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―

1515050　山本聖也

発表の流れ

1. はじめに
2. 研究手法
3. 計算機実験
4. 結論

1. はじめに

- 中小企業向けコンサルティングの支援体制の需要が高まってきている.

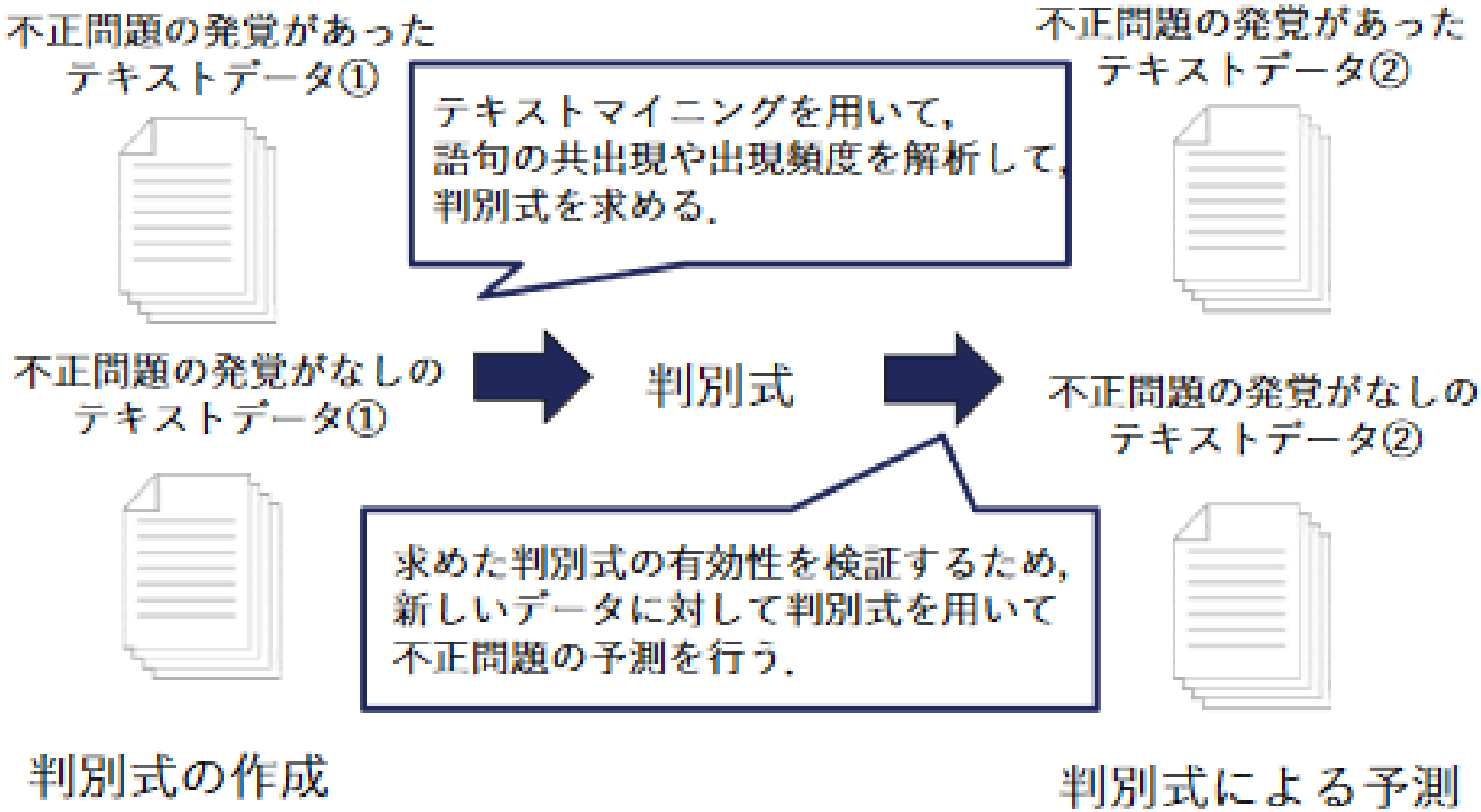
- 専門知識の欠如しているコンサルタントであっても専門性の高いサービスの提供を！

どうやって？

- 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.

