様式 C-19

科学研究費補助金研究成果報告書

平成２２年　４月　７日現在

研究種目：若手研究（Ｂ）
研究期間：　2008　～　2009
課題番号：20700214
研究課題名（和文）　ファジィクラスタリングに基づくテキストデータの分析に関する研究

研究課題名（英文）　Studies on Text Data Analysis Based on Fuzzy Clustering

研究代表者
本多　克宏　（HONDA　KATSUHIRO）
大阪府立大学・工学研究科・准教授
研究者番号：80332964

研究成果の概要(和文):非構造的なテキストデータから有益な情報を抽出することを課題とし，線形ファジィクラスタリング（局所的主成分分析）に基づく標本や変量の分類，視覚化などを通して，分析者が潜在的な相関ルールを直感的に理解することができる分析手法の開発を目的に研究を行った．テキスト－単語マップ作成におけるキーワードの自動選別法や，ノイズ文書を無視しながら類似したテキストからなる群ごとに核となる文書を強調する手法などを開発した．

研究成果の概要（英文）：With the goal being to extract meaningful information from non-structural text data, several techniques for intuitively understanding intrinsic association rules were developed considering sample classification, variable selection and visualization based on linear fuzzy clustering (or local principal component analysis). Useful text-keyword maps are constructed through such techniques as automated keyword selection or core document identification by noise rejection in document grouping processes.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,700,000	510,000	2,210,000
２００９年度	1,600,000	480,000	2,080,000
総　計	3,300,000	990,000	4,290,000

研究分野：知能情報学
科研費の分科・細目：情報学　・　感性情報学・ソフトコンピューティング
キーワード：データマイニング，クラスタリング，パターン認識，ニューラルネットワーク，感性情報処理

１．研究開始当初の背景

大規模なデータベースから将来の意思決定に有益な情報を探し出すデータマイニングでは，重要な要素として「四つのＳ：測定，視覚化，相関，層別」があげられる．研究代表者は，層別と相関の抽出という二つの要素を同時に考慮する手法として，局所的な主成分分析と同一視できる線形ファジィクラスタリングに基づくアプローチに関する研究を行ってきた．近年，インターネットを通じたデータの収集・蓄積の技術が発展・普及するにつれて，テキスト情報の分析技術に対する要請が非常に強まっており，テキスト文書間に内在する関連性の分析技術の開発が重要な課題となっている．

２．研究の目的

インターネット上でやり取りされるデータは，電子メールを代表として，多くがテキスト形式であることから，高度情報化社会においてテキスト情報の処理法が不可欠な技術となっている．本研究では，非構造的なテキストデータから有益な情報を抽出することを課題とし，線形ファジィクラスタリング（局所的な主成分分析）に基づく標本や変量の分類，次元圧縮（視覚化）などを通して，分析者が潜在的な相関ルールを直感的に理解することができる分析手法の開発を目的とした．

３．研究の方法

本研究では，分析者が潜在的な相関ルールを直感的に理解することができる分析手法の開発を目的に，非構造的なテキストデータから有益な情報を抽出するアプローチを探った．そのために，以下の四つの主たるテーマの下で，研究を行った．

(1) 初期設定値の影響を受けない決定論的な手順に基づくクラスター構造の把握を目的に，主成分分析に基づく k-Means クラスタリング法をテキスト分析へ応用した．テキスト文書間の関連性に基づく群構造の把握において，複数のテキスト群と関連性を有していたり，もしくはいずれとも関連性が薄かったりするノイズ文書とみなされるものの影響を排除するために，ノイズファジィクラスタリング機構との融合を行った．

(2) キーワードの自動選別によるテキスト－単語マップの構造明確化を目的に，主成分分析による低次元情報縮約に変量選択の機構を導入した手法を応用した．テキスト－単語マップの作成では，個々のテキスト文書を個体，単語を変量とみなし，単語の出現頻度のデータに主成分分析を施すことで 2 次元平面への情報縮約が行われるが，その際に，群構造の把握に対する各変量の寄与度を考慮し，寄与度の大きい変量（単語）の影響を強調することで，構造を明確化することを試みた．

