半教師付きスペクトラルクラスタリングの高度化
～特に、ソーシャルデータの解析を目的として～

代表研究者　　濵　砂　幸　裕　　近畿大学　理工学部情報学科　講師

1 はじめに

クラスタリングとは対象とするデータ集合を複数のクラスタに分割するデータ解析手法である。クラスタとは複数のデータの集まりを意味し、類似した特徴や性質を持つデータが同じクラスタに含まれるよう分類し、そうでないデータは異なるクラスタに含まれるよう分類を行うことがクラスタリングの目的である。クラスタリングの応用分野は検索エンジン・遺伝子解析などの自然科学分野をはじめとし、マーケティングや社会科学など多岐にわたる。特に近年では、計算機・記憶メディアの性能向上・低コスト化、センサー機器・ネットワークの普及などの情報通信技術の発達により、様々な自然現象や社会現象が観測・数値化され、大規模・複雑なデータベース上に蓄積されている。現状では蓄積した膨大なデータを解析し、特徴的な構造や有用な規則性を発見することが、先に述べた領域をはじめ様々な分野で求められている。そのような分野の1つにTwitterやFacebookなどのソーシャルメディア上に蓄積されるソーシャルデータが挙げられる。ソーシャルデータは行動予測やSNS上のコミュニティ発見における宝庫と言えるが、多様かつ大規模・複雑であり、絶えず更新されているため、高い処理能力を持つ解析ツールの開発が強く望まれている。しかしながら、ソーシャルデータは「年齢や性別」といった個体の情報に加えて、「SNS上の友人関係」といった個体間に与えられる情報など多様な形式のデータを含んでいるため、データの種類・量・曖昧さ・更新頻度などの点で、これまでとは比較にならないほどの大規模・複雑化が進んでいる。現状では、これまでと様相の異なるソーシャルデータを対象とした有効なデータ解析のツールは未だ開発されていない。

そこで本研究では、大規模・複雑なデータに対して、データの事前情報やユーザの意図といった**先験的知識を活用した解析を行う半教師付きクラスタリングを開発**する。特に、TwitterやFacebookなどのソーシャルメディア上に蓄積されるソーシャルデータが持つネットワーク構造などの特徴を半教師として活用することを考える。それらの半教師と従来手法および関係データに対するクラスタリング手法に加えて、グラフの分割問題として定式化されるスペクトラルクラスタリング[12]との融合を図り、新たなデータ解析の方法論を構築する。半教師付きクラスタリングの高度化により、得られる半教師のモデルと数理的な方法論は、クラスタリングに限らず、他のデータ解析手法にも適用可能であると考えられる。そのため、データ解析手法の研究者・利用者の両者にとって非常に有益である。さらに、ユーザの意図を反映したデータの可視化など、クラスタリングを用いた応用分野の発展も期待される。

本研究課題の遂行を目的とし平成25年度は、（1）逐次抽出型クラスタリングの新規開発、（2）関係データに対するクラスタリング手法の構築、（3)データの事前情報やユーザの意図を活用する半教師付きクラスタリングの開発の3点に取り組んだ。本研究課題において対象とするスペクトラルクラスタリングはデータをグラフ構造で表し、その行列に対して固有値問題を解く過程が必要となるため、データサイズが大きくなると、膨大な計算時間が必要となる。また、ノイズや外れ値を含むことで、固有値問題の対象となる行列が巨大になり、計算時間が著しく増大するため、個体間の情報のみで表される関係データからノイズや外れ値の影響を削減することも必要となる。そのような観点から、ノイズや外れ値の影響を受けにくい手法の構築および逐次抽出型アルゴリズムの開発を進めた。次に、多用な形式で表されるソーシャルデータを柔軟に扱うことを目的とし、関係データに対するクラスタリング手法の検討を行った。特に、関係データに対する従来のクラスタリング手法はノイズや外れ値の影響を考慮せずに分類を行う手法がほとんどであるため、構築した可能性クラスタリングのモデルを拡張し、関係データに対する可能性クラスタリングおよび逐次抽出型アルゴリズムを複数構築した。ここまでの検討により、ノイズや外れ値の検出および関係データに対する逐次抽出型クラスタリングの考えを援用することで、スペクトラルクラスタリングの計算時間を削減することが可能であると考えている。以上の検討をもとに、データの事前情報などの先験的知識を活用する半教師付きクラスタリングの開発を進めている。現状では、データの事前情報を個体間の類似性に反映させることで、任意のデータ対を意図したクラスタに分類することを可能とする手法の構築を進めている。以降、本申請課題の関連研究および実施内容を述べ、最後に今後の展望と自己評価を述べる。

