科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ９ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５００２６９

研究課題名（和文）半教師付きクラスタリングの包括的研究と制約混合分布モデルへの挑戦

研究課題名（英文）Study of semi-supervised clustering and challenge to constrained mixture distributions

研究代表者
宮本　定明（Miyamoto, Sadaaki）

筑波大学・システム情報系・教授

研究者番号：６０１４３１７９

交付決定額（研究期間全体）：（直接経費）　　4,000,000円、（間接経費）　　1,200,000円

研究成果の概要（和文）：既存の半教師付き分類技法と申請者らが開発した技法とを比較し、その特質について考察した。主要な成果は次の通り（１）階層的技法に制約を加えると、既存技法で代表的な制約付き混合分布と性能がほぼ等しい。（２）ファジィ技法を拡張することで、基本的な半教師付き混合分布よりも一般的な定式化ができる。（３）ＣＯＰ　Ｋ－Ｍｅａｎｓを拡張することはできるが、性能は単純なＣＯＰ　Ｋ－Meansをはっきり上回ることはなかった。（４）インダクティブクラスタリングという半教師付きクラスタリングにおける新たな概念を提唱し、その有用性を示した。（５）Twitterなどの実テキストデータに適用し、半教師の効果を調べた。

研究成果の概要（英文）：Existing methods of semi-supervised clustering and proposed methods by the authors are compared using artificial and repository data. Main results are as follows. 1. Agglomerative clustering performs as well as mixture distribution models in constrained clustering. 2.Methods of fuzzy clustering generalize basic mixture distribution models for semi-supervised classification. 3. Extensions of COP K-means have been proposed but they did not perform as well as constrained mixture distribution method. 4. The concept of inductive clustering has been proposed and its methodological usefulness has been shown. 5. Real data using Twitter have been analyzed using semi-supervised clustering and its effects have been investigated.

研究分野：情報学

科研費の分科・細目：人間情報学・ソフトコンピューティング

キーワード：半教師付き分類　制約クラスタリング　階層的技法　混合分布モデル　COP K-means　ファジィクラスタリング　インダクティブクラスタリング

１．研究開始当初の背景

教師付き分類と教師なし分類（クラスタリング）は、対比的に捉えられるが、技法群の内容にはかなりな違いがみられる。教師付き分類において標準的な SVM や K-最近隣法をクラスタリングに用いることは可能であるが直接的ではない。一方、クラスタリングにおいて標準的な階層的技法は、教師付き分類には使えない。

最近注目されている半教師付き分類は、教師付き分類と教師なし分類の中間に位置している。従来開発されてきた技法は、教師付き分類とクラスタリングの両面から研究されてきているが、クラスタリングの観点からみると、いまだ確立した体系にはなっていない。半教師付き分類の代表的文献として、Chapelle et al.による半教師付き分類の集成や Basu et al.による制約クラスタリングの論文集が挙げられるが、それらは主に K-means や混合分布モデルにおける制約の導入と、正定値カーネルによる SVM 技法を中心に論じており、階層的クラスタリングの考察が欠けている。また、K-means をはじめとする最適化分類と混合分布モデルとの特質の比較や、アルゴリズムの検討も十分にはなされていない。

一方で、クラスタリングにおける主要な技法である階層的方法に制約を導入する研究もなされているが、個別的技法の考察と、現実的でない数学的議論にとどまっており、半教師付き分類における階層的技法の意義はいまだ明らかにされていない。さらに、ファジィクラスタリングにおける半教師の導入についても、個別的検討例があるのみで、全体像が明らかになってはいない。SVM については、インダクティブ SVM とトランスダクティブ SVM との差が論じられており、SVM にもとづくクラスタリング技法も提案されてはいるが、クラスタリング技法としては制限が多く、その意義は明確ではない。

２．研究の目的

申請者を中心とする研究グループは、ファジィクラスタリングの研究で国内外に知られているが、ファジィ技法にとどまるものではなく、クラスタリングの主要な諸技法を包括的に考察している点で、他の研究グループと一線を画している。これまで刊行した英文専門書（Miyamoto,1990, Miyamoto et al.,2008）と邦文専門書（宮本,1999）は入門書のレベルを超えて、クラスタリング研究の基本的観点と独自の理論的方向性を示している。この観点からみると、本研究における基本的動機付けとして、「クラスタリング技法を包括的に俯瞰する立場にたって従来の諸方法を見渡したとき、従来の研究より一段高いレベルの研究の可能性があるのではないか」という動機が生まれてくる。