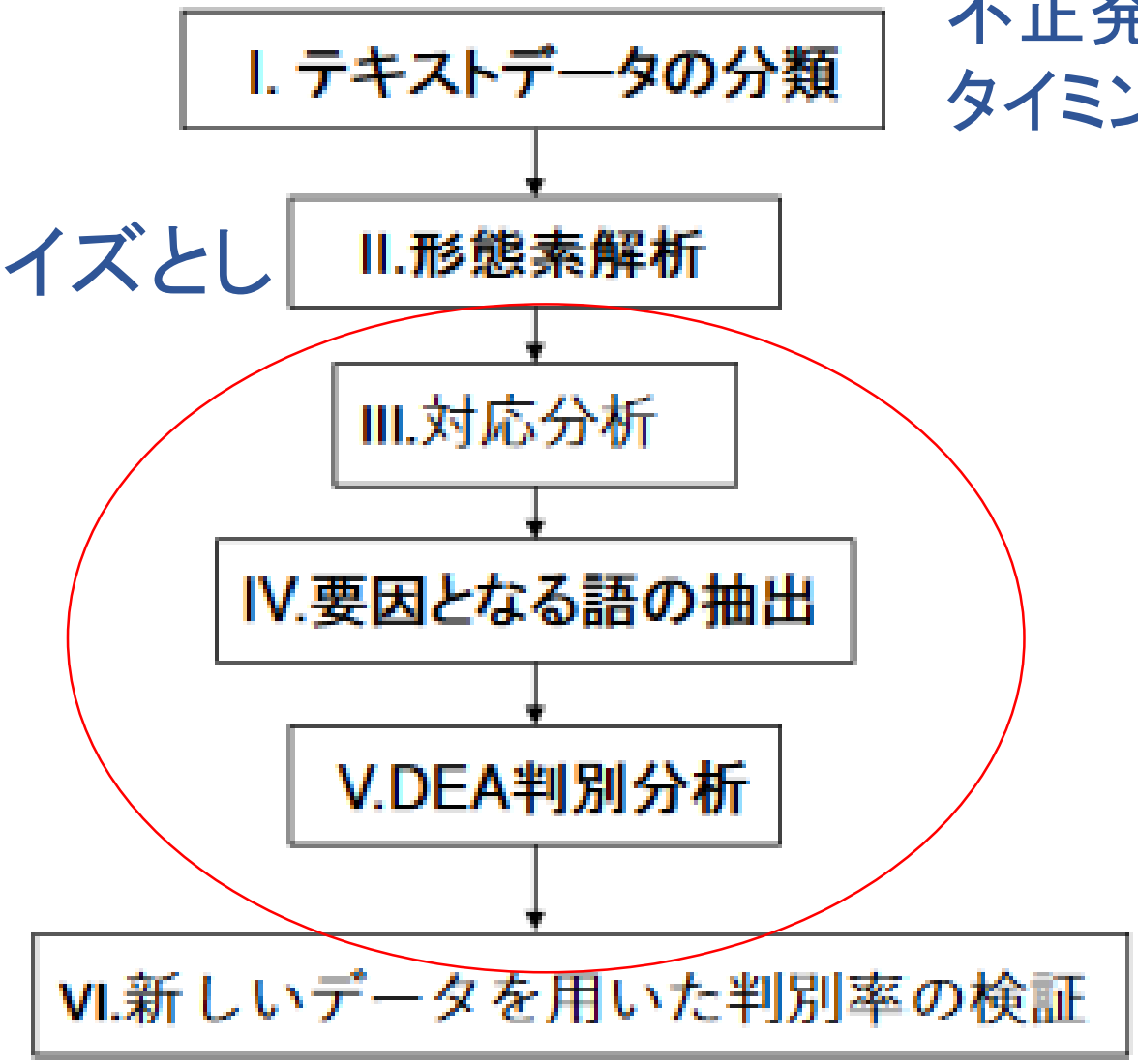
- これによりクライアント企業における不正問題発生の有無を判別する

2. 研究手法


不正問題の発覚があった
テキストデータ①
不正問題の発覚があった
テキストデータ②
テキストマイニングを用いて，
語句の共出現や出現頻度を解析して，
判別式を求める．
不正問題の発覚がなしの
テキストデータ①
判別式
不正問題の発覚がなしの
テキストデータ②
求めた判別式の有効性を検証するため，
新しいデータに対して判別式を用いて
不正問題の予測を行う．
判別式の作成
判別式による予測

