1C1-3

オブジェクト識別におけるクラスタ数決定方式

Determining the Number of Clusters in Object Identification

小山 聡　　田中 克己
Satoshi Oyama　Katsumi Tanaka

京都大学大学院 情報学研究科 社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

We deal with the problem of determining the number of clusters in object identification. In the case of disambiguating abbreviated names in the DBLP data set, we show that the number of clusters (full names) for each abbreviated name obeys a power law distribution, which indicates that using the average number of clusters as a fixed parameter in clustering is not appropriate. We propose building a model for predicting the number of clusters in a given data set. Several statistics of a data set to cluster are introduced as explanatory variables used in the model. We conduct experiments using a linier regression model as well as an improved two-stage model.

1. はじめに

オブジェクト識別は，文書集合やデータベース中に現れる名前が実世界の同じオブジェクトに対応しているか否かを判定する問題である．この問題は，曖昧な名前を含むデータ（文書やデータベースレコード）をクラスタリングし，同一クラスタに含まれるデータを同じオブジェクトに対応すると判定することで解決されることが多い [Oyama 06]．

様々なクラスタリング手法の中で最もよく用いられるものとして，K-Means 法と階層的クラスタリング法がある．K-Means 法はクラスタ数をあらかじめパラメータとして設定する必要がある．Single Linkage 法を含む階層的クラスタリング法は，階層的なクラスタリング木を結果として出力する．木によるクラスタの表現は，文書集合をトピックの類似度に従って分類するような場合には適している．しかし，オブジェクト識別の場合には，データが同一のオブジェクトを参照しているか否かを明確に判定することが目標であり，クラスタ数を決定する必要がある．そのため，階層的なクラスタリング法の場合でも，クラスタ数や閾値をパラメータとして与えて，あるクラスタ数でクラスタリングを停止させることが必要となる．

しかし，オブジェクト識別において，最適なクラスタ数を事前に決定することは容易ではない．その理由として，正解クラスタ数（オブジェクト数）が問題毎に大きく異なることがある．以下では例として，引用データベースにおける省略された著者名の曖昧性解消の問題を挙げる．これは，ファーストネームがイニシャルに省略された名前（例えば，D. Johnson や J. Smith）に対して，著者を対応させる問題である．計算機科学分野の文献データベースである DBLP データは多くの著者名がフルネームで与えられており，異なるフルネームは異なる人物であると仮定して，オブジェクト識別問題におけるテストコレクションとして用いられる．

図 1 は 2003 年の DBLP データにおける雑誌論文と国際会議論文の著者において，省略された著者名がいくつのフルネームの著者名に対応しているかを示したものである．(論文数は 448056 件，異なるフルネームの著者名は 300167 件，異なる省略名は 200251 件であった.）縦軸横軸ともに対数目盛が用いられていることに注意したい．データ点は両対数グラフにおいて直線に載っているように見える．これは，データが冪乗分布

連絡先: 〒 606-8501 京都市左京区吉田本町 京都大学大学院 情報学研究科 社会情報学専攻 小山 聡

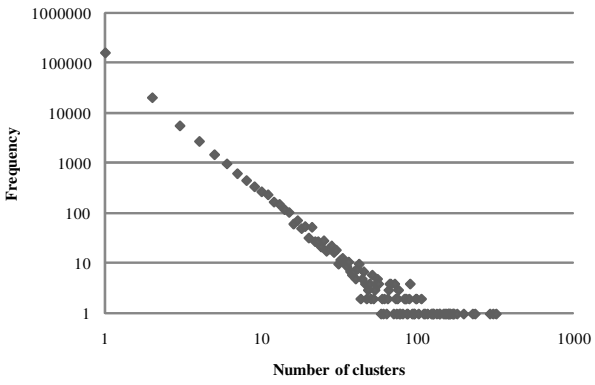

図 1: 1 つの省略名に対する著者名数の分布

に従っていること，すなわち，大きな割合の省略名において，正解クラスタ数（対応するフルネーム数）は小さい一方で，非常に多くのクラスタが対応する省略名も存在することを示している．実際，82%の省略名には単一のフルネームしか対応しない一方で，“J. Lee” という省略名に対しては，316 もの異なるフルネームが存在する．

冪乗分布のような歪んだ分布を示すデータにおいては，データを平均値で代表させることは適切ではない．例えば，この DBLP データにおける省略名あたりのフルネーム数（クラスタ数）の平均は約 1.5 であるが，全ての省略名に対して，常にクラスタ数のパラメータを 1 （クラスタリングを全く行わない）や 2 に設定することは，非常に大きなクラスタ数を持つ省略名を無視することになり，有効ではない．