その一方で、半教師付きクラスタリングは、Wagstaff らの COP-K means (ICML,2001) や Basu や Shental らの制約クラスタリング技法にとどめを指すという意見が既にあるかも知れない。この見方からすれば、申請者らは、階層的技法とファジィ技法、拡張 K-means 技法などを包括的に論じることによって、Basu らによる制約付き混合分布モデルに挑戦し、これを超えることが目的となる。

本研究では、クラスタリング技法が置かれている現実的状況（たとえば、階層的方法が今後も多くの分野で愛用されるであろうこと）に鑑み、次の基本的立場に則って研究を進める。

「クラスタリングの技法は教師付き分類とは異なり、混合分布モデルや SVM だけが中心となるのではなく、階層的技法、K-Means、ファジィクラスタリング、など性質の異なるモデルを融合的に用いるべきである。」

従って、半教師付き分類を研究するにあたっても、この立場に基づいて研究を進める。また同時に、次の点を明らかにすることが必要である。

「現在最もよく知られている制約付き混合分布モデルによるクラスタリングと本研究で提案する方法との得失の解明。」

これらをまとめると、（１）制約付き混合分布モデルへの挑戦、と（２）半教師付き手法の包括的考察による体系化、の 2 項目が研究目的となる。さらに、（３）これらの技法の特定のモデル（ファジィ近傍モデル）への応用、を検討する。

３．研究の方法

（１）制約付き混合分布モデルへの挑戦

①階層的クラスタリングアルゴリズムにおける対制約の導入と、数値例の計算による制約付き混合分布モデルとの性能比較

②拡張 K-means/fuzzy c-means 技法への対制約導入と制約付き混合分布モデルとの性能比較

（２）半教師付き手法の包括的考察による体系化

半教師導入の技法として（A）対制約導入（Must-Link, Cannot-Link ）による方法、(B) 距離修正法、(C) 目的関数修正法、などに注目し、次の諸技法の開発を目的とする。

①階層的クラスタリングアルゴリズムにおける対制約と距離修正法の提案

②拡張 K-means 技法と fuzzy c-Means における対制約、距離修正、目的関数修正技法の提案

③逐次クラスター抽出技法における半教師付きアルゴリズムの提案

（３）応用例としてのファジィ近傍モデルによる半教師付きテキストデータ解析技法の開発

４．研究成果

前項の各項目について、以下に得られた成果を要約して述べる。

（１）

① 階層的クラスタリングアルゴリズムにおける対制約の導入と、数値例の計算による制約付き混合分布モデルとの性能比較：

ペナルティ追加による対制約導入と距離修正法による対制約導入法を比較し、ペナルティ法が優っていることを数値例により確認した。次に、ペナルティ法を用いた階層的技法と制約付き混合分布モデル（Shental による方法）の効果を様々な数値例により比較し、全体的には、階層的技法は混合分布モデルに匹敵する性能を有すること、非線形性の強い数値例については、混合分布モデルを上回ることを実証した。

② 拡張 K-means/fuzzy c-means 技法への対制約導入と制約付き混合分布モデルとの性能比較

クラスターサイズ変数と共分散変数を含む拡張 K-means 技法について、制約付きクラスタリングアルゴリズムを提案し、制約付き混合分布モデルと比較した。その結果、数値例によっては制約付き混合分布モデルを上回る性能を発揮するが、初期値やパラメータ設定が難しく、全体的には制約付き混合分布モデルに匹敵するとはいえなかった。ただし、今後の研究により、適切な初期値・パラメータ設定法がみつかれば、性能が向上する可能性は認められる。一方、fuzzy c-means については、以下に述べるような拡張を行ったが、未だ満足すべき成果までは得られていないため、研究をさらに継続する必要がある。

（２）

① 階層的クラスタリングアルゴリズムにおける対制約と距離修正法の提案

ペナルティ追加による対制約導入と距離修正法による対制約導入法を標準的な結合法について比較検討した結果、主に重心法と Ward 法を利用することとなった。距離修正法では、正定値カーネルの使用が前提となるが、ペナルティ法では、カーネルは利用してもしなくても良く、方法論的にはペナルティ法がより優れていることがわかった。数値例を処理した結果についても、ペナルティ法が上回り、距離修正法は使用しないという結論となった。

