イジングマシンを用いたクラスタリング手法の研究

理学専攻・情報科学コース　2040655　松本 奈紗

1　はじめに

近年，E コマース市場の成長による宅配便取扱数や輸送交通量の増加に伴い，運送業の人手不足や排気ガスによる環境汚染が社会問題となっている．この問題の解決策として，宅配の物流拠点から個人宅までの配送における配送計画問題を解き，配送を効率化することが挙げられる．配送計画問題についての研究を進めていく中で，複数の車両に荷物を割り当てたり，各車両の配送経路を求める際に荷物をグループ分けしたりするためのクラスタリングの精度が最終的な配送ルートに影響を及ぼし，重要であることがわかった．そこで本研究では，2 次元空間上の頂点を 2 点間距離をもとに分割するクラスタリング問題に着目し，クラスタリングの精度を向上させるために，イジングマシンを用いたクラスタリング手法を 2 種類提案する．ここでは 1 つ目の手法について述べる．イジングマシンとは，組合せ最適化問題を解くための専用計算機である．イジングマシンは，目的関数を定義するパラメータを入力とし，目的関数を最小化するような二値変数の組を解として返す．イジングマシンは組合せ最適化問題を高速に解くことが期待されている．クラスタリングは組合せ最適化問題として定式化できる．

2　モデル

2.1　単純コスト法

イジングマシンを用いたクラスタリングの従来手法の目的関数は次式のように表される．

$$H = \sum_{1\le i<j\le N} d_{ij} \sum_{g=1}^{G} x_{i,g}x_{j,g} + \alpha \sum_{i=1}^{N} \left(\sum_{g=1}^{G} x_{i,g} - 1 \right)^2 \quad (1)$$

ここで，d_{ij} は点 i と点 j 間の距離，N は点の数，G はグループ数である．点 i がグループ g に所属するなら $x_{i,g} = 1$，そうでなければ $x_{i,g} = 0$ とする．したがって，(1) の右辺第 1 項は，同じグループに所属する点の 2 点間距離の総和となる．右辺第 2 項は，各点はちょうど 1 つのグループに所属するという制約を満たすための制約項である．この制約を満たす解を有効解と呼ぶことにする．α は正の定数で，制約の強さを表す．この目的関数では各グループに含まれる点のペアの組合せの数だけペア間の距離が加算されるため，各グループに含まれる点の数が均等になりやすい傾向がある．そのため，不均一に分布したデータセットに対して，うまくクラスタリングすることができないという問題がある．

2.2　反復分数コスト法

前節の問題を解決するために，次式で表される別の目的関数を提案する．

$$H = \sum_{g=1}^{G} \frac{\sum_{i<j} d_{ij}x_{i,g}x_{j,g}}{N_g(N_g-1)} + \alpha \sum_{i=1}^{N} \left(\sum_{g=1}^{G} x_{i,g} - 1 \right)^2 \quad (2)$$

ここで，$N_g = \sum_{i=1}^{N} x_{i,g}$ はグループ g の点の数である．式 (2) の右辺第一項は各グループ内の点のペア間の平均距離の総和を表している．

イジングマシンは式 (2) のような非 QUBO 形式を直接最小化することができない．そこで，QUBO 形式で記述された別の目的関数を反復的に最小化するハイブリッドアルゴリズムを採用した．この手法を反復分数コスト法と呼ぶことにする．反復分数コスト法は，配送計画問題を解くために提案されたハイブリッド・パラメトリック法を基盤としている [1]．ハイブリッド法では，元の分数目的関数を最小化する代わりに，イジングマシンを用いて，離散最適化ステップで，対応するパラメトリック問題を反復的に解く．