2. クラスタ数予測モデルの構築

以上の理由から本研究では，与えられたオブジェクト識別（クラスタリング）対象のデータ集合における，オブジェクト数（クラスタ数）を予測するモデルを構築することを試みる．すなわち，$S = \{d_1, d_2, \ldots, d_{|S|}\}$ をクラスタリング対象のデータ集合とした時，

$$y = f(S)$$

の形でクラスタ数 y を予測する．

具体的には，オブジェクト数（正解クラスタ数）が既知のデータ集合 S^j とそのクラスタ数 y^j のペアの集合 $T = \{(S^1, y^1), (S^2, y^2), \ldots, (S^{|T|}, y^{|T|})\}$ を教師付き学習の訓練集合として用い，未知のデータのクラスタ数を予測するモデルを構築するアプローチを採る．

クラスタ数を決定する既存の方法として，クラスタの自然さを表すヒューリスティックな指標を導入し，その指標の値を最大化するクラスタ数を求めるものがいくつか提案されている [Milligan 85]．この他に，ギャップ統計量という指標を用いて最適なクラスタ数を決定する方式も提案されている [Tibshirani 01]．いずれの方式においても，全てのクラスタ数でクラスタリングを行い，指標の値を計算する必要がある．また，自然なクラスタが，必ずしもオブジェクト識別のような特定な問題でのクラスタに対応しているとはかぎらない．

クラスタ数を外からのパラメータとして与えずに自動的に決定するクラスタリング方式として，相関クラスタリングが提案されている [Bansal 04]．相関クラスタリングにおいては，同じクラスタに属するべきデータ対，異なるクラスタに属するべきデータ対が与えられたときに，ある評価関数を最適化するように，分割を決定し，その結果としてクラスタ数が決定される．しかしながら，問題設定の一般性やアルゴリズムの単純さなどの点から，依然として K-Means 法や階層的クラスタリング方が幅広く用いられている．

我々のアプローチの利点をまとめると以下になる．

- モデルを用いてクラスタ数を予測することで，クラスタ数が 1 のデータ集合に対して，不要なクラスタリングを避けることができ，計算負荷を低減することができる．これは，特に冪乗分布のように，クラスタ数 1 のデータ集合が多くの割合（DBLP データでは 82%）を占める場合には効果的である．
- クラスタ数をパラメータとするヒューリスティックな指標で最適なクラスタ数を決定する従来の方式では，全てのクラスタ数でクラスタリングを行う必要があったが，提案手法ではその必要がない．
- オブジェクト識別問題に固有の特徴を反映した，精度の高いモデルを構築することが可能になる．

最後の点に関しては，次章で DBLP データにおける省略著者名の曖昧性解消問題を例として，問題の特徴を説明変数としてモデルに反映させることを説明する．

以降では，まず最もよく用いられるモデルの一つである線形回帰モデルを用いる．線形回帰モデルでは，以下の式を用いてクラスタ数 y を予測する．

$$y = \sum_{n=1}^{N} w_n x_n + w_0 \quad (1)$$

ここで，x_n は説明変数であり，与えられたデータ集合を特徴づける変数を用いる．w_n は各変数に対する重みである．

重みを決定する方法としては，良く知られた最小自乗法を含めていくつか存在する．今回は，サポートベクトル回帰 [Drucker 97] を用いる．

3. モデル構築に用いる説明変数

我々は，データ集合（例えば，同一省略名による論文の集合）が与えられた時に，これがいくつのクラスタ（異なる人物の論文）からなるかを予測したい．そのため，説明変数として，個々のデータ点（論文）ではなく，データ集合（同一省略名による論文の集合）の特徴を表すものを用いる．今回は DBLP データを例に，説明変数としてクラスタ数に影響を与えると考えられるものを複数導入する．表 1 にモデル構築に用いる説明変数を示す．説明変数は大きく 2 つのグループに分けられる．1 つは上に述べた，クラスタリング対象データ集合の統計量である．

3.1 データ数

クラスタリング対象データ集合の統計量の最も基本的なものとして，クラスタリング対象のデータ数そのものがある．すなわち，変数 NumPapers は対象の省略名を含む論文数である．これは，データ数自体が，クラスタ数と関連を持つという予想に基づいている．ある省略名での文献数が 500 件もあれば，1 人の人物ではなく，多数の人物が存在する可能性が高い．逆に，論文数が少なければ少数の人物しか対応しない．極端な場合，ある省略名での論文が 1 件しかなければ，1 人の人物しか存在し得ない．