2 関連研究

2-1 クラスタリング

代表的なクラスタリング手法である *k*-means[10,13]やファジィ *c*-平均法[2,15,16]（以下、FCM）は対象とするデータ個体を *p* 次元の実ベクトル $x_k \in R^p$ で表し、各クラスタの代表点を意味するクラスタ中心 v_i と所属する個体間の非類似度が最小になるようにデータを分類する。このとき、各個体の各クラスタへの所属する度合いを帰属度 u_{ki} で表す。また、各個体と各クラスタ中心間の類似性を示す尺度として、ユークリッド距離の自乗やマハラノビス距離などの非類似度 d_{ki} を用いる。最も古典点なクラスタリング手法である *k*-means は、帰属度を個体がクラスタに属すとき 1 とし、属さないとき 0 とするクリスプな分割を行う。一方、FCM は帰属度が 0 から 1 の連続値を取るようにファジィな分割を行う。これらをはじめ、多くのクラスタリング手法は、各個体と各クラスタ中心間の類似性を示す非類似度を用いて構成した目的関数を最小化することでクラスタの分割を行う。以下に代表的なクラスタリング手法である *k*-means[10,13]、標準型 FCM[2]、エントロピー型 FCM[15]の目的関数を示す。

$$J_h(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}\|x_k - v_i\|^2, \tag{1}$$

$$J_s(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} (u_{ki})^m \|x_k - v_i\|^2, \tag{2}$$

$$J_e(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}\|x_k - v_i\|^2 + \lambda_u \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}\log u_{ki}. \tag{3}$$

式(1)、(2)、(3) はそれぞれ、*k*-means、標準型 FCM、エントロピー型 FCM の目的関数である。また、式 (2) の $m>1$、式 (3) の $\lambda>0$ はファジィ化パラメータである。上記の目的関数を帰属度 u_{ki} に関する以下の制約条件の下で交互最適化し最小化することで、*n* 個のデータを *c* 個のクラスタに分割する。

$$\mathcal{U}_h = \left\{ (u_{ki}) : u_{ki} \in \{0,1\},\ \sum_{i=1}^{c} u_{ki} = 1,\ {}^{\forall}k \right\}, \tag{4}$$

$$\mathcal{U}_f = \left\{ (u_{ki}) : u_{ki} \in [0,1],\ \sum_{i=1}^{c} u_{ki} = 1,\ {}^{\forall}k \right\}. \tag{5}$$

式 (4) は *k*-means における制約条件であり、帰属度 u_{ki} は 0 もしくは 1 の 2 値を取る。一方、式 (5) は標準型 FCM、エントロピー型 FCM における制約条件であり、帰属度 u_{ki} は 0 から 1 の連続値を取る。これらの手法をベースとし、様々なクラスタリング手法が開発され、実データの分析やアプリケーションへの応用など様々な面で実用に供されている。

2-2 可能性クラスタリングと逐次抽出型アルゴリズム

従来のクラスタリング手法の多くは式 (4)、(5) で表される帰属度に関する制約条件を満たしながら、クラスタ分割を行う必要がある。そのため、クラスタとなるデータの集まりから外れたノイズや外れ値と考えられる個体についても帰属度の総和が 1 となるようにクラスタ分割を行うこととなる。つまり、*k*-means や FCM などのクラスタリング手法の分類結果はノイズや外れ値に大きく影響を受けることとなる。そのような問題の解決を目的とし、式 (4)、(5) の制約条件を取り除き、代わりに帰属度に関する正則化項を目的関数に付加することで、クラスタ代表点の近くでは帰属度が大きな値を取り、クラスタから離れたノイズや外れ値に対しては帰属度が小さな値を取るように修正した可能性クラスタリング[11]が提案されている。可能性クラスタリングは目的関数に付加される帰属度に関する正則化項により分類特徴が大きく異なるため、様々

な正則化手法が提案されている[9,16]。代表的な可能性クラスタリングの目的関数を以下に示す。

$$J_p(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m d_{ki}+\sum_{i=1}^{c}\eta_i\sum_{k=1}^{n}(1-u_{ki})^m, \qquad (6)$$

$$J_{pe}(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}\left\{d_{ki}+\lambda u_{ki}\left(\log u_{ki}-1\right)\right\}. \qquad (7)$$

