小 特 集

The 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD2006)

井手 剛
Tsuyoshi Idé
IBM 東京基礎研究所
IBM Research, Tokyo Research Lab.
goodidea@jp.ibm.com, http://www.trl.ibm.com/people/ide/

keywords: semi-supervised clustering, constraint, distance metric

1. クラスタリング界の若きヒロイン

昨年の 9 月にベルリンで開催された機械学習・データマイニングの分野での国際会議 ECML/PKDD 2006 *1 に参加した．名前上は二つの会議の併催のようになっているが，実質的に両者は不可分である．論文の競争率もなかなか高く*2，この分野での権威ある国際会議のひとつになっている．

この会議で私は，時系列データのクラスタリングに関する若干の結果を発表した [Idé 06]．私のセッションでは 4 件の発表があり，私以外の 3 件が半教師つきクラスタリングに関するものであった．

このセッションでひときわ光彩を放っていたのが，“Measuring constraint-set utility for partitional clustering algorithms”と題された講演であった [Davidson 06]．登壇したのは第 2 著者の Kiri Wagstaff である．事例レベルの拘束付きクラスタリングの創始者として，クラスタリング業界では彼女を知らぬ者はない．大げさに言えば，クラスタリング界の若きヒロインといってよいかもしれない．研究の中心地アメリカで，そうやって活躍してきたという自負もあるのだろう，自信に満ち溢れた彼女の様子が，場を彼女の招待講演の会場のように変えるのを，私は軽い羨望と共に眺めた．

2. ふたつのベストペーパー

さて，先に掲げた彼女の発表の題目をあえて日本語に訳すと，「クラスタリングにおいて**拘束条件の “使い甲斐” を測る**」とでもなろうか．機械学習の伝統的な分類では，クラスタリングは，教師なし学習の代表的なタスクとして知られている．しかし，ここ数年のトレンドは，クラスタリングを「半教師つき学習」として定式化すること，すなわち，ある種の事前知識を使ってクラスタリングの性能を向上させようとする試みである．

∗1 The 17th European Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases.

∗2 2006 年の投稿論文数は 685 本，受理率はフルとショートあわせて 16.5%の由．

このアイディアに先鞭をつけたのが Wagstaff らであった [Wagstaff 00, Wagstaff 01]．彼女らは，既存の階層型クラスタリングや k 平均クラスタリングを，事例の対についての事前知識を取り込めるように拡張した．すなわち，クラスタリングの対象とする標本の一部に，「この対は必ず同じクラスタに属する」「この対は同じクラスタにはならない」というような**拘束条件**が付いていると考え，たとえば k 平均アルゴリズムの近傍標本を探すステップに，拘束条件をチェックする部分を入れ込む．これはクラスタリングの正解が，標本の一部において事前にわかっているということを意味するから，直感的に考えて，拘束条件を入れれば入れるほど，クラスタリングの質は上がるはずである．Wagstaff らは，同一クラスタに属するという制約を**必須結合（must-link）**，属さないという制約を**禁則結合（cannot-link）**と呼んだ．

半教師つきクラスタリングは，目的関数に直接拘束を入れ込むもの（拘束式）と，距離計量の任意性を利用して事前知識を入れ込むもの（距離式）に 2 大別される．前者の代表例が Wagstaff らの拘束つき k 平均法：COP (constraint-partitioning) k-means であり，後者についての最初期の仕事のひとつが，Xing らの距離計量学習法 [Xing 03] である．2000 年以降になって急速に発展してきたこれらの 2 つの方向を，確率的枠組みに統合することを目指したのが，共著者 Basu らの 2004 年の仕事 [Basu 04] で，これは KDD ’04 の Best Research Paper に輝いている．一方，もうひとりの共著者 Davidson は，拘束つき k 平均アルゴリズムを計算量的に検討し，新たなクラスのアルゴリズムを提案した [Davidson 05]．この論文は SDM ’05 のベストペーパーになっている．

これら二つのベストペーパー，特に Basu の仕事は，半教師つきクラスタリングのその後の研究の流れに大きく影響を及ぼしただけでなく，クラスタリングという古典的なタスクが，半教師つきという場面設定を得て，機械学習の中心的なテーマのひとつになっていることを強くコミュニティーに印象付けた．このセッションでの Wagstaff の講演は，その立役者 3 名の共著であり，注目度満点だったのである．セッションに参加した大部分の人間は，彼女の講演を目当てに来たはずである．

表 1 COP k 平均クラスタリング [Wagstaff 01] において精度が悪化した試行の割合（原論文の Table 2 の一部を転載）.