不正発覚の有無，
タイミングで分類

不要な語をノイズとして省く



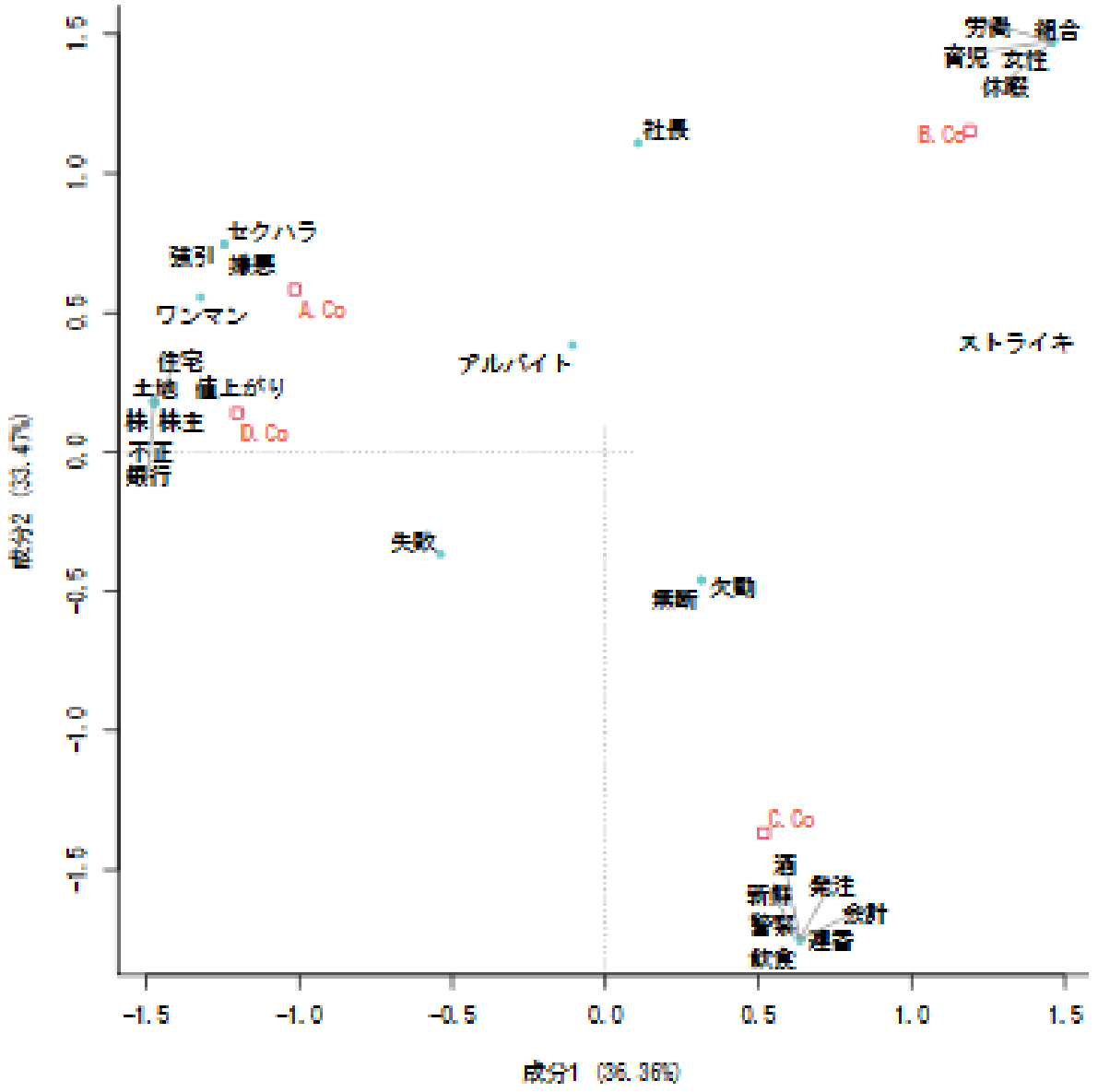
- 対応分析

データ票の行や列に含まれる情報を小数の成分に圧縮する手法

行に語句，列に企業
重みとなる変数をそれぞれ設定し，それらの変数の相関が最大となるように計算を行う

	語句	AAA	BBB	⋯	KKK
企業名		a_1	a_2	⋯	a_K
A Co.	b_1	t_{11}	t_{12}	⋯	t_{1K}
B Co.	b_2	t_{21}	t_{22}	⋯	t_{2K}
⋮	⋮	⋮	⋮	⋱	⋮
N Co.	b_N	t_{N1}	t_{N2}	⋯	t_{NK}

各成分のサンプルスコア，カテゴリスコアの値を求め，散布図に配置する．これにより語句の対応関係が可視化できる



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_1G_2において単語iと原点の距離d_{iG}を式（1）で算出

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

- 続いて(2)，(3)よりdを更新
- 各グループの要因となる語をDEA分析の変数に用いる

$G_1 < G_2$のとき

$$\begin{cases} e_{iG_1} = d_{iG_1} + M - d_{iG_2} \\ e_{iG_2} = \infty \end{cases} \quad (2)$$

$G_2 \geq G_1$のとき

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

DEA判別法

- テキストデータを利用して不正予測を行う手法として採用
- 二段階で行われる判別法
- 一段階目

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

- 二段階目

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判別の制度を高める

- Stage1は誤判別を最小化するようになっており判別境界にηの幅をもたせることでグループ1、グループ2のどちらに半別されるか不明なデータを洗い出すことができる.

stage1

$$\min \quad \sum_{j\in G_1} S_{1j}^{+} + \sum_{j\in G_2} S_{2j}^{-}$$

$$\text{s.t.} \quad \sum_{i=1}^{k} \lambda_i z_{ij} + S_{1j}^{+} - S_{1j}^{-} = d + \eta \quad (j \in G_1)$$

$$\sum_{i=1}^{k} (\lambda_i^{+} - \lambda_i z_{ij} + S_{2j}^{+} - S_{2j}^{-} = d \quad (j \in G_2)$$

$$\sum_{i=1}^{k} |\lambda_i| = 1$$

$$S_{1j}^{+}, S_{1j}^{-}, S_{2j}^{+}, S_{2j}^{-} \geq 0 \qquad ($$