式（6)、(7）はそれぞれ、べき乗正則化、エントロピー正則化を用いた可能性クラスタリングの目的関数である。また、$\eta_i>0$、$m>1$、$\lambda>0$ は可能性クラスタリングのためのパラメータである。可能性クラスタリングでは、式（4)、(5）で表される帰属度の総和が 1 となる制約条件は考慮されないため、クラスタから大きく外れたデータに対しては小さな帰属度を示す。

ノイズや外れ値の影響を受けにくいクラスタリング手法として、クラスタを 1 つずつ逐次的に抽出する逐次抽出型アルゴリズムがある[5,14]。逐次抽出型アルゴリズムは個体が密集した領域からクラスタを逐次的に抽出することで、通常のクラスタとノイズや外れ値のみで構成されるノイズクラスタとの 2 つのクラスタへの分割を繰り返す手法である。逐次抽出型アルゴリズムには様々な変種が存在するが、本研究課題では分類特徴を数理的に考察しやすいことおよび従来手法との比較を考え、可能性クラスタリングに基づく逐次抽出型アルゴリズムを用いる。可能性クラスタリングでは分類するクラスタ数を 1 つに設定し、データの分割を行うことが可能である。このとき、全ての個体のクラスタへの帰属度は 0 から 1 の数値を取る。そのため、帰属度にしきい値を定めることで、抽出するクラスタとノイズクラスタの 2 つに分割するクラスタリング手法とみなすことができる。本研究課題では、新たな可能性クラスタリング手法を提案し、それを基に逐次抽出型アルゴリズムを構築する。さらに、構築したアルゴリズムを通常のベクトルデータのみならず関係データを扱えるように拡張することで、様々な形式で表されるソーシャルデータを解析する分類手法の土台を構成する。

2-3 関係データに対するクラスタリング

従来のクラスタリング手法は p 次元の実ベクトルで表されたデータを対象としている。しかしながら、クラスタリングの対象となるデータには様々な形式のデータが存在する。そのようなデータの 1 つに個体間の類似性のみで表される関係データがある。関係データとは、国家間の輸出入額、Web サイト上のページ遷移などのように、クラスタリングの対象となる個体間の類似度あるいは非類似度のみが与えられたデータを指す。このような関係データに対するクラスタリング手法として、ファジィノンメトリックモデル[6,18]（以下、FNM)、リレーショナルファジィ c-平均法[8]（以下、RFCM)、AP アルゴリズム[20]（以下、AP）など様々なクラスタリング手法が提案されている。スペクトラルクラスタリングもそのような関係データに対するクラスタリング手法の 1 つであり、データ分類の性能・数理的な特徴など様々な理由から近年大きな注目を集めている。関係データに対するクラスタリング手法では、個体間の類似性 r_{kt} を用いた行列を距離行列あるいは非類似度行列などと呼び、その行列を対象にクラスタ分割を行う。このとき、n 個のデータに対して、$n\times n$ の距離行列が生成される。一般にクラスタリングで対象となるデータ対には対称性が成り立つが、関係データの中には非対称となるものも数多く存在するため、非対称データの扱いも重要な課題の一つである。

FNM、RFCM、AP などのクラスタリング手法では、k-means などのようにクラスタ内の代表点を決めることなく、上記の行列で示された個体間の類似性を用いて定式化された目的関数を最小化することでクラスタ分割を行う。一方スペクトラルクラスタリングは、上記の行列の固有値問題を解く過程が必要となる。固有値問題はデータサイズが大きくなると計算時間も増大することが知られている。そのため、より実用的な解析を行うには関係データに含まれるノイズや外れ値などの不要なデータを削減し、計算時間の短縮と意味のあるデータで構成されるようなクラスタへの分割が必要となる。しかしながら、関係データに含まれるノイズや外れ値の検出は、通常のベクトルデータほど盛んに研究されていない。そのため、今後の増大が予想されるソーシャルデータの解析が求められる現状においては、ベクトルデータや関係データに含まれるノイズや外れ値の検出を統一的なアプローチで行うことが重要であると考えられる。そのような観点から、上記の可能性クラスタリングや逐次抽出型アルゴリズムをスペクトラルクラスタリングへ援用あるいは前処理に用いることで、ノイズや外れ値を検出し、意味のあるソーシャルデータに対してクラスタ分割を行うことが必要不

