様式 C-19

科学研究費補助金研究成果報告書

平成　23　年　5　月　10　日現在

機関番号：10101
研究種目：基盤研究(C)
研究期間：2008　～　2010
課題番号：20500123
研究課題名（和文）グラフの彩色に基づき局所的な制約と大域的な制約を統合するクラスタリング手法
研究課題名（英文）A Clustering Method based on graph Coloring for integrating Local and Global Constraints
研究代表者
吉田　哲也（ TETSUYA YOSHIDA ）
北海道大学・大学院情報科学研究科・准教授
研究者番号：80294164

研究成果の概要（和文）：
本研究では，近年グラフ理論の分野で提案された b 彩色という概念に着目し， この彩色が満たすべき 2 つの制約をそれぞれ局所的な制約と大域的な制約に対応させることにより，グラフの彩色に基づくクラスタリング手法の開発に取り組んだ．具体的には，データ対の非類似度に基づいてデータ集合をグラフ表現に変換する手法を定式化し，構築したグラフに対する彩色を通じてクラスタリングを行うアルゴリズムを開発した．開発した手法を実データに適用して評価し，その有効性を確認した．

研究成果の概要（英文）：
We have developed a clustering method based on b-coloring of a graph in Graph Theory. This coloring requires that two kinds of constraints should be satisfied in the coloring of a graph. By regarding one kind of constraint as a local constraint and the other as a global constraint, we have developed a clustering method which reflects both constraints in a unified manner. For a given dataset, we define a graph structure based on the pairwise dissimilarities so that the coloring of the dataset can be conducted over the graph. The developed algorithm conducts the coloring of the graph for conducting the clustering of the dataset. The algorithm has been implemented as a prototype system, and experiments using the prototype system were conducted over several datasets. The results of the experiments indicate the effectiveness of the developed clustering algorithm.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,300,000	390,000	1,690,000
2009 年度	1,100,000	330,000	1,430,000
2010 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：グラフ・クラスタリング・彩色・制約

１．研究開始当初の背景

入手可能なデータの量や種類の爆発的な増加に伴い，データから規則性やパターンを半自動的に取り出す「データマイニング」に関する研究が近年活発に行われている．その際，膨大なデータ集合から個々のパターンを

発見するだけではなく，データの全体像を把握することにより，データ全体を俯瞰して活用するニーズが高まっている．これを実現する技術として，データ全体をいくつかの類似したグループ(各グループはクラスタと呼ばれる) に分割するクラスタリング手法がある．クラスタリング手法に関する研究は従来から統計，機械学習，データマイニングなどの様々な分野で研究が行われてきているが，大別すると，以下の２つのアプローチに分けられる．

局所的な観点からは， データ間やクラスタ間での距離に基づく手法がある．これらの手法では類似したデータやクラスタを併合などが可能であるが，大域的なクラスタ間非類似度を考慮することはできないという課題がある．

他方，大域的な観点からは， k-means 法に代表されるような最適化に基づく様々な研究がある．これらの手法はデータ全体としてのズレを極力少なくするような分割を得ることが可能であるものの，各データ対間での局所的な制約の反映は困難であるという課題がある．

上記のように，データのクラスタリングに対して局所的な観点と大局的な観点からそれぞれ様々な手法が提案されてきてはいたが，それぞれ個別に研究が進められることが多かった．このため，両者を統一的に扱うクラスタリング手法の開発が強く求められていた．

２．研究の目的

近年，グラフ理論の分野で，グラフの彩色に対して b 彩色という新しい概念が提案された．グラフの b 彩色とは，i) 頂点間での適正彩色，ii) 各色ごとでの支配頂点の存在，という２つの制約を満たすような頂点の彩色である．

一般にグラフの彩色においては，彩色に関する制約のもとで，同じ色を持つ複数の頂点が存在する．同じ色を持つ頂点は同じグループ（クラスタ）に属するとみなすことにより，グラフの彩色の観点からデータのクラスタリングを行うことが可能となる．たとえば，適正彩色に基づいてグラフにおけるクリークを同定し，クリークをクラスタと見なす研究なども提案されていた．しかし，このアプローチでは局所的な制約のみが考慮され，大域的な制約を反映してクラスタリングを行うことは困難であった．

このため，本研究では，上記での前者（適正彩色）を局所的な制約と捉え，後者（支配頂点の存在）を大域的な制約と捉えることにより，局所的な制約と大域的な制約の両者を同時に反映したクラスタリング手法を実現することに取り組む． それぞれの制約をグラフの彩色という統一的な観点から捉えることにより，両者を統一的に扱う手法の開発が可能になると期待される．