Glass	Ionosphere	Iris	Wine
28%	26%	29%	38%

3. 事前知識がクラスタリングの質を悪くする

著者らはまず,（正解ラベル既知の）いくつかの標準データセットについて，拘束つきクラスタリングを 1000 回繰り返した．各回において拘束は，25 ペアの標本を無作為に取り出し，ラベルを確認することで生成する．

その結果，なんと，少なからぬ試行において，拘束を入れることによってクラスタリングの質（Rand Index で測る）が悪化していることがわかった．表 1 に結果例を載せる．もちろん，1000 回の平均でクラスタリング精度を比べれば，拘束つきの方がよい結果を出すのだが，実用上のシナリオでは，拘束のセットを無作為に生成するなどということは不可能で，ひとつ確定した拘束条件が与えられるわけであるから，表 1 の結果は深刻に捉えるべきである．

こうして著者らは問題をこのように述べる．なぜある種の拘束条件は（それが正解そのものであるにもかかわらず）クラスタリング精度を下げる方向に働くのだろうか．拘束つきクラスタリングが真に有用なものであるためには，拘束条件の「使い甲斐」とでも言うべきものを調べなければならない．その評価尺度を確立しなければならない．

4. 評価尺度：「珍しさ」と「一貫性」

やや天下り的な仕方で，著者らは，拘束を特徴付ける尺度として，**珍しさ**（informativeness）と，**一貫性**（coherence）という 2 つの量を定義する．珍しさというのは，拘束がない時のクラスタリング結果を前提にした時，それに比べて意外だという意味である．その場合，拘束を与えることによって，より有効な正解情報を得られるということであるから，珍しさの高い拘束は結果の改善に有効なはずである．式として表すと，クラスタリング手法 $\mathcal{A}$ の下での，拘束の集合 C の珍しさ $\mathcal{I}_{\mathcal{A}}(C)$ は

$$\mathcal{I}_{\mathcal{A}}(C) = \frac{1}{|C|}\sum_{c\in C}\mathrm{unsat}(c, P_{\mathcal{A}})$$

のようになる．ここで，$P_{\mathcal{A}}$ は $\mathcal{A}$ による無拘束下でのクラスタリングの結果を表し，unsat は，ひとつの拘束 c が，$P_{\mathcal{A}}$ における結果と異なっていた場合に 1 をとる関数である．

一方，一貫性の方は，$\mathcal{A}$ によらずに拘束の集合 C だけから決まる量で，必須結合と禁制結合の干渉の少なさを定量化したものである．たとえば図 1 のように，禁制結合と必須結合が，ある重なりを持って存在しているとする．著者によれば，両者の重なりは拘束同士に矛盾があることを示唆する．したがって，一貫性の高い拘束集合とは，この種の重なりが少ない集合のことである．

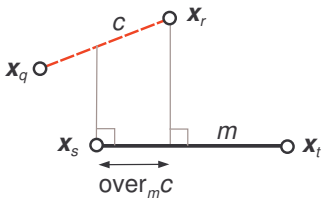

図 1 禁制結合 c と必須結合 m との重なりの定義．丸が標本を表す．4 つの標本に対し 2 つの拘束がある状況が描かれている．

C に含まれる必須結合の集合を C_{ML} と表し，禁制結合の集合を C_{CL} と表す．図 1 のようにして，互いの重なり $\mathrm{over}_m c$ を定義する（同様にして $\mathrm{over}_c m$ も定義できる）．この時，拘束集合 C の一貫性は，重なりの ない ものの割合として

$$\mathcal{COH}(C) = \frac{1}{|C_{\mathrm{CL}}||C_{\mathrm{ML}}|}\sum_{c\in C_{\mathrm{CL}}}\sum_{m\in C_{\mathrm{ML}}}\delta(c,m)$$

と定義される．ただし $\delta(c,m)$ は，$\mathrm{over}_m c = \mathrm{over}_c m = 0$ の時に 1 となり，それ以外は 0 となる関数である．

5. 実験的検証

珍しさと一貫性による特徴づけと，クラスタリング結果の関係を調べるために，著者らは大きく分けて 3 種類ほどの実験を行っている．ひとつは，表 1 同様に，25 個の拘束を無作為に取り出すことを 1000 回行い，その際，拘束の一貫性の高い上位 500 番までの結果について議論したものである．論文では 4 種類の拘束付きクラスタリング手法が比較されているのだが，表 1 と同じく COP k 平均の結果だけを表 2 に記す．表に示されているように，一貫性の高い拘束を課されたものはクラスタリングの精度が良いことがわかる．

もうひとつの実験は，3 つの拘束を無作為に取り出し，珍しさや一貫性をコントロールしながらクラスタリング結果を議論したものである．著者らは，一貫性が 1 の（つまり禁則結合と必須結合に全然重なりがない）拘束集合については，珍しさが高いほどクラスタリング精度が高いという一般的な傾向があると主張している．ただし実験結果には例外もあり，著者らの特徴づけが必ずしも完璧ではないことを示唆している．

最後の実験は，顔写真のデータを使って拘束同士の相互作用を視覚的に論じたものであるが，ここでは説明を省略する．

表 2 一貫性が高い拘束とそうでない場合との比較（原論文の Table 6 において COP k 平均の結果のみを転載）.