可欠であると考えている。

3 スパース性を示す可能性クラスタリング

3-1 L1 正則化可能性クラスタリング

はじめに、本研究課題で取り組んだ逐次抽出型アルゴリズムの基となる L1 正則化可能性クラスタリングについて説明する。先に述べたとおり、ノイズや外れ値の影響を受けにくいクラスタリング手法である可能性クラスタリングは、目的関数に正則化項を付加することで構成される。本研究課題では、ノイズや外れ値といったクラスタから遠く離れた個体からの影響を可能な限り軽減し、意味のあるデータ集合をクラスタとして分類する新たな手法として、L1 正則化可能性クラスタリング[7]を提案した。L1 正則化は機械学習などの分野で盛んに研究が進められている正則化手法であり、不要な特徴を 0 とすることで解の一部あるいは大部分が 0 となるスパースな解を得る手法である[3]。L1 正則化を用いたクラスタリング手法として、小さな帰属度を 0 とするように正則化項を加えたスパース可能性クラスタリングが以前に提案されている[9]。本研究課題では、それとは逆のアプローチを取り、帰属度の大きな個体に対して、帰属度を 1 とするように正則化項を加えることで蜜なデータ構造を持つクラスタを形成する手法を提案した。本手法のアプローチは可能性クラスタリングをベースにしているため、ノイズや外れ値に対しては帰属度が小さく、クラスタの代表点に近いデータに対しては帰属度が 1 となるようにクラスタ分割を行う手法である。L1 正則化を用いた可能性クラスタリング（以下、L1PCM）の 2 種類の目的関数を以下に示す。

$$J_{pl}(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^m\|x_k-v_i\|^2+\gamma\sum_{k=1}^{n}\sum_{i=1}^{c}|1-u_{ki}|, \tag{8}$$

$$J_{epl}(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}\left\{u_{ki}\|x_k-v_i\|^2+\lambda u_{ki}\left(\log u_{ki}-1\right)\right\}+\gamma\sum_{k=1}^{n}\sum_{i=1}^{c}|1-u_{ki}|. \tag{9}$$

ここで、$m>1$、$\gamma>0$、$\lambda>0$ は L1PCM に対するパラメータである。上記の目的関数において、帰属度 u_{ki} は絶対値項を含んでいるため、通常のようにラグランジュ関数の偏微分より解を導出することはできない。そこで、可能性クラスタリングにおける帰属度は各クラスタについて独立であることから、差分法を用いて絶対値項を置き換えることで偏微分可能な以下の関数が得られる。

$$J_{pl}^{ki}(u_{ki})=(u_{ki})^m d_{ki}+\gamma\left(\xi^{+}+\xi^{-}\right).$$

ここで、ξ^{+}、ξ^{-} は次の制約条件を満たすパラメータである。

$$1-u_{ki}\le\xi^{+},\quad 1-u_{ki}\ge-\xi^{-},\quad \xi^{+},\xi^{-}\ge 0.$$

上記の目的関数と制約条件から、ラグランジュ乗数 β^{+}、β^{-}、ψ^{+}、$\psi^{-}\ge 0$ を導入することで、以下のラグランジュ関数が得られる。

$$L_{pl}=(u_{ki})^m d_{ki}+\gamma\left(\xi^{+}+\xi^{-}\right)+\beta^{+}\left(1-u_{ki}-\xi^{+}\right)+\beta^{-}\left(-1+u_{ki}-\xi^{-}\right)-\psi^{+}\xi^{+}-\psi^{-}\xi^{-}.$$

上記のラグランジュ関数を最適性の必要条件から導かれる関係式を用いて式を整理し、再び最適性の必要条件を用いることで、帰属度 u_{ki} に対する解が求まる。その解を主問題に代入し、ラグランジュ双対問題と解が満たす条件を考慮することで、最適解が導出される。以上の手順を用いて導出した目的関数(8)、(9)における最適解を以下に示す。

$$u_{ki} = \begin{cases} 1 & \left(0 \leq d_{ki} \leq \frac{\gamma}{m}\right) \\ \left(\frac{\gamma}{m d_{ki}}\right)^{\frac{1}{m-1}} & \left(\frac{\gamma}{m} < d_{ki}\right) \end{cases} \tag{10}$$

$$u_{ki} = \begin{cases} 1 & (0 \leq d_{ki} \leq \gamma) \\ \exp\left(-\frac{d_{ki} - \gamma}{\lambda}\right) & (\gamma < d_{ki}) \end{cases} \tag{11}$$

式(10)、(11)は各個体とクラスタ中心間の非類似度がパラメータ内に収まる場合には帰属度 1 を示し、そうでない場合にはクラスタから離れるほど小さな帰属度を示すことを表している。本手法はパラメータ γ の値により、帰属度が大きな値を示す領域が変化する手法となっており、これまでの可能性クラスタリングとは異なる特徴を持つ分類規則を示す手法である。L1PCM のアルゴリズムは最適解の必要条件から導かれた解を用いて交互最適化を行い、目的関数の最小化を行うことで構成される。