反復分数コスト法のアルゴリズムを以下に示す．

1. 誤差パラメータ δ を設定, 反復カウンタ n を $n = 0$, パラメータ λ を初期値 $\lambda_0 = 0$ とする．
2. イジングマシンを用いて次式で表された目的関数を最小化する．

$$H = \sum_{g=1}^{G}\sum_{i<j}^{N} d_{ij}x_{i,g}x_{j,g} + \alpha \sum_{i=1}^{N} \left(\sum_{g=1}^{G} x_{i,g} - 1 \right)^2 - \lambda_n \sum_{g=1}^{G} \left(\sum_{i=1}^{N} x_{i,g} \right) \left(\sum_{i=1}^{N} x_{i,g} - 1 \right) \quad (3)$$

ここで，$\boldsymbol{x}$ は二値変数を表す．また，得られた解を $\hat{\boldsymbol{x}}$ とする．

3. $\hat{\boldsymbol{x}}$ が有効解であれば，次の反復での λ を次式のように設定する．

$$\lambda_{n+1} = \sum_{g=1}^{G} \frac{\sum_{i<j}^{N} d_{ij}\hat{x}_{i,g}\hat{x}_{j,g}}{\hat{N}_g(\hat{N}_g-1)} \quad (4)$$

ここで，$\hat{N}_g = \sum_i \hat{x}_{i,g}$ とする．$\hat{\boldsymbol{x}}$ が有効解でなければ，解なしで終了する．

4. $|\lambda_{n+1} - \lambda_n| \le \delta$ の場合，$\hat{\boldsymbol{x}}$ を最終解として終了する．それ以外の場合，$n+1 < n_{\max}$ の間は，$n \to n+1$ としてステップ 2 に戻り，$n+1 = n_{\max}$ のときは，解なしで終了する．

アルゴリズムが正常に終了した場合，λ_{n+1} は最小コストを与えることが期待される．誤差パラメータ δ は 10^{-6}，最大反復回数 $n_{\max}$ は 10 とした．

3　結果

本節では単純コスト法と反復分数コスト法の実行結果を示す．目的関数を最小化するためのアルゴリズムとして，シミュレーテッド・アニーリング法 (SA) とシミュレーテッド分岐アルゴリズム (SB) [2] の 2 種類を用いる．ここでは，SB の結果のみを述べる．正方形領域に一様に分布する点の集合と，同じ大きさの領域に不均一に分布する点の集合の 2 種類のデータセットに対して実行した．使用したデータセットは 200 点で構成され，これを 10 グループにクラスタリングした．ハイパーパラメータ α の値は単純コスト法では一様分布と不均一分布でそれぞれ 6 と 5，反復分数コスト法

では一様分布と不均一分布の両方で 5.5 とした．また，離散最適化の計算時間を制御するパラメータである時間ステップ数を変化させる．時間ステップごとに単純コスト法では 100 回，反復分数コスト法では 50 回のシミュレーションを行なった．

3.1　単純コスト法

図 1 は単純コスト法のクラスタリング結果の中で最も良いものを示している．同じ色の点は同じグループに属していることを表す．不均一分布では，見かけ上のクラスターが二つのグループに分割されていたりした．

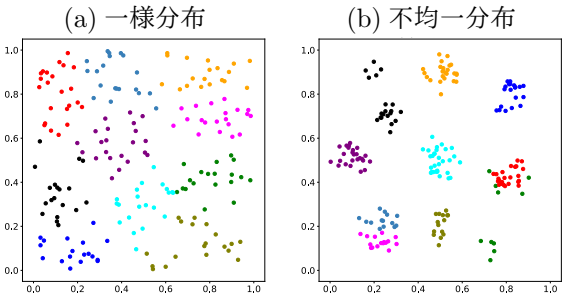

図 1: 単純コスト法のクラスタリング結果．

図 2 は，クラスタリング結果が実行時間にどのように依存するかを示したものである．箱ひげ図は，有効解の平均シルエット係数の分布を示している．平均シルエット係数はクラスタリング評価指標であり，高いほどクラスタリング結果が良いとされる．単純コスト法での実行時間は，1 つの離散最適化問題の解を得るために必要な時間であるワンショット実行時間と定義する．図 2 での実行時間は時間ステップ数ごとの有効解の平均値である．平均シルエット係数は，一様に分布したデータセットでは 0.35–0.39 程度またはそれ以下であるのに対し，不均一に分布したデータセットでは 0.45–0.6 程度またはそれ以上と高くなっている．この差は，データ分布の違いを反映している．