(3) 個体の特徴ベクトルの代わりに個体間の関連性の強さについての情報のみが与えられる関係性データに内在する相関構造を抽出することを目的に，k-Medoids クラスタリング法を局所的な主成分分析に応用した．web 上のホームページデータのように，非構造なテキストを含む文書データを分析する際には，文書間の相互の類似性をもとにした関連文書の分類が必要となるが，文書間の関連性を直感的にとらえるためには，低次元視覚化が有効である．単一の 2 次元平面で構造を表現することが困難な場合でも直感的理解が容易となるように，線形ファジィクラスタリグに k-Medoids 法におけるクラスタープロトタイプ同定法を導入することで，複数の低次元空間を用いて構造を表現する多クラスター型の多次元尺度構成法の開発を試みた．

(4) 量的尺度を構成しない言語値で与えられるカテゴリー変量を含むデータの分析法を目的に，ファジィクラスタリングにおける最適尺度構成法の開発を試みた．

４．研究成果

前項で示した主たる四つのテーマにおける研究成果は，以下のとおりである．

(1) 主成分分析に基づく k-Means クラスタリング法にノイズファジイクラスタリング機構を導入することで，ノイズの影響を排除したロバストな k-Means クラスタリング法を開発した．提案手法は，以下の特性や意義を有する．

① おのおののデータ点が k-Means クラスタリングに寄与する度合をノイズファジィクラスタリングの枠組みで推定し，寄与度合をファジィメンバシップとみなしてファジィ主成分分析を施すことで，寄与度を考慮した k-Means クラスタリング指標を算出している．

② k-Means クラスタリングへの寄与度とデータ分類についてのクラスタリング指標を繰り返し最適化の枠組みで実装しているが，その手順は決定論的であり，従来の k-Means 型手法における初期分割依存の影響を受けない分析手法となっている．

③ テキスト分析への応用においては，内容の類似した文書群の抽出の際に，群（クラスター）ごとに核を構成する文書（k-Means クラスタリングへの寄与度の大きい文書）を強調することで，群構造をより明確化できることを示した．

④ 国内外の学術会議で発表を行ったほか，IEEE（米国電気電子技術者協会）の論文誌においても論文発表を行った．特に，ファジィシステムに関する IEEE 論文誌（IEEE Transaction on Fuzzy Systems）では，近年，クラスター分析に関する成果が数多く発表されているが，本研究における成果は決定論的な手順に基づく点で独自の提案を行っており，初期値依存性の排除という優位性を持っている．

⑤ これまでに広く研究されてきた k-Means 型クラスタリングにおける種々の改善モデルを，決定論的手順に組み込むことが可能であることを示唆している本研究成果は，k-Means クラスタリングの実用性を向上させる意義を有しており，今後の応用可能性の展開が大いに期待される．

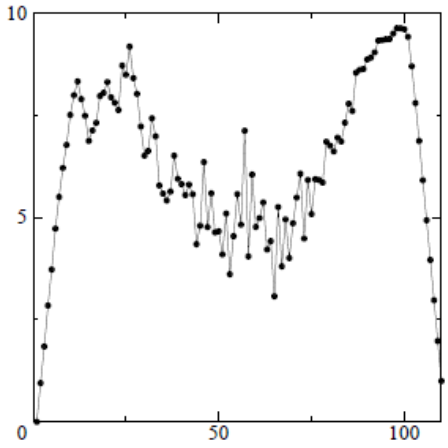

図 1　従来の k-Means モデルによるクラスター指標の例（夏目漱石の「こころ」の文書データ分析の結果. どの部の文書であるかを使用せずに作成したクラスター指標. 山状の部分がクラスター(群構造)を示す. ノイズ文書のために, 本来の 3 群構造が見出せない）

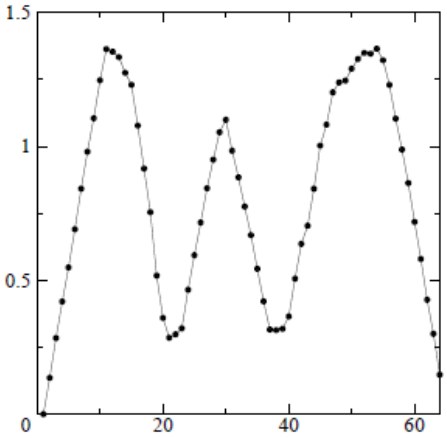

図 2　提案法によるクラスター指標の例(ノイズ文書を分析から除外することで, 3 部構成に対応する群構造が見出された)