３．研究の方法

本研究では，グラフの彩色に基づくクラスタリングの基礎理論の確立と，この理論に基づくアルゴリズムの開発を目指した． 具体的には，
(1) b 彩色に基づくクラスタリングの基礎理論の確立，
(2) クラスタリングアルゴリズムの研究開発，
(3) 実データへの適用と評価，
の 3 項目に対する研究開発を行った． 具体的には，それぞれの項目に対して以下の課題に取り組んだ．

(1) b 彩色に基づくクラスタリングの基礎理論の確立

データ集合のクラスタリングを行うためには，まず，与えられたデータ集合をグラフ構造として表現する必要がある．とくに，b 彩色は無向グラフに対する概念であるため，与えられたデータ集合を無向グラフとして表現する必要がある．

このため，データ集合の表現形式を無向グラフとしてのグラフ表現に変換する手法の定式化に取り組んだ．さらに，グラフ理論における彩色問題に関する文献調査を行い，クリークや独立部分集合などと b 彩色との関係を考察した．

(2) クラスタリングアルゴリズムの研究開発

b 彩色における２つの制約を満たす最大の彩色数の決定は，一般には NP 困難な問題である．このため，与えられたデータ集合に対して構築したグラフにおいて，彩色数が最大となるような彩色（クラスタリング）アルゴリズムの開発に取り組むことは現実的ではない．

実用的な計算時間で動作するクラスタリングアルゴリズムを開発するために，(1)で行った考察，特に彩色数と頂点に接続する辺数の最大値との関係に基づき，辺数の最大値を活用する効率的なクラスタリングアルゴリズムを検討した．

グラフの彩色は一般に多大な計算量を要する処理である．このため，最初から「良い」クラスタリングとなるグラフの彩色を探索するのではなく，制約を満たす彩色に対して，制約を充足しつつクラスタの質を逐次的に改良する再彩色を行うアルゴリズムを検討した．

さらに，計算資源（使用可能な計算時間やメモリなど）に余裕がある場合には，計算資

源を積極的に活用して良いクラスタリングを得ることが重要となる．このため，彩色に対する探索空間を拡大してクラスタリングの精度を高めるアルゴリズムの開発に取り組んだ．

(3) 実データへの適用と評価

開発したアルゴリズムの有効性を評価するため，開発したアルゴリズムをプロトタイプシステムとして計算機上に実装する．次に，実装したプロトタイプシステムを機械学習の分野での標準的なベンチマークデータに適用し，他手法との比較実験を行い，開発した手法の有効性を確認する．

さらに，得られた実験結果をもとに，開発するアルゴリズムでの計算時間や挙動を調べ，開発したアルゴリズムの更なる改良を行う．

４．研究成果

(1) b 彩色に基づくクラスタリングの基礎理論の確立

与えられたデータ集合をグラフ構造（無向グラフ）として表現するため，データの表現形式をグラフ表現に変換する手法を定式化した．具体的には，与えられたデータの性質に応じて個々のデータ間での非類似度を測る非類似度関数が与えられると仮定し，非類似度に対する閾値に応じて，閾値を超える非類似度を持つデータ同士を辺で繋ぎ，その非類似度を辺ラベルとすることにより，データ集合を無向グラフに変換して表現する手法を提案した．

なお，提案法においては，非類似度関数および閾値は扱うデータの性質に依存して決めるべきものであり，データに応じて与えられると仮定している．この意味で，これらは提案法におけるハイパーパラメータに対応する．

(2) クラスタリングアルゴリズムの研究開発

彩色数と頂点に接続する辺数の最大値との関係に基づき，辺数の最大値から彩色数の上限を推定することが可能である．この性質に基づき，推定した上限の彩色数のもとで一旦適正彩色の制約を満たすように頂点を彩色し，その後に支配頂点の制約を満たすように色を付け替える，という２パスで全頂点を彩色するクラスタリングアルゴリズムを開発した．

上記の手法では，主に彩色における制約充足のみを考慮しているため，彩色の結果得られるクラスタの質（クラスタリングの精度）が良いとは限らない．この課題に対処するため，クラスタリング精度の観点から，これまでに提案されている様々な有効性指標を調査し，検討した有効性指標を活用できるように上記で開発したアルゴリズムを改良した．この結果，データ分割の質（クラスタリング精度に対応する）を単調に向上させることを保証しながら，色の変更（再彩色）を行うアルゴリズムを開発した．