② 拡張 K-means 技法と fuzzy c-Means における対制約、距離修正、目的関数修正技法の提案

制約付き拡張 K-means については、上記の通り、クラスターサイズ変数と共分散変数を含む方法に COP K-means のアルゴリズムを利用する方法を提案し、評価を行った。結果は、上記（１）②に既に述べた。この方法は、fuzzy c-means に利用することも可能であるが、クラスターをクリスプ化する必要があるため、計算量が増え、方法としての明晰さに欠けている。その一方で、ファジィ技法の良さはそれほどみられないため、考察を中止することとなった。さらに、fuzzy c-means における目的関数修正法については、基本的な半教師付き混合分布モデルを特殊なケースとして含むという理論的成果が得られた。しかしながら、基本的な半教師付き混合分布モデルよりも制約付き混合分布モデルのほうが性能が優るため、fuzzy c-means における目的関数修正法が実際に役立つケースは限られていると考えられる。Fuzzy c-means における距離修正法も試みたが、他の技法に比べて良い性能を示さなかったため、考察を中止した。

③ 逐次クラスター抽出技法における半教師付きアルゴリズムの提案

ノイズクラスタリングの技法を利用して、制約付き逐次クラスター抽出技法を提案し、簡単な数値例についてその効果を確認した。また、逐次回帰モデル抽出に同技法を適用し、制約の効果について考察し、アルゴリズムの提案を行った。

（３）応用例としてのファジィ近傍モデルによる半教師付きテキストデータ解析技法の開発

Twitter の実データを収集し、ファジィ近傍モデルに基づく COP K-means 技法および拡張 COP K-means 技法を主に利用してクラスターを抽出し、制約のある場合とない場合の相違について考察した。データ量はまだ小さいが、制約の効果がはっきりみてとれる場合がいくつかあった。

上記計画に挙げていない研究として、階層的技法が多量のデータに向かない点を改善するため、2 段階技法を提案した。第 1 段階では、COP K-means で多数のクラスターを生成し、第 2 段階では、制約付き重心法あるいは制約付き Ward 法を用いる。この方法により、多量のデータに対する制約付き階層的クラスタリングが可能となった。

また、非対称類似度データに対する階層的技法と制約付き技法についても考察を行い、提案手法の効果を調べた。

さらに、半教師付きクラスタリングに関係する重要な概念であるインダクティブクラスタリング(inductive clustering)を提唱し、その利用法と理論的有用性について示した。インダクティブクラスタリングとは、クラスタリング技法から分類器（classifier）が自然に得られる技法のことで、K-means や fuzzy c-means，混合分布モデルがこれにあたる。一方、階層的技法は産短距離法以外、この性質をもたず、インダクティブではない。

このように、初めに計画した事項についてすべて検討し、かつ計画に挙げていない事項についても検討した。階層的技法については、標準的な混合分布モデルに匹敵し、他の技法はまだ混合分布には及ばないという結論となったが、当該分野に関する一定の水準の研究成果が得られたと思われる。また、インダクティブクラスタリングの概念は今後のクラスタリング研究に重要な概念であり、さらに進んで考究すべきものと考えられる。Twitter など web 上の実テキストデータについても、ファジィ近傍モデルに対する制約による一定の効果が得られた。これについても解析と検討を継続したい。