- Stage1で得られた最適解をλi*とd*としたとき，要因iの出現回数Zijは次の判断基準により5種類に分類される
- C1，C2は正しく分類されたこと
- 誤判別の集合

G1∩R2，G2∩R1

R0はオーバーラップ領域にある

データ

$$
\begin{aligned}
R_1 &= \left\{ j \in G \middle| \sum_{i=1}^{k} \lambda_i^* z_{ij} \geq d^* + \eta \right\} \\
R_0 &= \left\{ j \in G \middle| d^* + 1 > \sum_{i=1}^{k} \lambda_i^* z_{ij} > d^* \right\} \\
R_2 &= \left\{ j \in F \middle| d^* \geq \sum_{i=1}^{k} \lambda_i^* z_{ij} \right\} \\
C_1 &= \{ j \in R_1 | j \in G_1 \} \\
C_2 &= \{ j \in R_2 | j \in G_2 \}
\end{aligned}
$$

- stage2では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

・オーバーラップ領域に
存在していたデータの判別
が可能となっている

stage2

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

- Stage2の最適解より，次の基準で判別される

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

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

- 上式のとき企業jはG1
- 下式のとき企業jはG2

3. 計算機実験

- 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

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

実験結果

学習良

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

予測良

- 条件4の場合の結果

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

- 1の条件より判別率が向上
- 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
- １～３と同様にG1の判別率の方が高くなった

4. 結論

- 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
- 不正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る．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こと．