	Glass	Ionosphere	Iris	Wine
all	69.4	58.6	87.8	70.9
Top 500	70.4	59.3	88.3	71.5

6. む　す　び

残念ながら，著者らの実験結果は，総合的に見てかなり微妙である．珍しさと一貫性で拘束を特徴付けられるという著者らの結論は，実験により十分裏付けられているとは思えない．そもそも彼らの指標の定義自体に検討の余地は大いにある.「珍しさ」の方はある程度許せるとしても,「一貫性」については突込みどころ満載である．例えば，図 1 におけるクラスタ分けが $\{\{\boldsymbol{x}_r\},\{\boldsymbol{x}_q,\boldsymbol{x}_s,\boldsymbol{x}_t\}\}$ となっている場合と，$\{\{\boldsymbol{x}_q\},\{\boldsymbol{x}_r,\boldsymbol{x}_s,\boldsymbol{x}_t\}\}$ となっている場合を区別しないことの妥当性など，正直，疑問は尽きない．拘束つきクラスタリングに潜む魔物を，まだ十分に追い詰められていないということであろう．

論文自体は若干腰砕けになっている感が否めないものの，学習アルゴリズムを半教師つき学習の方向に拡張する際，何らかの危険の萌芽があるという主張は，十分面白いと思う[*3]．一歩先に進むために，不可思議な現象の「なぜ」を問おう——この論文のトーンはそういうものである．極端な実用主義者の中には,「なぜ」を問うことに意味を認めない者もいるが，研究をヒューリスティックスの塊にしてしまっては，学会を退屈にさせるばかりだと私は思う．このヒロインの（実質的に未完の）論文は私に，ある種の勇気も与えてくれた．むしろその意味で思い出に残る講演であった．

◇ 参　考　文　献 ◇

[Basu 04] Basu, S., Bilenko, M., and Mooney, R. J.: A probabilistic framework for semi-supervised clustering, in *Proc. ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining*, pp. 59–68 (2004)

[Davidson 05] Davidson, I. and Ravi, S. S.: Clustering With Constraints: Feasibility Issues and the k-Means Algorithm, in *Proc. SIAM Intl. Conf. Data Mining (SDM)*, pp. 138–149 (2005)

[Davidson 06] Davidson, I., Wagstaff, K. L., and Basu, S.: Measuring Constraint-Set Utility for Partitional Clustering Algorithms, in *Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pp. 115–126 (2006)

[Idé 06] Idé, T.: Why does subsequence time-series clustering produce sine waves?, in *Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pp. 311–322 (2006)

[Wagstaff 00] Wagstaff, K. and Cardie, C.: Clustering with Instance-level Constraints, in *Proc. Intl. Conf. Machine Learning (ICML)*, pp. 1103–1110 (2000)

[Wagstaff 01] Wagstaff, K., Cardie, C., Rogers, S., and Schroedl., S.: Constrained K-means Clustering with Background Knowledge, in *Proc. Intl. Conf. Machine Learning (ICML)*, pp. 577–584 (2001)

[Xing 03] Xing, E., Ng, A., Jordan, M., and Russell, S.: Distance metric learning, with application to clustering with side-information, in *Advances in Neural Information Processing Systems*, Vol. 15, pp. 505–512 (2003)

[新納 06] 新納浩幸, 佐々木稔, 村上浩司：制約を修正に用いた半教師有りクラスタリング, 第９回情報論的学習理論ワークショップ（IBIS 2006), pp. 77–82 (2006)

[神嶌 06] 神嶌敏弘：教師ありクラスタリングと絶対/相対クラスタリング, 第９回情報論的学習理論ワークショップ（IBIS 2006), pp. 83–88 (2006)

*3 直接関係する和文文献としては [新納 06] を挙げておく．また，クラスタリング諸種法の分類については [神嶌 06] が日本語で読める優れた文献である．

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理

著　者　紹　介

井手　剛(正会員)

苫小牧工業高等専門学校機械工学科，東北大学工学部機械工学科を経て，2000 年東京大学大学院理学系研究科・物理学専攻博士課程修了．同年より IBM 東京基礎研究所研究員．博士（理学）．2004，2006 年人工知能学会全国大会優秀賞．実数値時系列データからの知識発見技術に興味を持つ．