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 20 件）

① S. Miyamoto, N. Obara, Algorithms of Crisp, Fuzzy, and Probabilistic Clustering with Semi-supervision or Pairwise Constraints, Proc. of 2013 IEEE International Conference on Granular Computing, 225-230, 2013（査読有）
② Y. Komazaki, S. Miyamoto, Variables for Controlling Cluster Sizes on Fuzzy c-Means, Lecture Note in Artificial Intelligence 8234, 192-203, 2013（査読有）
③ Hengjin Tang, Sadaaki Miyamoto, Semi-supervised Sequential Kernel Regression Models with Pairwise Constraints, Lecture Note in Artificial Intelligence 8234, 166-178, 2013（査読有）
④ Y. Tamura, N. Obara, S. Miyamoto, A Method of Two-Stage Clustering with Constraints Using Agglomerative Hierarchical Algorithm and One-Pass k-Means++, Advances in Intelligent Systems and Computing, 245, 2013, 9-19（査読有）
⑤ S. Miyamoto, S. Takumi, Journal of Advanced Computational Intelligence and Intelligent Informatics, 17, 2013, 504-510（査読有）
⑥ S. Miyamoto, S. Takumi, Inductive Clustering and Twofold Approximations in Nearest Neighbor Clustering, LNAI 7647, 355-366, 2012（査読有）
⑦ S. Takumi, S. Miyamoto, Comparing Different Methods of Agglomerative Hierarchical Clustering with Pairwise Constraints, Proc. of SCIS-ISIS 2012, 1545-1550, 2012（査読有）
⑧ N. Obara, S. Miyamoto, A Method of Two-Stage Clustering with Constraints Using Agglomerative Hierarchical Algorithm and One-Pass K-Means, Proc. of SCIS-ISIS 2012, 1540-1544, 2012（査読有）
⑨ Y. Komazaki, S. Suzuki, S. Takumi, S. Miyamoto, Analysis of Disaster Information on Twitter Using Different Methods of Clustering Based on a Fuzzy Neighborhood Model, Proc. of 5th International Symposium on Computational Intelligence and Industrial Applications, 2012, pp.1-5 (CD-ROM)（査読有）
⑩ S. Takumi, S. Miyamoto, Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering, Proc. of 2012 IEEE International Conference on Granular Computing, 2012, 542-547（査読有）
⑪ S. Miyamoto, S. Takumi, Hierarchical Clustering Using Transitive Closure and Semi-supervised Classification Based on Fuzzy Rough Approximation, Proc. of 2012 IEEE International Conference on Granular Computing, 2012, 438-443（査読有）
⑫ S. Miyamoto, S. Suzuki, S. Takumi, Clustering in Tweets Using a Fuzzy Neighborhood Model, Proc. of WCCI 2012 IEEE World Congress on Computational Intelligence, 2012, 251-256（査読有）
⑬ Hengjin Tang, Sadaaki Miyamoto, Sequential Regression Models with Pairwise Constraints Using Noise Clusters, Journal of Advanced Computational Intelligence and Intelligent Informatics, 16, 2012, 814-818（査読有）
⑭ S. Miyamoto, A. Terami, Inductive vs. Transductive Clustering Using Kernel Functions and Pairwise Constraints, Proc. of 11th Intern. Conf. on Intelligent Systems Design and Applications (ISDA 2011), 1258-1264,

2011（査読有）

⑮ S. Takumi, S. Miyamoto, Text Clustering Using a Multiset Model, Proc. of the 2011 IEEE International Conference on Granular Computing, 630-635, 2011 【Best Paper Award】（査読有）

⑯ S. Takumi, S. Miyamoto, Agglomerative Hierarchical Clustering Using Asymmetric Similarity Based on a Bag Model and Application to Information on the Web, Lecture Notes in Artificial Intelligence 7027, 187-196, 2011（査読有）

⑰ S. Miyamoto, A. Terami, Constrained Agglomerative Hierarchical Clustering Algorithms with Penalties, Proc. of 2011 IEEE International Conference on Fuzzy Systems, 422-427, 2011（査読有）

⑱ S. Takumi, S. Miyamoto, Agglomerative Clustering Using Asymmetric Similarities, Lecture Notes in Artificial Intelligence 6820, 114-125, 2011（査読有）

⑲ Y. Kanzawa, Y. Endo, S. Miyamoto, KL-Divergence-Based and Manhattan Distance-Based Semisupervised Entropy-Regularized Fuzzy c-Means Journal of Advanced Computational Intelligence and Intelligent Informatics, 15, 1057-1064, 2011（査読有）

⑳ H.J. Tang, S. Miyamoto, Sequential Extraction of Fuzzy Regression Models: Least Squares and Least Absolute Deviations, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 19, 53-63, 2011（査読有）

〔学会発表〕（計 19 件）招待講演 6 件含む

① Zhang Canlun, Sadaaki Miyamoto, Text Clustering of Chinese Documents Using Fuzzy Neighborhood, 第 40 回ファジィ・ワークショップ, 2014 年 3 月 7 日～8 日 首都大学東京, 南大沢, 東京.

② S. Miyamoto, Semi-supervised Clustering Using Hierarchical or Non-hierarchical Algorithms, IEEE International Conference on Granular Computing (invited talk), 2013 年 12 月 15 日, Beijing, China.

③ S. Miyamoto, Neighborhood System and Term-Text Clustering with Application to Mining Risks (invited talk), The 3rd International Workshop on Soft Computing and Disaster Control, 2013 年 11 月 9 日, Bali, Indonesia.

