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 ー対応分析とDEA判別分析による不正予測ー

○渡邉るりこ　藤井信忠　国領大介　貝原俊也　(神戸大学)
大西由訓　安部洋一 ((株) エフアンドエム)　山東良子 ((公財) 新産業創造研究機構)

Support Method of Consulting Service Using Text Mining -Prediction of Corruption Using Correspondence Analysis and DEA Discriminant Analysis-

∗R. Watanabe, N. Fujii, D. Kokuryo, T. Kaihara (Kobe Univ.) ,
Y. Onishi, Y. Abe (F&M Corp.), R. Snato (The New Industry Research Organization)

Abstract− This study aims to build a supporting method for consulting service companies so that the companies can respond to client's demand regardless of the expertise of the companies. Occurrence of future problems in client companies is predicted by using text mining with data taken from a consulting company; correspondence analysis and DEA discriminant analysis are employed. Computer experiments are conducted to verify the effectiveness of the proposed method.

Key Words: Text mining, Correspondence Analysis, DEA Discriminant Analysis

1　はじめに

近年中小企業の活性化が重要視される [1] 中で，中小企業が自社内で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 [2]．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，専門的な分野の指導は必ずしも得意ではないという問題がある [3]．専門知識の欠如しているコンサルタントであっても専門性の高いサービスを素早く提供するための支援手法が求められている．

本研究ではコンサルティング企業の専門性に依存しないコンサルティングサービスを実現することを目的とする．顧客から受けた相談事項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，クライアント企業における不正問題発生の有無を判別する手法を提案する．

2　研究手法

Fig.1 に研究の概要を示す．まず，テキストデータを不正問題発覚の有無で分類し，テキストマイニングを行うことにより判別式の作成を行う．求めた判別式の有効性を検証するため，新たに分類したテキストデータに対して不正問題発生の有無を判別する．

図 2 に研究手法の流れを示す．

2.1　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．テキストデータは不正問題発覚の時系列を考慮しながら，不正問題発生の有無で分類する．不正問題発覚の時系列とグループ分類の関係を Fig.3，4 に示す．コンサルティング中に不正問題が発覚した場合，不正問題発覚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分析対象から除外する．これは本研究が不正

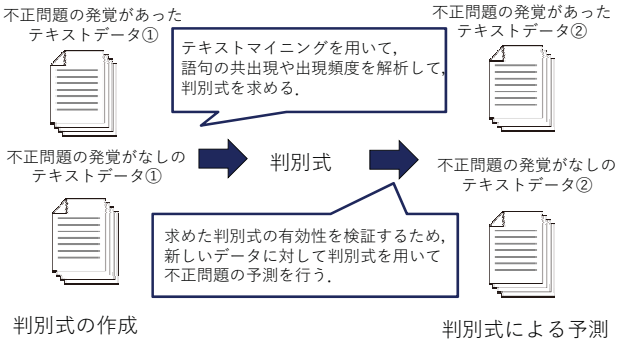

Fig. 1: Overview

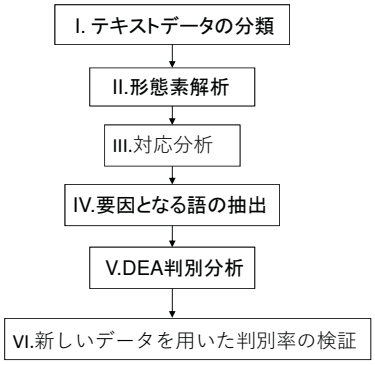

Fig. 2: Algorithm

問題発生の予測を目的とするためである．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，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．

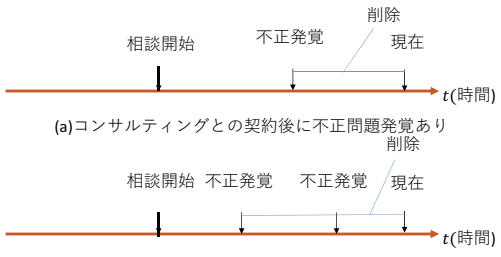

Fig. 3: Fraud problem group

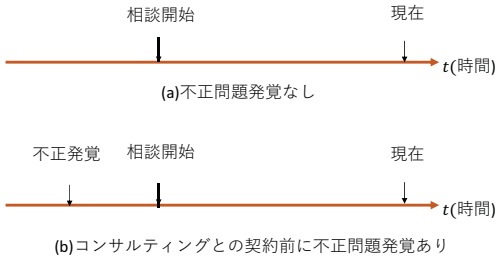

Fig. 4: Normal group

2.2 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．本研究では，工藤拓氏が開発した形態素解析器として公開されている MeCab を使用している [4]．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．

