループの要因となる語が実際に抽出されていることを確認した．さらにコンサルタントが気づくことのできなかった語句についても，計算機実験から得られた抽出語をもとに議論することによって，新たな関連性を推測することできた．

- 抽出語数の変化と判別率の関係を検討するための計算機実験により，判別に用いる語句数は少なすぎると判別できないが，増やしすぎてもノイズとなる語句が抽出され判別率が上昇しないことを確認した．また，企業数の変化と判別率について検討すると，判別式作成のために用いる企業数が多いほど汎用的な判別式が作成できていたことがわかった．また，判別率は解約グループに比べ，継続グループの方が判別率が低いことを示した．

- 提案手法の予測結果とコンサルタントの予測結果との比較により，コンサルタントが問題発生予測をおこなう際に捉えいてる特徴を提案手法においても考慮できていることが示唆された．また，コンサルタントが自身のない予測50% 付近の企業について，提案手法による解約予測支援が可能であることが示唆された．

- 実現場適用に向けて手法を拡張することにより，グループ間に企業数の差があり語句の重なりが少ない実規模問題に対して，75%以上判別することを確認した．

- 不確実性を含む問題に対して，ロジスティック曲線を用いることにより発生確率を算出した．発生確率を提示することにより，コンサルタントが注目すべき企業を特定する支援となることが示唆された．

- 不確実性を含む問題構造を検討するため，ロジスティック曲線を用いて算出した不正問題発生確率とコンサルタントの予測及びクライアント企業へのヒアリン

グをおこなった．比較結果から，コンサルタントに相談していなくても不正問題が発生していた企業は存在した．計算機の不正問題発生確率とヒアリングの結果が一致していることから，提案手法の支援可能性が示唆された．

本研究では，コンサルタントのクライアント企業における状態認識を支援するため異なる 2 種の問題構造に着目し，確実な問題として解約問題を，不確実性を含む問題として不正問題を対象として支援手法を提案し有効性を検証した．以上の結果から，コンサルタントがクライアント企業の抱える問題で察知しなければならない多様な問題について，支援できることが示唆される．テキストマイニングを用いたコンサルティングサービスのための支援システム構築により，抽出された語の解析を利用した問題発生の要因発見や，コンサルタントが支援すべき企業の特定が可能となったことが示唆された．このシステム構築により新サービスの提案，質の安定したコンサルティングサービスの提供，新人の早期育成などが期待される．

## 6.2 今後の展望

今後の課題として以下の項目が挙げられる．

- 自然言語処理の高度化
  本研究ではテキストデータを単語に分割し，出現回数の値を用いて解析している．しかし，例えば“残業”という語が出現していたとしてもその語に続く語が“多い”と“ない”では全く意味が異なる．このように同じ単語が出現したとしても前後の語句の関係によって表す意味が異なってしまう場合がある．これを解析するために自然言語処理の意味解析や構造解析の手法を用いる検討をする必要がある．

- 定量的データの使用
  本研究では判別式の変数にキストデータから得られた語句の出現回数のみを用いている．しかし例えば企業の従業員数や売上高など，対象問題に関連のある定量的データが存在する．これらのデータも判別式の変数に加えることにより判別の精度が上がると考えられる．

- 機械学習の適用
  本研究では，作成した判別式や判別係数から意味を考察し，コンサルタントの理解やクライアント企業の説明に用いることができるようにするため線形の判別式を用いた．しかし，予測データにおいて 100% の判別はできていないため非線形な判別手法についても検討する必要がある．

- 支援手法の拡張

  本稿では，コンサルタントのクライアント企業に関する問題察知に関する支援手法を提案した．これによりクライアント企業の状態認識については支援可能であるが，状態認識後におけるコンサルタントの行動判断についての支援ついても検討する必要がある．

# 謝辞

本論文の執筆にあたり，多くの方々にご指導ご鞭撻を賜りました．

著者の指導教員である藤井信忠准教授に心より厚く御礼を申し上げます．研究のみならず，日常生活におきましても多くのご指導・ご鞭撻を賜りました．国内外での講演など数多くの挑戦する機会を与えて頂き，困難に直面することもありましたが，常に意義を感じながら研究に取り組むことができました．また，様々な分野で活躍する方々との交流の場を多く設けて頂き，広い視野を持ちながら研究に取り組むことができました．