④ 鈴木昭平, 松崎慧太 , 宮本定明, ファジィ近傍モデルを用いたウェブテキスト解析, 第 29 回ファジィシステムシンポジウム, 2013 年 9 月 9 日～11 日, 大阪国際大学, 大阪.

⑤ 田村優友, 小原伸広, 宮本定明, One-pass k-means++を用いた対制約付き二段階階層的クラスタリング, 第 29 回ファジィシステムシンポジウム, 2013 年 9 月 9 日～11 日, 大阪国際大学, 大阪.

⑥ S. Miyamoto, H. Kondoh, Generalizations of K-Means Algorithms for Constrained Clustering, Intl. Conf. on Future Trends in Computing and Communication - FTCC 2013, 2013 年 7 月 13 日, Bangkok, Thailand.

⑦ 近藤晴香, 宮本定明, 松崎慧太, 半教師付き分類のための K-means 関連クラスタリング技法, 第 28 回ファジィシステムシンポジウム, 2012 年 09 月 11 日～2012 年 09 月 14 日, 名古屋工業大学、名古屋.

⑧ 小原伸広, 宮本定明, One-pass K-means を用いた対制約付き二段階階層的クラスタリング, 第 28 回ファジィシステムシンポジウム, 2012 年 09 月 11 日～2012 年 09 月 14 日, 名古屋工業大学、名古屋.

⑨ 宮本定明, 佗美怜, ファジィラフ近似を生成するクラスタリングと半教師付き分類, 第 28 回ファジィシステムシンポジウム, 2012 年 09 月 11 日～2012 年 09 月 14 日, 名古屋工業大学、名古屋.

⑩ S. Miyamoto, An Overview of Hierarchical and Non-hierarchical Algorithms of Clustering for Semi-supervised Classification, MDAI 2012 (invited talk) 2012 年 11 月 22 日, Girona, Spain.

⑪ S. Miyamoto, Statistical and Non-statistical Models in Clustering: An Introduction and Recent Topic (invited talk), JCS-CLADAG 2012, 2012 年 09 月 03 日, Capri Island, Italy.

⑫ S. Miyamoto, A Family of Methods for Asymmetric Nearest Neighbor Clustering, JCS-CLADAG 2012, 2012 年 09 月 03 日, Capri Island, Italy.

⑬ 鈴木昭平, 宮本定明, ファジィ近傍に基づいたテキストマイニングと Twitter への応用, 27th Fuzzy System Symposium, 2011 年 9 月 12 日 福井大学, 福井市.

⑭ 近藤晴香, 宮本定明, 階層的アルゴリズムと逐次的アルゴリズムによる制約クラスタリング, 27th Fuzzy System Symposium, 2011 年 9 月 12 日 福井大学, 福井市.

⑮ 佗美怜, 宮本定明, 非対称類似度を用い

た階層的クラスタリングのための平均結合法，27th Fuzzy System Symposium，2011年9月12日 福井大学，福井市．

⑯ S. Miyamoto, Inductive and Non-inductive Methods of Clustering, 2012 IEEE International Conference on Granular Computing （invited talk），2012年08月11日，Hangzhou，China．

⑰ S. Miyamoto, Agglomerative Clustering Using Asymmetric Measures without Reversals in Dendrograms, Fourth Japanese-Germen Symposium on Classification (JGSC2012), 2012年3月9日，同志社大学，京都市．

⑱ S. Miyamoto, Two Classes of Algorithms for Data Clustering, IUKM2011, (invited talk) 2011年10月27日 Zhejiang University, Hangzhou, China.

⑲ Hengjin Tang, Sadaaki Miyamoto, Semi-Supervised Sequential Regression Models with Pairwise Constraints, Modeling Decisions for Artificial Intelligence 2011 (MDAI 2011), 2011年7月28日，Empark Grand Hotel (Changsha, China).

〔その他〕
ホームページ等
http://www.soft.risk.tsukuba.ac.jp/miyamoto/

６．研究組織
(1)研究代表者
宮本　定明（MIYAMOTO, Sadaaki）
筑波大学・システム情報系・教授
研究者番号：６０１４３１７９

(2)連携研究者
遠藤　靖典（ENDO, Yasunori）
筑波大学・システム情報系・教授
研究者番号：１０２６７３９６