3-2 逐次抽出型アルゴリズム

次に、提案した L1PCM を用いた逐次抽出型アルゴリズムについて説明する。逐次抽出型アルゴリズムは目的関数をベースとした手法、アルゴリズムをベースとした手法など様々な手法が提案されている[5,14]。本研究課題では、提案した L1PCM を用いて、ノイズや外れ値の影響を受けにくい逐次抽出型アルゴリズムを構築した。構築した逐次抽出型アルゴリズムは分類するクラスタ数を 1 と設定し、L1PCM を用いてクラスタ分割を行い、クラスタを逐次的に抽出するアルゴリズムとなっている。特に、本アルゴリズムはクラスタ代表点の付近では帰属度が 1 となる領域が広く、ノイズや外れ値などのクラスタから離れたところでは帰属度がかなり小さな値を取るように構成されている。そのため、これまでの逐次抽出型アルゴリズムよりも複雑な分類境界を示すことが可能である。以下に提案手法を用いた逐次抽出型アルゴリズムを示す。

Algorithm 2 Sequential cluster extraction by L_1PCM

STEP 1	Give X, initial values u_{ki} and v_i and parameters m or λ and γ.
STEP 2	Repeat L_1PCM algorithm with $c = 1$ for calculating u_{ki} and v_i until convergence.
STEP 3	Extract $\{x_k \mid u_{ki} = 1\}$ from X.
STEP 4	If $X = \emptyset$, stop. Otherwise, give initial values and go back to **STEP 2**.

図 1　提案手法を用いた逐次抽出型アルゴリズム

本研究課題で構築する逐次抽出型アルゴリズムは、L1PCM の正則化パラメータを用いて、帰属度が大きな値を取る領域を調節することで、任意の大きさのクラスタを抽出するようにアルゴリズムを実行することが可能である。表 1 から 3 に提案手法を用いた逐次抽出を行った結果を示す。数値実験に用いたデータは UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) 上で公開されているベンチマークデータの 1 つである Breast Cancer Wisconsin data set (以下、BCW) である。BCW データは 569 個体、9 次元のベクトルとして表され、2 クラスタに分類される。表 1 中の sL1PCM は目的関数(8)による提案手法、表 2 中の eL1PCM は目的関数(9)による提案手法、表 3 中の SHCM は比較に用いた逐次抽出型ハード c-平均法[5]である。表 1、2、3 で用いている、Rand Index[17]は 2 つのクラスタ分割の一致度を表す指標であり、分類が既知のデータセットとクラスタリングによる分類結果を用いて、クラスタ分割の精度を評価するのに用いられる。値が 1 に近いほど 2 つのクラスタ分割は一致しており、0 に近いほどクラスタ分割は一致していない。また、Num. of clusters は抽出したクラスタ数を示している。Ave、SD、Max、Min はそれぞれ、平均値、標準偏差、最大値、最小値を意味している。表 1、2 と表 3 を比較すると、提案手法を用いた場合に Rand Index の平均値が大きいことが確認できる。また、抽出するクラスタ数は表 1 の sL1PCM が最も安定しており、次に eL1PCM、

SHCM となっていることが確認できる。これらの結果より、BCW データセットに対しては提案手法を用いた逐次抽出型アルゴリズムが有効であることがわかる。

表 1　Breast Cancer Wisconsin data set に対して、sL_1PCM による逐次抽出を行った結果（100 回試行）

	Rand Index		Num. of clusters			
γ/m	Ave	SD	Ave	SD	Max	Min
5.0	0.865	1.54×10^{-2}	4.66	11.58	116	2
7.5	0.894	1.90×10^{-2}	4.33	4.16	36	3
10.0	0.827	1.86×10^{-2}	4.27	1.22	10	3
12.5	0.789	1.40×10^{-2}	3.29	0.67	6	3
15.0	0.766	1.14×10^{-2}	3.09	0.32	5	3

表 2　Breast Cancer Wisconsin data set に対して、eL_1PCM による逐次抽出を行った結果（100 回試行）

	Rand Index		Num. of clusters			
γ	Ave	SD	Ave	SD	Max	Min
5.0	0.853	1.07×10^{-2}	18.35	5.84	43	12
7.5	0.883	5.57×10^{-3}	12.10	2.93	25	9
10.0	0.894	1.05×10^{-2}	9.93	2.67	19	6
12.5	0.837	2.25×10^{-2}	7.15	0.92	9	4
15.0	0.786	2.10×10^{-2}	4.00	0.00	4	4