分解した形態素の中から，分析のノイズとなる以下の項目の語句は省略した．

- コンサルティングサービスの名称：(例) 適性診断サービス
- 定型文：(例)FAX 送信済み
- 意味をなさない語：(例) 年，月，件

2.3 対応分析

対応分析はフランスの研究者，Jean-Paul-Ben-zeccri 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 [5]．本研究では，行項目 (サンプル) に語句を列項目 (カテゴリ) に企業名をいれ，企業 i のテキストデータにおける語句 j の出現回数 t_{ij} の二次元データを対象とする．二次元データを Table.5 に示す．

Table 1: Correspondence analysis

	語句	AAA	BBB	$\cdots$	KKK
企業名		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}

サンプル，カテゴリのそれぞれに重みとなる変数（サンプルスコア a_i, カテゴリスコア b_j）を設定し，それらの変数の相関が最大となるように計算を行い，各成分におけるサンプルスコア，カテゴリスコアの値を求める．成分を軸として，その成分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．散布図の原点から離れるほど，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現れるため，原点付近の語が対象グループの特徴を強く表す．

散布図の例を Fig.5 に示す．Fig.5 は不正問題が起

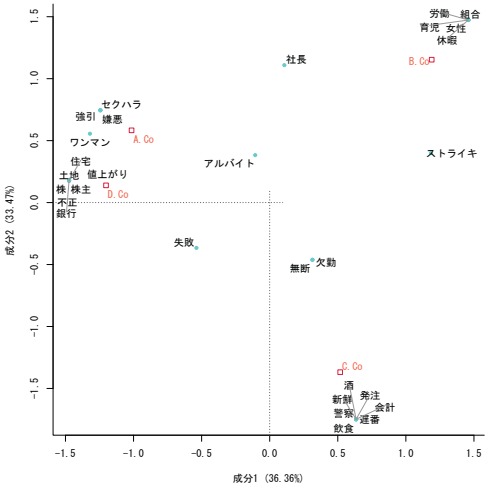

Fig. 5: Correspondence Analysis sample

こった企業四社に対して対応分析を行い，寄与率の高い成分上位二つを軸とした散布図である．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する企業には“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．

そこで本研究では不正問題発覚の有無で分けたそれぞれのグループのデータに対して対応分析を行い，原点に近い語に注目し次節で示す方法で各グループの要因となる語を抽出する．

2.4 要因となる語の抽出

対応分析の結果から全成分を考慮し，各グループ G_1, G_2 において単語 i と原点の距離 d_{iG} を式 (1) で算出する．ただし，D_G は成分の総数，x_{ijG} は成分 j の単語 i のサンプルスコア，C_{jG} は成分 j の寄与率である．

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

それぞれの単語に関して式 (2)，(3) の操作により，d 値を更新する．ただし M は定数である．

$$\begin{cases} & G_1 < G_2 \text{のとき} \\ e_{iG_1} = d_{iG_1} + M - d_{iG_2} & \\ e_{iG_2} = \infty & \\ & G_2 \geq G_1 \text{のとき} \end{cases} \tag{2}$$

$$\begin{cases} e_{iG_2} &= d_{iG_2} + M - d_{iG_1} \\ e_{iG_1} &= \infty \end{cases} \quad (3)$$

各グループについて e 値の小さい語句から一定数を抽出して，各グループの要因となる語として DEA 判分析の変数に用いる．

2.5 DEA 判別分析

DEA 判別分析法は末吉ら [6] 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，グループ間に文章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．