学部4年次に貝原研究室に配属されて以来，4年半に渡りご指導・ご鞭撻を賜りました貝原俊也教授に心より厚く御礼を申し上げます．研究を進める上で困難に直面した際には，時には厳しく時には諭すように軌道修正をして頂きました．

常に丁寧にご指導・ご鞭撻を賜りました國領大介助教に心より厚く御礼を申し上げます．研究者として，指導者としての姿勢を学びました．

お忙しい中副査を引き受けて下さり，論文全体に渡って貴重なご助言，ご指導を多々頂きました，大川剛直教授，鳩野逸生教授に厚く御礼を申し上げます．

共同研究を通じて数々の貴重な御意見を受け賜りました，(株)エフアンドエムの安部洋一氏，公共財団法人新産業創造研究機構の山東良子女史に深く感謝いたします．

研究室に配属されてからの4年半の間に出会い，私を支えて頂き，暖かいご支援・励ましのお言葉を頂きました皆様に感謝致します．特に，貝原研究室の卒業・修了生及び在学生の皆様との充実した研究室生活はかけがえのない思い出です．

システム工学に関する興味を抱くようになったきっかけ及び，後期課程への進学を決心するきっかけを頂きました皆様に感謝の意を表します．

最後に，私が後期課程へ進学することを快く許して頂き，これまで私を育てて頂いた両親と暖かく見守って頂いた妹に心より感謝致します．

2020年7月

渡邉　るりこ

# 参考文献


[1] 経済産業省. 中小企業憲章について. http://www.meti.go.jp/committee/summary/0004655/kensho.html.

[2] 中小企業庁. 認定経営革新等支援機関. http://www.chusho.meti.go.jp/keiei/kakushin/nintei/.

[3] 中小企業庁. 認定経営革新等支援機関. http://www.chusho.meti.go.jp/keiei/kakushin/nintei/.

[4] 本村陽一, 竹中毅, 石垣司. サービス工学の技術 -ビッグデータの活用と実践-. 東京電機大学出版局, 2012.

[5] Feng Duan, Masahiro Morioka, Jeffery Too Chuan Tan, and Tamio Arai. Multi-mode assembly-support system for cekk production. *International Journal of Automation Technology*, Vol. 2, pp. 384–389, 2008.

[6] Toshiro Kamma, Takafumi Saito, and Shogo Abe. Analysis and adaptation for exaggeration types of animation motion. *Graphic Science*, Vol. 47, pp. 13–23, 2013.

[7] 林田英雄, 脇森浩志. テキストマイニング技術とその応用. *UNISYS TECHNOLOGY REVIEW*, Vol. 84, pp. 29–44, 2005.

[8] MartiA. Hearst. Untangling text data mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 3–10, 1999.

[9] 喜田昌樹. 『テキストマイニング入門—経営研究での活用法』. 白桃書房, 2008.

[10] Hoang T. P. Thanh and Phayung Messad. Stock market trend prediction based on text mining of corporate web and time siries data. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 18, pp. 22–31, 2014.

[11] Junji Yano and Kenji Araki. Performancs evaluation for a method of generating bussiness reports from call center speech dialogues using inductive learning. *Imformation Proccessing Society of Japan*, 2007.

[12] Haruhiko Takase, Hiroharu Kawanaka, and Shinji Tsuruoka. Supporting system for quiz in large class -automatic keyword extraction and browsing interface-. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 18, pp. 152–157, 2015.

[13] Manabu Nii, Kazunobu Takahama, and Shota Miyake. Rule representation for nursing-care process evaluation using decision tree techniques. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 18, pp. 918–925, 2015.

[14] Hajime Murai and Akifumi Tokosumi. Network analysis of the four gospels and the catechism of the catholic church. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 11, pp. 772–779, 2007.

[15] 増田正. 地方議会の会議録に関するテキストマイニング分析 高崎市議会を事例として. 地域政策研究, Vol. 15, pp. 17–31, 2011.

[16] 深澤克明, 沢澄千恵子. 中世古典文学におけるテキストマイニングの試み. 情報処理学会誌, Vol. 26, pp. 152–157, 2016.

[17] 長屋秀幸, 黒岩公考, 井手尾俊宏, 杉本学. テキストマイニングを利用した化学情報抽出ツールの開発. ケモインフォマティクス討論会予稿集, p. 26, 2014.

[18] Hiroshi Wakamori. Text mining techniques for analyzing big data. *UNISYS Technology Review*, Vol. 115, , 2013.