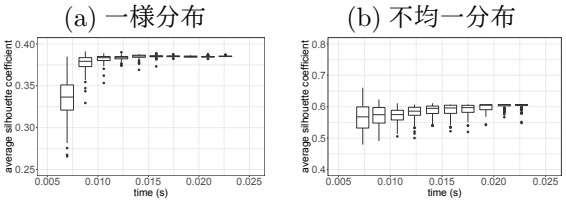

図 2: 平均シルエット係数の箱ひげ図．横軸はワンショット実行時間．

3.2　反復分数コスト法

図 3 は反復分数コスト法のクラスタリング結果の中で最も良いものを示している．図 1 と比較して，不均一分布においてうまく分類されていることがわかる．

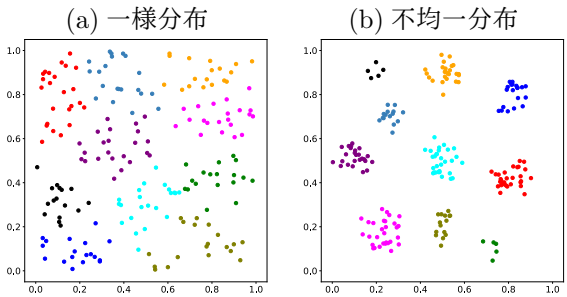

図 3: 反復分数コスト法のクラスタリング結果．

式 (3) の有効解をいくつか得るために，離散最適化ステップ (ステップ 2) で 100 回の最適化サンプリングを行った．最終的な有効解率は，時間ステップ数が 1600 以上の場合は 80–100%となった．最終解が得られるまでの，ステップ 2–4 の反復回数はほとんどの場合においては，3 回だった．

図 4 は反復分数コスト法による離散最適化にかかる時間を示したものである．反復分数コスト法での実行時間は，ワンショット実行時間の合計と定義する．実行時間は，反復回数，離散最適化ステップでの最適化サンプリング数，時間ステップ数に依存する．平均シルエット係数は，図 4 に対応する領域では，ハイパーパラメータ α と時間ステップ数にほとんど依存しない．一様分布では，平均シルエット係数は 0.392–0.394 となり，単純コスト法と同程度の値となった．一方で，不均一分布では，平均シルエット係数は 0.709–0.716 となり，単純コスト法よりも約 18 %高くなった．

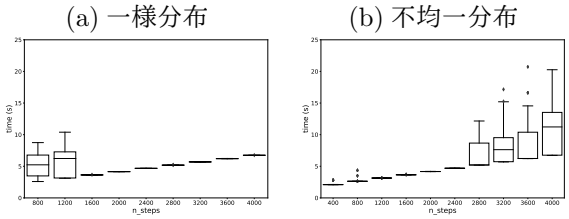

図 4: 実行時間の箱ひげ図. 横軸は時間ステップ数.

4　まとめ

本研究では，2 次元空間上の頂点を 2 点間距離をもとに分割するクラスタリング問題に着目し，クラスタリングの精度を向上させるために，イジングマシンを用いた反復分数コスト法と呼ばれるクラスタリング手法を提案した．この手法は，従来手法である単純コスト法と比べて，不均一に分布したデータセットに対し，クラスタリング精度が高いことが確認できた．

参考文献

[1] A. Ajagekar, T. Humble and F. You, Computers & Chemical Engineering **132**, 106630 (2020).

[2] H. Goto, K. Endo, M. Suzuki, Y. Sakai, T. Kanao, Y. Hamakawa, R. Hidaka, M. Yamasaki and K. Tatsumura, Science Advances **7**, eabe7953 (2021).