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．ただし決定変数が，グループ G,，企業を特徴付ける要因 i，企業 j，企業 j のテキストデータにおける要因 i の出現回数 z_{ij}，オーバーラップの幅 η であり，従属変数が判別係数 λ_i，判別境界 d，スラック変数 S_{ij}^+, S_{ij}^- である．

stage1

$$\begin{aligned} \min \quad & \textstyle\sum_{j\in G_1} S_{1j}^+ + \sum_{j\in G_2} S_{2j}^- \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^k \lambda_i z_{ij} + S_{1j}^+ - S_{1j}^- = d + \eta \ \ (j \in G_1) \\ & \textstyle\sum_{i=1}^k (\lambda_i^+ - \lambda_i z_{ij} + S_{2j}^+ - S_{2j}^- = d \ \ (j \in G_2) \\ & \textstyle\sum_{i=1}^k |\lambda_i| = 1 \\ & S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \geq 0 \end{aligned} \quad (4)$$

式 (4) 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 η 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．stage1 で得られた最適解を λ_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．

$$R_1 = \left\{ j \in G \middle| \sum_{i=1}^k \lambda_i^* z_{ij} \geq d^* + \eta \right\} \quad (5)$$

$$R_0 = \left\{ j \in G \middle| d^* + 1 > \sum_{i=1}^k \lambda_i^* z_{ij} > d^* \right\} \quad (6)$$

$$R_2 = \left\{ j \in F \middle| d^* \geq \sum_{i=1}^k \lambda_i^* z_{ij} \right\} \quad (7)$$

$$C_1 = \{ j \in R_1 | j \in G_1 \} \quad (8)$$

$$C_2 = \{ j \in R_2 | j \in G_2 \} \quad (9)$$

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．stage2 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．

stage2

$$\begin{aligned} \min \quad & \textstyle\sum_{j\in G_1\cap(R_0\cup R_2)} S_{1j}^+ + \sum_{j\in G_2\cap(R_0\cup R_1)} \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^k \lambda_i z_{ij} \geq d + \eta \ \ (j \in C_1) \\ & \textstyle\sum_{i=1}^k \lambda_i z_{ij} + S_{1j}^+ - S_{1j}^- = c \ \ (j \in G_1 \cap (R_0 \cup R_2)) \\ & \textstyle\sum_{i=1}^k \lambda_i z_{ij} + S_{2j}^+ - S_{2j}^- = c \ \ (j \in G_2 \cap (R_0 \cup R_1)) \\ & \textstyle\sum_{i=1}^k \lambda_i a_{ij} \leq d \ \ (j \in C_2) \\ & \textstyle\sum_{i=1}^k |\lambda_i| = 1 \\ & d \leq c \leq d + \eta \\ & S_{1j}^+, S_{1j}^-, S_{2j}^+, S_{2j}^- \geq 0 \end{aligned} \quad (10)$$

stage1 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，stage1 の結果を保障している．目的関数では stage1 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，stage1 で誤判別されたデータが stage2 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．

stage2 の最適解を c^*, λ_i^* とすると次の基準によって判別される．式 (11) のとき企業 j は G_1 に属するものと判別される．式 (12) のとき企業 j は G_2 に属するものと判別される．

$$\sum_{i=1}^k \lambda_i^* z_{ij} \geq c^* \quad (11)$$

$$\sum_{i=1}^k \lambda_i^* z_{ij} < c^* \quad (12)$$

3 計算機実験

3.1 実験内容

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，判別式作成のため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．

条件 1 業種に偏りを持たせる

1. 製造業のみ
2. 卸売業のみ

条件 2 地域に偏りを持たせる (a 地域のみ)

条件 3 ランダム (5 回)

条件 4 不正問題発覚ありのテキストデータに
業種・地域をそろえる

3.2 実験条件

実験条件についてまとめる．

- テキストデータ
  - 不正問題発覚あり G_1/不正問題発覚なし G_2
  - 総企業数：130(G_1：40/G_2：90)
  - 段落（相談内容件数）：3731/3946
  - 文：11978/14550
  - 総単語数：236880/269980
  - 異なり語数：8592/9074
  - 業種について： 各業種の企業数を Table.2 に示す．