表 3　Breast Cancer Wisconsin data set に対して、SHCM による逐次抽出を行った結果（100 回試行）

	Rand Index		Num. of clusters			
D	Ave	SD	Ave	SD	Max	Min
5.0	0.846	7.02×10^{-3}	43.63	6.94	72	38
7.5	0.888	1.06×10^{-2}	20.52	3.20	31	17
10.0	0.890	1.05×10^{-2}	11.97	1.88	20	10
12.5	0.844	2.43×10^{-2}	8.69	0.83	11	5
15.0	0.801	1.10×10^{-2}	5.24	0.67	6	4

3-3 関係データに対する逐次抽出型アルゴリズム

上記の検討を基に、FNM、RFCM、AP などの関係データに対するクラスタリング手法に対して、L1 正則化を用いた新たなクラスタリング手法を構築した。それぞれの手法の目的関数に式(8)、(9)と同様に L1 正則化項を付加し、最適解の導出を行い、分類アルゴリズムを提案した。さらに、提案したアルゴリズムを用いて関係データに対する逐次抽出型アルゴリズムを構築した。構築したアルゴリズムはノイズや外れ値の影響を低減し、関係データから多くのデータが含まれる密なクラスタを抽出することが可能であり、計算時間においても、従来手法と同程度で実行可能であるため、膨大なソーシャルデータの解析を行う実用的な手法であると考えている。また、これまでの検討により、関係データに対する逐次抽出型アルゴリズムが構築できたことから、外れ値検出および逐次抽出の考えを援用することで、スペクトラルクラスタリングの計算時間を削減し、より実用に適した手法の構築が可能であると考えている。現状では、手法の構築が完了した段階であるため、今後の計画として複数のベンチマークデータや実データを用いた評価・検証を引き続き継続し、提案手法がどのようなデータの分類に適しているかを明らかにする。さらに、従来手法との比較および数理的なつながりを明らかにすることでソーシャルデータ解析のみならず、ネットワーク構造を持つデータを扱う他の分野への応用を進めたいと考えている。

4 関係データの半教師モデルを用いたクラスタリング

これまでの検討をもとに、データの事前情報などの先験的知識を活用する半教師付きクラスタリングを開発した。半教師付きクラスタリングはデータの事前情報やユーザの意図をクラスタリングの枠組みで扱うことで、膨大なデータを柔軟に処理する手法である。半教師の代表的な例として、「２つのデータ対は同じクラスタに属する」という must-link、「２つのデータ対は違うクラスタに属する」という cannot-link といった対制約が知られている[1,4,19]。項目（２）の遂行により得られた知見をもとに、関係データの半教師モデルを構築し、エントロピー型ファジィ c-平均法の正則化項として扱う手法を提案した。提案手法の目的関数を以下に示す。

$$J_{eefap}(U,V,W) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}\|x_k - v_i\|^2 - \sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{k=1}^{n} u_{ki}w_{ti}\alpha_{kt} + \sum_{i=1}^{c}\sum_{k=1}^{n}\sum_{k=1}^{n} u_{ki}w_{ti}\beta_{kt} + \lambda_u \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ki}\log u_{ki} + \lambda_w \sum_{i=1}^{c}\sum_{t=1}^{n} w_{ti}\log w_{ti} \quad (12)$$

式（12）において、w_{ti} は prototype weight と呼ばれる関係データの代表点に対する重みであり、λ_u、λ_w は帰属度および prototype weight のファジィ化パラメータである。また、α_{kt}、β_{kt} は個体間の半教師を反映させるための正則化項であり、以下で定義される。

$$\alpha_{kt} = \begin{cases} \alpha & ((x_k, x_t) \in ML) \\ 0 & (\text{otherwise}) \end{cases} \qquad \left(\because \alpha = \gamma_{ml} \max_{p,q} \|x_p - x_q\|^2\right), \quad (13)$$

$$\beta_{kt} = \begin{cases} \beta & ((x_k, x_t) \in CL) \\ 0 & (\text{otherwise}) \end{cases} \qquad \left(\because \beta = \gamma_{cl} \max_{p,q} \|x_p - x_q\|^2\right). \quad (14)$$