さらに，上記で開発したアルゴリズムは，一旦彩色（あるいは再彩色）が行われた頂点の色は固定される，という意味で，欲張り探索に基づくアルゴリズムである．クラスタリングの精度を向上させるため，彩色に対する探索空間を拡大してより積極的に再彩色を行うアルゴリズムを検討した．この検討をもとに，再彩色に対する最良優先探索，および色交換を行うクラスタリングアルゴリズムを開発した．

(3) 実データへの適用と評価

上記(2)で開発したアルゴリズムを，Java 言語を用いてプロトタイプシステムとして計算機上に実装した．実装したプロトタイプシステムを，機械学習の分野での標準的なベンチマークデータとして用いられる UCI レポジトリでのデータセットに適用し，他手法との比較実験を行い，開発した手法の有効性を確認した．評価指標としては，クラスタリングの精度に対応する Precision（この指標は情報検索でも広く用いられている），クラスタ間のへだたりに対応する Distinctness，クラスタの質として用いられる一般化 Dunn 指標を評価した．

たとえば，UCI レポジトリでの zoo データセットに対しては，開発したアルゴリズムはクラスタリングの精度に対応する Precision Precision が 0.812 となり，kmeans 法(Precision=0.723)や EM 法(Precision=0.673)などの代表的な従来法を大きく上回る精度を示すことを確認した．このため，開発した手法は効果的であると言える．

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　4 件）

1. 吉田哲也，岡谷一宏：制約を反映するグラフ表現に基づく射影による半教師ありクラスタリング，情報処理学会論文誌：数理モデル化と応用，vol. 4, No. 1, 62-71 (2011)，査読有り

2. 吉田哲也：相互情報量に基づくクラスタリングに対するグラフモデルとその評価，情報処理学会論文誌：数理モデル化と応用，vol. 3, No. 3, 1-11 (2011)，査読有り

3. Hacid, H. and Yoshida, T.: Neighborhood Graphs for Indexing and Retrieving Multi-dimensional Data, Journal of Intelligent Information System, Vol.34,No.1 93-111 (2010), 査読有り

4. Yoshida, T., Elghazel, H., Deslandres, V., Hacid, M.S., and Dussauchoy, A.: Toward Improving b-Coloring based Clustering using a Greedy re-Coloring Algorithm, Greedy Algorithm, Bednorz, W. (Ed), chapter 29, 553-568 (2009), 査読有り

〔学会発表〕(計 7件)

1. Yoshida, T.: Toward Finding Hidden Communities based on User Profile, IEEE workshop on Social Interactions Analysis and Services Providers (SIASP), in conjunction with IEEE International Conference on Data Mining 2010, 380-387, December 13th, 2010, Sydney, Australia, 査読有り

2. Yoshida, T.: A Graph Model for Clustering based on Mutual Information, 11th Pacific Rim International Conference on Artificial Intelligence, LNAI 6230, 339-350, September 1st, 2010, Daegu, Korea, 査読有り

3. Ogino, H. and Yoshida, T.: Toward improving re-coloring based clustering with graph b-coloring, 11th Pacific Rim International Conference on Artificial Intelligence, LNAI 6230, 206-218, September 1st, 2010, Daegu, Korea, 査読有り

4. Yoshida, T. and Okatani, K.: A Graph-based projection approach for Semi-Supervised Clustering, 11th Pacific Knowledge Acquisition Workshop, LNAI 6232, 1-13, August 31st, 2010, Daegu, Korea, 査読有り

5. Yoshida, T.: Performance Evaluation of Constraints in Graph-based Semi-Supervised Clustering, The 2010 International Conference on Active Media Technology (AMT-2010), LNAI 6335, 138-149, August 29th, 2010, Tronto, Canada, 査読有り

6. Musaraj, K, Yoshida, T., Daniel, F., Hacid, M.S., Casati,F. and Benatallah, B.: Message Correlation and Web Service Protocol Mining from Inaccurate Logs, The IEEE 8th International Conference on Web Services, 259-266, July 8th, 2010, Florida, USA, 査読有り

7. Elghazel, H., Yoshida, T. and Hacid, M.S.: An Integrated Graph and Probability Based Clustering Framework for Sequential Data, The 11th International Conference on Discovery Science (DS-08), LNAI 5255, 246-258, October 15th, 2008, Budapest, Hungary, 査読有り

６．研究組織

(1)研究代表者

吉田　哲也　(TETSUYA　YOSHIDA　)
北海道大学・大学院情報科学研究科・准教授
研究者番号：80294164

(2)研究分担者

なし

(3)連携研究者

なし