(2) 主成分分析に基づくテキスト－単語マップの作成において，文書間の関連性を示す群構造をより明確化するために，相関ルールの抽出を目的とした変量選択機構を導入したモデルを応用した．提案手法は，以下の特性や意義を有する．

① テキスト－単語の共起関係データに主成分分析を施す際に，変量（単語）間の相関を強調する選択機構を導入することで，内容の類似した文書群がより相互に隣接して配置される 2 次元散布図が得られることを示した．

② 主成分分析に基づく k-Means クラスタリングにおける決定論的なデータ分割手順に変量選択の機構を導入した手法を提案し, 得られるアルゴリズムが本質的に主成分分析において変量間の相関ルール抽出を目的とした変量選択モデルと同等となることを明らかにした. すなわち, データ中の群構造をとらえるデータ分割手法であるクラスタリングモデルと, 他次元データの情報縮約を目的とした主成分分析との同等性という理論的特性を明らかにした.

③ 国内外の学術会議で発表を行ったほか, 引用頻度の高い Springer の Lecture Note シリーズの雑誌にも掲載された.

④ クラスター分析と主成分分析は, おのおの, 個別に広く研究開発がなされてきたが, 本研究の成果はおのおのの改善モデルが相互に有効性を持っていることを示唆しており, 両分野でのこれまでの研究成果の融合により, 実用性の向上につながる発展が大いに期待される.

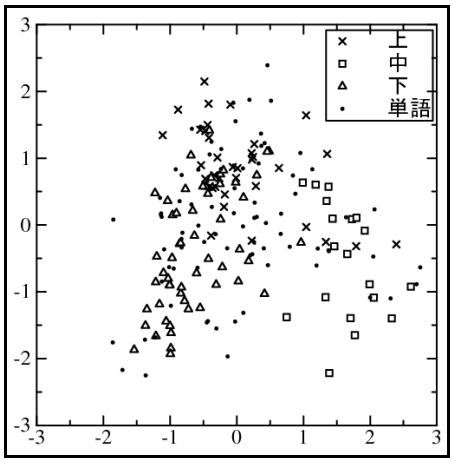

図 3　従来の主成分分析によるテキスト－単語マップの例（夏目漱石の「こころ」の文書データ分析の結果. どの部の文書であるかを使用せずに作成した散布図. 不要な単語の影響で, 3 部のテキスト群の境界があいまいである.）

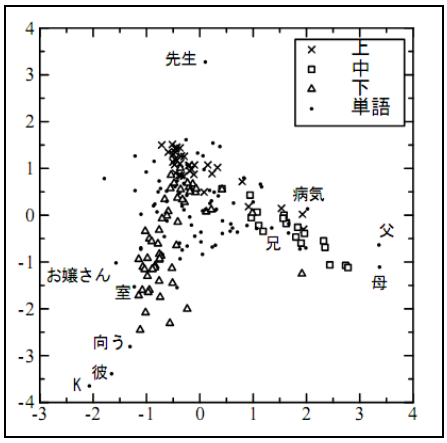

図 4　提案法によるテキスト－単語マップの例（図中に注記した単語が, 文書分析において重要と自動選別されたキーワードを表す. キーワードを強調することで, 3

部のテキスト群の境界を明確化できている．また，特定の部との関連が強い単語がキーワードとして選別されており，テキスト・単語ともに内容の要約に有益なマップ作成が実現された．）

(3) 線形ファジィクラスタリングにおけるクラスターのプロトタイプを，k-Medoidsクラスタリング法と同様に標本中のいずれかを用いて定義することにより，関係性データについても適応可能な局所的主成分分析法を開発した．提案手法は，以下の特性や意義を有する．

① 標本相互間の距離のみを用いてプロトタイプとなる直線や平面と各標本との距離を算出することで，標本の座標値が与えられない場合にも適用可能な線形クラスタリング法を提案した．