Table 2: Business type

	A. 農業	B. 漁業	C. 鉱業	D. 建設 A
G_1	0	0	0	5
G_2	0	0	18	21
	E. 製造業	F. 電気業	G. 情報通信業	H. 運輸業
G_1	3	0	1	1
G_2	21	0	1	3
	I. 卸売業	J. 金融業	K. 不動産業	L. 学術研究
G_1	18	0	3	0
G_2	24	1	2	5
	M. 宿泊業	N. 娯楽業	O. 教育	P. 医療
G_1	2	1	0	1
G_2	2	1	2	0
	Q. 複合サービス	R. サービス業	S. 公務	合計
G_1	0	4	0	40
G_2	1	8	0	90

  - 地域について：各地域の企業数を Table.3 に示す．

Table 3: Region

地域	a	b	c	d	e	合計
G_1	8	10	7	11	4	40
G_2	27	26	8	20	9	90

- 判別式を作成するために用いた企業 (学習データ)
  - 企業数：40
  - オーバーラップの幅：0.5
  - 抽出語数 20
- 判別式を検証するために用いた企業 (予測データ)
  - 企業数：40

3.3 実験結果

各条件で抽出して作成した判別式の判別率を表 4 に示す．学習データにおいて，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，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．

また，予測データの判別率について，不正問題発覚ありのグループに比べて不正問題発覚なしのグループの判別率が低い．これは，不正問題発覚なしのグループの中に現在までに発覚していないだけで，不正問題が発生している企業を含むためであると考えられる．条件 1 製造業に業種を偏らせたときのみ，予測データにおいて不正問題発覚なしのグループのほうが判別率が良い．これは，条件 1-1 では製造業に特徴的な語が抽出され判別が行われているが，不正問題発覚ありグループにおいて製造業が少ないため判別がうまくいかなかったと考えられる．

Table 4: Condition 1〜3

抽出条件	条件 1-1	条件 1-2	条件 1 平均	条件 2
学習 (%)	97.5	90	93.75	95
予測 G_1(%)	35	75	60	65
予測 G_2(%)	55	30	42.5	30

抽出条件	条件 3-1	条件 3-2	条件 3-3	条件 3-4	条件 3-5	条件 3 平均
学習 (%)	90	87.5	92.5	92.5	95	91.5
予測 G_1(%)	70	55	65	60	55	61
予測 G_2(%)	46	50	40	40	35	42.2

各グループの業種と地域をそろえたときの判別率を表 5 に示す．学習データ・予測データともに結果 1 の条件より判別率が向上した．これは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．また実験 1 と同様に，予測データにおいて不正問題発覚なしのグループより，不正問題発覚ありのグループのほうが判別率が良い結果となった．

Table 5: Condition4

	学習	予測 (G_1)	予測 (G_2)	予測平均
判別率 (%)	97.37	80	40	60

4 結論

本研究では，コンサルティング企業の業務分野は幅広く，専門的な分野の指導は必ずしも得意ではないという現状から，専門知識の有無に左右されずに中小企業で発生する問題に対応，予測できるサービスの必要性を考え，テキストデータの解析によって不正問題発生の判別を目的として，テキストマイニングを用いた手法を提案し，計算機実験を行った．計算機実験では，クライアント企業の業種や地域の偏りが判別にどのような影響を及ぼすか検証した．得られた結果を以下にまとめる．

1. 業種・地域によって文章に特徴があることが分かった．判別式を作成する際は業種や地域の差のノイズが判別分析に影響しないように，各グループの業種と地域の比率をそろえるべきである．

2. 予測データにおいて不正問題発覚ありのグループよりも不正問題発覚なしのグループの判別率が低下していた．これは，不正問題発覚なしのグループの中に不正問題発生企業が存在するという，本研究の不確実性を含む問題構造によるものと考えられる．

参考文献

1) http://www.meti.go.jp/committee/summary/0004655/kensho.html

2) http://www.chusho.meti.go.jp/keiei/kakushin/nintei/
3) http://www.chusho.meti.go.jp/pamflet/hakusyo/H26/h26/
4) http://taku910.github.io/mecab/
5) 石田慎一郎, 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.
6) 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.
7) Kh coder,http://khc.sourceforge.net/.