[19] Hossein Hassania, Christina Beneki andStephan Unger, Maedeh Taj Mazinani, and Mohammad Reza Yeganegi. Text mining in big data analytics. *Big data and Cognitive Computing*, 2020.

[20] 大森寛文. 企業・組織における知識発見の実践手法に関する研究～テキストマイニングと知識の構造化論の融合による知識の発見. 埼玉大学大学院　経済科学研究科　博士論文, 2014.

[21] 市村由美, 長谷川隆明, 渡部勇, 佐藤光弘. テキストマイニング-事例紹介-. 人工知能学会誌, Vol. 16-2, pp. 192–200, 2001.

[22] Michael John Jones Mark Clatworthy. Financial reporting of good news and bad news: evidence from accounting narratives. *Accounting and business research*, 2006.

[23] Murray Z. Frank Werner Antweiler. All that talk just noise? the information content of internet stock message boards. *The American Finance Association*, 2004.

[24] 荒川俊博, 若山邦紘, 中村洋一. テキストデータを利用したDEA判別分析による倒産予測. 法政大学修士論文, 2010.

[25] 和泉潔 後藤卓松井藤五郎. テキスト分析による金融取引の実評価. 人工知能学会論文誌, Vol. 26, pp. 313–317, 2011.

[26] 小室達章. テキストマイニングを活用したリスク概念の分析. 金城学院大学論集 社会科学編, Vol. 12, , 2016.

[27] B. Chu-Caroll, J.andCarpenter. Vector based natural language call routing. *Computational Linguistics*, Vol. 25, pp. 361–388, 1999.

[28] Harp S. Tan P., Blau HJ. and Goldman R. Textual data mining of service center call records. *in Proc. of KDD-2000*, pp. 417–423, 2000.

[29] 三末和男岡本青史, 西野文人. カスタマーセンター支援システム. 人工知能学会誌, Vol. 15, pp. 1027–1034, 2000.

[30] 長野那須川. Takmi: Text analysis and knowledge mining-知識発見のためのテキストマイニング技術-. 情報処理学会第59回全国大会, Vol. 4N-06, , 1999.

[31] 中山市村, 赤羽, 三好, 関口, 藤原. 日報分析システムの開発. 電子情報通信学会技術研究報告, Vol. NLC2000-26, pp. 31–38, 2000.

[32] 堀口真宏. 災害支援者sctにおけるイメージの検討 : テキストマイニング分析を用いて. 東洋学園大学紀要, Vol. 28, pp. 1–20, 2020.

[33] 大山泰史, 青柳領, 八坂昭仁, 田方真哉. テキストマイニングを用いたバスケットボールクリニックに参加した選手の意識についての事例研究. 日本体育学会大会予稿集, Vol. 70, p. 269, 2019.

[34] Ratchakoon Oruengkarn, Kok Wai Wong, and Chun Che FUng. A review of data mining techniques and applications. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 21, pp. 31–48, 2017.

[35] 原聡. 機械学習における解釈性. 人工知能学会誌, Vol. 33, pp. 366–369, 2018.

[36] Matt Turek. Explainable artificial intelligence (xai).

[37] 原聡. 説明可能ai. 人工知能学会誌, Vol. 34, pp. 577–581, 2019.

[38] 渡部勇. テキストマイニングの技術と応用. 情報の化学と技術, Vol. 53, pp. 28–33, 2003.

[39] 石田基広, 小林雄一郎. Rで学ぶ日本語テキストマイニング. ひつじ書房, 2013.

[40] Mecab: Yet another part-of-speech and morphological analyzer mecab (和布蕪) とは. *http://mecab.sourceforge.net/*.

[41] Michael Greenacre. Correspondence analysis in practice. *CRC Press*, 2017.

[42] 石川慎一郎, 前田忠彦, 山崎誠. 言語研究のための統計入門. くろしお出版, 2010.

[43] F. E, Grubbs. Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, Vol. 21, pp. 27–58, 1950.

[44] 古橋武. 多変量解析の基礎iii（判別分析（改訂版））-理論とrによる演習-. 電子書籍, 2017.

[45] 末吉俊幸, 多賀谷英明, 渡辺伸輔. 相対効率分析法と目標計画法から見た3種の判別分析法の比較・考察:日本企業の格付け評価への応用. 日本オペレーションズ・リサーチ学会, pp. 122–123, 1997.

[46] 末吉俊幸. DEA -経営効率分析法-. 朝倉書店, 2001.

[47] 森平爽一郎. 信用リスクの測定と管理. 中央経済社, 2011.