式（13）、（14）はそれぞれ、must-link、cannot-link の定義式となっている。上記の目的関数と帰属度および prototype weight に対する制約条件から、新たな半教師付きクラスタリングのアルゴリズムを構築した。本手法はデータの事前情報などの先験的知識を半教師として扱い、対制約として個体間の類似性に反映させることで、任意のデータ対を意図したクラスタに分類することが可能である。本手法はベクトルデータあるいは関係データを扱うクラスタリング手法のモデルをもとに構成しており、従来手法の拡張が容易に行えるという点に特徴がある。

構築した半教師付きクラスタリング手法を用いた数値実験の結果を表 4、5 に示す。数値実験に用いたデータは UCI Machine Learning Repository 上で公開されているベンチマークデータの 1 つである Iris data set である。Iris データは 150 個体、4 次元のベクトルとして表され、3 クラスタに分類される。Iris データは分類結果が既知であるため、その情報を用いて 100、300、500 個の対制約をランダムに発生させ、実験を行った。また、γ_{ml}、γ_{cl} は半教師の影響度合いを定めるパラメータであり、その値が大きいほど対制約は満たされやすくなり、そうでない場合には制約を違反することもある。表 4、5 の結果から、多くの対制約を与え、パラメータの値を大きくすることで分類結果が大きく向上していることが確認できる。また、Iris データに対しては must-link の方が cannot-link よりも効果的であることが Rand Index、制約違反の 2 つの結果から確認できる。この他にも同様のアプローチにより構築された複数の半教師付きクラスタリングを比較したところ、本項で示した手法がデータの事前情報を半教師として扱うのに適した手法であることが確認された。

今後の計画として、データに関する先験的知識が明らかでない場合、つまり少数の半教師のみが対象となる際に、局所的な対制約を大域的に波及することで、全体の分類結果を向上させる推論型の半教師付きクラスタリング手法の構築が必要であると考えている。実問題のデータに対して半教師を付与することは人的・時間的コストが必要となるため、大量の半教師を与えることは現実的な手法とは言い難い。そこで、少量の半教師が持つ局所的な情報を大域的に波及することで、データ集合全体の分類性能向上を行う手法を構築することが、実用的な解析を行う上で重要となると考えている。また、それらの検討で得られたデータの事前情報をスペクトラルクラスタリングに活用することで、ネットワーク構造などの付加情報を持つソーシャルデータを柔軟に処理し、有効な解析を行う半教師付きクラスタリング手法が開発できると考えている。

表 4　eFCM-eFAP を用いて分類を行った結果（Rand Index 100 回試行）

Number of pairwise constraints	$\gamma_{ml}, \gamma_{cl} = 0.50$		$\gamma_{ml}, \gamma_{cl} = 1.00$	
	ML	*CL*	*ML*	*CL*
100	0.835 ± 0.005	0.831 ± 0.003	0.847 ± 0.008	0.834 ± 0.004
300	0.859 ± 0.007	0.838 ± 0.004	0.955 ± 0.028	0.845 ± 0.010
500	0.933 ± 0.015	0.845 ± 0.004	0.988 ± 0.031	0.878 ± 0.021

表 5　eFCM-eFAP を用いて分類を行った結果（制約違反 100 回試行）

Number of pairwise constraints	$\gamma_{ml}, \gamma_{cl} = 0.50$		$\gamma_{ml}, \gamma_{cl} = 1.00$	
	ML	*CL*	*ML*	*CL*
100	24.96 ± 4.22	12.47 ± 3.45	22.21 ± 4.63	12.15 ± 3.44
300	62.35 ± 8.19	36.74 ± 5.73	12.83 ± 7.96	33.91 ± 6.79
500	42.83 ± 11.82	57.38 ± 6.71	5.20 ± 12.72	41.32 ± 14.48

5 まとめ

本研究課題では、大規模・複雑なデータに対して、データの事前情報やユーザの意図といった**先験的知識を活用した解析を行う半教師付きクラスタリングの開発**を目的とし、（1）逐次抽出型クラスタリングの新規開発、（2）関係データに対するクラスタリング手法の構築、（3）データの事前情報やユーザの意図を活用する半教師付きクラスタリングの開発の 3 点に取り組んだ。以上の検討で得られた知見をスペクトラルクラスタリングなどのクラスタリング手法に援用することで、膨大なソーシャルデータやネットワーク構造などの特徴を持つデータを柔軟に処理する新たなクラスタリング手法が構築できると考えている。