② クラスタリング基準を 2 乗距離の代わりに絶対値距離とすることにより，ノイズの影響を受けにくいロバストな構造把握が可能となる．

③ クラスターのプロトタイプの推定が，直線や平面を張る標本（medoid）の組合せ最適化問題となるが，メンバシップの大きな標本に絞って探索することで計算量削減と同時に初期値依存性の軽減が可能であることを示した．

④ 国内外の学術会議で発表を行ったほか，国際学術雑誌にも掲載される．

⑤ web ページ間の関連性を分析する web マイニングなどでは，web ページの内容を直接的に分析するのではなく，ページ間の相互の関連性から全体の特性を考察するアプローチが多く用いられている．関係性データのクラスタリングは近年，活発に研究がすすめられている分野であり，本研究の成果も同様の応用分野への発展が大いに期待される．

(4) 回帰誤差を予測式の推定とデータ分類の両方の基準として併用する FCM 型のスイッチング回帰モデルに，カテゴリー変量の逐次的な最適変換による数量化の過程を組み込んだ分析手法を開発した．提案手法は，以下の特性や意義を有する．

① 予測誤差の最小化と分類精度の向上を同時に考慮しながらカテゴリー変量を数量化する機構を提案した．

② 数量化のモデルとしては，2 種類のアプローチを提案した．単一の数量的空間を構築するアルゴリズムは，データ標本間の近接度を知る際に有効なアプローチとなる．一方，分割ごとに名義変量の数量化を行うアルゴリズムは，局所的な回帰モデルにおける量的変量と名義変量との関連をとらえるのに有効なアプローチとなる．

③ 国内外の学術会議で発表を行ったほか，学術雑誌にも掲載された．

④ カテゴリー変慮の取り扱いはテキスト分析の基盤となる技術であり，単語間の相対的な関連の分析において大いに発展が期待される．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計６件）

① K. Honda、A. Notsu、H. Ichihashi、Fuzzy PCA-guided Robust k-Means Clustering、IEEE Transactions on Fuzzy Systems、査読有、Vol.18、2010、67－79

② N. Haga、K. Honda、A. Notsu、H. Ichihashi、Local Sub-space Learning by Extended Fuzzy c-Medoids Clustering 、International Journal of Knowledge Engineering and Soft Data Paradigms、査読有、Vol.2、2010、印刷中

③ K. Honda、A. Notsu、H. Ichihashi、PCA-guided k-Means with Variable Weighting and Its Application to Document Clustering、Modeling Decisions for Artificial Intelligence、査読有、Vol. LNAI-5861、2009、281－292

④ 本多克宏、市橋秀友、野津亮、最適尺度法に基づく尺度混在データのためのFCM型スイッチング回帰モデル、システム制御情報学会論文誌、査読有、21 巻、2008、269－275

〔学会発表〕（計２５件）

① K. Honda、A. Notsu、H. Ichihashi、Visual Assessment of Cluster Tendency in Relational Data Considering Sample Responsibilities、10th International Symposium on Advanced Intelligent Systems、2009 年 8 月 18 日、Busan（Korea）

② 本多克宏、松井智宏、野津 亮、市橋秀友、ファジィ主成分分析に基づくロバスト k-Meansによるテキスト文書の分類、第 25 回ファジィシステムシンポジウム、2009 年 7 月 15 日、筑波大学（茨城）

③ K. Honda、T. Matsui、A. Notsu、H. Ichihashi、Application of Kernel Trick to Fuzzy PCA-guided Robust k-Means、Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on Advanced Intelligent Systems、2008 年 9 月 18 日、名古屋大学（愛知）

④ 和田秀樹、本多克宏、市橋秀友、野津 亮、単語の重要度を考慮した局所的主成分分析によるテキストー単語マップ作成、第 24 回ファジィシステムシンポジウム、2008 年 9 月 3 日、阪南大学（大阪）

⑤ K. Honda、T. Ohyama、H. Ichihashi、A. Notsu、FCM-type Switching Regression With Alternating Least Squares Method、2008 IEEE International Conference on Fuzzy Systems、2008年6月4日、香港(中華人民共和国)

〔図書〕(計０件)

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕

６．研究組織
(1)研究代表者
本多 克宏（HONDA KATSUHIRO）
大阪府立大学・工学研究科・准教授
研究者番号：80332964

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：