[48] 国立研究開発法人情報通信研究機構（NICT）. 日本語wordnet. http://compling.hss.ntu.edu.sg/wnja/.

[49] 金明哲. テキストデータの統計科学入門. 岩波書店, 2009.

[50] 日本公認不正検査士協会. 横領等の社内不正発生状況に関する調査結果報告書. 岩波書店, 2011.

[51] 東京商工リサーチ. Tsr業種コードブック. 2015.

# 本論文の構成

本論文の第3章は以下の論文からなる.

- Ruriko Watanabe, Nobutada Fujii, Daisuke Kokuryo, Toshiya Kaihara, Yoichi Abe, Ryoko Santo: A Study on Support Method for Consulting Service using Text Mining, International Journal of Automation Technology (IJAT), Vol.12 No.4, 482-491, 2018.
- 渡邉るりこ，藤井信忠，國領大介，貝原俊也，安部洋一，山東良子：業務履歴の解析によるコンサルティングサービスの支援手法-抽出手法の比較検討-，日本機械学会第27回設計工学・システム部門講演会，山口海峡メッセ下関，2017.

本論文の第4章は以下の論文からなる.

- Ruriko Watanabe, Nobutada Fujii, Daisuke Kokuryo, Toshiya Kaihara, Yoichi Abe: A study on support method of consulting service using text mining-Application to real problem-, Special issue in Acta Electrotechnica et Informatica from Informatics'19(出版予定)
- Ruriko Watanabe, Nobutada Fujii, Daisuke Kokuryo, Toshiya Kaihara, Yoshinobu Onishi, Yoichi Abe, Ryoko Santo: A study on supporting method by combining correspondence analysis and DEA discriminant analysis-Application to real-scale problem-, ICServ International Conference on Serviceology, Asia University Taichungm, Taiwan, pp.13-15, Nov 2018.
- Ruriko Watanabe, Nobutada Fujii, Daisuke Kokuryo, Toshiya Kaihara, Yoichi Abe: A study on support method of consulting service usingcustomer information -Application to real problem-, IEEE 15th International Scientific Conference on Informatics, Poprad, Slovakia, November2019
- 渡邉るりこ，藤井信忠，國領大介，貝原俊也，安部洋一，山東良子：テキストマイニングを用いたコンサルティングサービスの支援手法（第2報)-実規模問題への適用-, サービス学会第6回国内大会, 明治大学, 2018.
- 渡邉るりこ，藤井信忠，國領大介，貝原俊也，安部洋一，山東良子：業務履歴の解析によるコンサルティングサービスの支援手法（第2報)-類義語辞書を用いたデー

タ拡張-, 日本機械学会第27回設計工学・システム部門講演会，東北大学片平キャンパス，2019.

- 渡邉るりこ，藤井信忠，國領大介，貝原俊也，安部洋一，山東良子：企業内テキストの解析によるコンサルタント支援手法- 類義語辞書の活用と実現場への適用-, 第64回システム制御情報学会研究発表講演会，WEB開催，2020.

本論文の第5章は以下の論文からなる．

- Ruriko Watanabe, Nobutada Fujii, Daisuke Kokuryo, Toshiya Kaihara, Yoichi Abe: A Support Method for Client Companies' State Recognition of Consulting Services Using Text Mining, International Journal of Automation Technology (投稿中)

- Ruriko Watanabe, Nobutada Fujii, Daisuke Kokuryo, Toshiya Kaihara, Yoshinobu Onishi, Yoichi Abe, Ryoko Santo: A Study on Supporting Method for Consulting Service using Text Mining, 11th CIRP Conference on Intelligent Computation in Manufactuaring Engineering, Gulf of Naples, Italy, 19-21 July, 2017.

- 渡邉るりこ，藤井信忠，國領大介，貝原俊也，大西由信，安部洋一，山東良子：テキストマイニングを用いたコンサルティングサービスの支援手法-対応分析とDEA判別分析による不正予測-，第59回自動制御連合講演会国内大会，北九州国際会議場，2016.

- 渡邉るりこ，藤井信忠，國領大介，貝原俊也，大西由信，安部洋一，山東良子：テキストマイニングを用いたコンサルティングサービスの支援手法-不正予測の実サービスにおける検証-，サービス学会第5回国内大会，広島県情報プラザ，2017.

神戸大学博士論文「テキストマイニングを用いた
コンサルティングサービス支援手法に関する研究」全 102 頁

提 出 日 2020 年 7 月 17 日