本研究課題の最終的な自己評価は、

（1）　膨大なソーシャルデータを解析する半教師付きクラスタリングの開発

（2）　関係データに対するクラスタリングの高度化

以上の 2 点から行うこととなる。項目（1）の半教師付きクラスタリングの開発については、平成 25 年度の検討により一定の成果が得られたと考えている。また、項目（2）については、関係データからのノイズ検出や非対称データの扱いなど、今後のデータ解析分野の重要な課題が残されていると考えている。上記項目のいずれか 1 つでも達成できれば十分な成果と考えているが、本研究課題の成果を実問題に適用し、より波及させることが現代の情報通信分野における本研究課題の意義と考えている。そのため、本研究課題の遂行により得られた関係データのクラスタリングに関する知見を、それに留まらず、より高次のレベルに発展させることを目標として設定している。

【参考文献】

[1] Basu, S., Davidson I., Wagstaff K., eds., `Constrained Clustering: Advances in Algorithms, Theory and Applications', Data Mining and Knowledge Discovery vol. 3, Chapman & Hall/CRC, 2008.

[2] Bezdek J. C., `Pattern Recognition with Fuzzy Objective Function Algorithms', Plenum Press, New York, 1981.

[3] Candes E. J., Wakin M. B., Boyd S., Enhancing Sparsity by Reweighted l1 Minimization, Journal of Fourier Analysis and Applications, Vol. 14, No. 5, pp. 877-905, 2008.

[4] Chapelle O., Schoelkopf B., Zien A., eds., `Semi-Supervised Learning', MIT Press, 2006.

[5] Dave R. N., Characterization and detection of noise in clustering, Pattern Recognition Letters, Vol. 12, No. 11, pp. 657-664, 1991.

[6] Endo Y., On Entropy Based Fuzzy Non Metric Model - Proposal, Kernelization and Pairwise Constraints -, Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol. 16, No. 1, pp. 169-173, 2012.
[7] Hamasuna Y., Endo Y., Sequential Extraction By Using Two Types of Crisp Possibilistic Clustering, Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013), pp.3505-3510, 2013.
[8] Hathaway R. J., Davenport J. W., Bezdek J. C., Relational Duals of the c-Means Clustering Algorithms, Pattern Recognition, Vol. 22, No. 2, pp. 205-212, 1989.
[9] Inokuchi R, Miyamoto S., Sparse Possibilistic Clustering with L1-Regularization, Proc. of The 2007 IEEE International Conference on Granular Computing (GrC2007), pp. 442-445, 2007.
[10] Jain A. K., Data clustering: 50 years beyond K-means, Pattern Recognition Letters, Vol. 31, No. 8, pp. 651-666, 2010.
[11] Krishnapuram R., Keller J. M., A possibilistic approach to clustering, IEEE Transactions on Fuzzy Systems, Vol. 1, No. 2, pp. 98-110, 1993.
[12] Kulis B., Basu S., Dhillon I., Mooney R., Semi-supervised graph clustering: a kernel approach, Machine Learning, Vol. 74, No. 1, pp. 1-22, 2009.
[13] MacQueen J. B., Some methods for classification and analysis of multivariate observations. Proc. of Fifth Berkeley Symp. on Math. Statist. and Prob., pp. 281-297, 1967.
[14] Miyamoto S., Kuroda Y., Arai K., Algorithms for Sequential Extraction of Clusters by Possibilistic Method and Comparison with Mountain Clustering, Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol. 12, No. 5, pp. 448-453, 2008.
[15] Miyamoto S., Mukaidono M., Fuzzy c-means as a regularization and maximum entropy approach, Proc. of the 7th International Fuzzy Systems Association World Congress} (IFSA'97), Vol. 2, pp. 86-92, 1997.
[16] Miyamoto S., Ichihashi H., and Honda K., `Algorithms for Fuzzy Clustering', Springer, Heidelberg, 2008.
[17] Rand W. M., Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, Vol. 66, No. 336, pp. 846-850, 1971.
[18] Roubens M., Pattern classification problems and fuzzy sets, Fuzzy Sets and Systems, Vol. 1, pp. 239-253, 1978.
[19] Wagstaff K., Cardie C., Rogers S., Schroedl S., Constrained k-means clustering with background knowledge, Proc. of the 18th International Conference on Machine Learning (ICML 2001), pp. 577-584, 2001.
[20] Windham M. P., Numerical classification of proximity data with assignment measures, J. of Classification, Vol.2, pp. 157-172, 1985.

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
Sequential Extraction By Using Two Types of Crisp Possibilistic Clustering	Proc. of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013)	2013 年 10 月
On Cluster Extraction from Relational Data Using Entropy Based Relational Crisp Possibilistic Clustering	Proc. of The Fifth International Conference on Knowledge and Systems Engineering (KSE 2013)	2013 年 10 月