3.2 属性値の多様性

データの属性値の多様性を示す変数をいくつか用意した．NumCoauthors は，クラスタリング対象データ集合に，いくつの異なる共著者名（共著者名も省略名で表される）が含まれるかを表したものである．これは，多くの異なる人物との共著論文がある場合，オブジェクト識別対象の省略名も多くの人物を含む可能性が高いという予想に基づいている．また，NumTitleWords は，論文のタイトルに現れる異なる単語の数である．タイトルに現れる単語の種類が多いことは，論文のトピックの多様性を表すと考えられ，一般的には多くの人物が存在することを示すと考えられる．NumJournal は論文が掲載された異なるジャーナル／国際会議の数である．通常同じ人物は限られた数のジャーナルや国際会議で論文を発表する傾向が高いため，ジャーナル／国際会議の種類が多いほど，異なる人物が存在する可能性が高いと考えられる．

3.3 時間的な多様性

データの時間的な多様性もオブジェクト識別において有用な指標である．例えば，最初の論文と最近の論文の出版年が 50 年も違っていた場合，異なる人物の論文である可能性が高い．YearRange は最も新しい論文と最も古い論文の出版年の差を表す変数である．YearSTD は論文の出版年の標準偏差を表す．

3.4 名前の頻度

以上の変数はクラスタリング対象のデータ集合から導かれる統計量であった．しかしながら，クラスタリング対象データ以外の統計量からも，クラスタ数決定に有効な情報が得られる可能性がある．特に人名の曖昧性解消においては，人名そのもののデータベース全体での頻度も重要な情報源となる．

例えば，“Tanaka” というラストネームは，“K. Tanaka” や “S. Tanaka” など，様々な省略名で現れる．これは，“Tanaka” というラストネームが一般的であることを示しており，“K. Tanaka” に複数の人物が対応する可能性が高いことを示す．一方，“Kitsuregawa” というラストネームは “M. Kitsuregawa” という名前にしか現れない珍しいラストネームであり，これは “M. Kitsuregawa” が単一の人物の可能性が高いと推測する材料となる．以上の理由から，全データセット中に現れる全ての省略名でのラストネームの頻度を表す LastFreq を説明変数として導入する．

また，ファーストネームのイニシャルも，アルファベットによって大きく頻度が異なる．最も頻度の多い M は全体の約

クラスタリング対象データ集合の統計量	データ数	NumPapers	対象省略名を含む論文数
	属性値の多様性	NumCoauthors	異なる共著者の数
		NumTitleWords	タイトルに現れる異なる単語の数
		NumJournals	異なるジャーナル／国際会議の数
	時間的な多様性	YearRange	最も新しい論文と最も古い論文の出版年の差
		YearSTD	論文の出版年の標準偏差
クラスタリング対象データ集合以外の統計量	名前の頻度	LastFreq	全省略名中でのラストネームの頻度
		InitialFreq	全省略名中でのイニシャルの頻度

表 1: モデルに用いる説明変数

10%の省略名で用いられるのに対し，最も頻度の少ない Q は全体の 0.1%の省略名でしか用いられない．ここから例えば，Q はイニシャルとして用いられることが稀なアルファベットであり，Q がイニシャルの場合，多くの人物が対応する可能性は低いと推測することが可能となる．InitialFreq は全データセット中に現れる全ての省略名でのイニシャルの頻度を表す変数である．イニシャルの頻度の値をそのままを用いると，ラストネームの頻度にくらべて全体的に大きな値になり，属性値の大きさに不均衡が生じる．そこで，相対頻度をパーセント表示した値を実際には用いている．

3.5 説明変数の値の例

表 2 にいくつかの省略名におけるクラスタ数と説明変数の値の例を示す．多くのクラスタ数が対応する “J. Lee” や “K. Tanaka” のような省略名に対しては，各説明変数の値が大きく，1 つのクラスタ数しか対応しない “V. Zissimopoulos” のような名前に対しては，説明変数の値が小さい傾向がある．もちろん，例外もあり，例えば “M. Kitsuregawa” の例では多くの説明変数が大きな値であるが，実際のクラスタ数は 1 である．式 (1) における各変数の重み w_n を適切に調整することで，クラスタ数を説明できるモデルを構築することが必要なことが分かる．

4. 実験

1 章で説明した DBLP データセットから，2000 件の省略名をランダムに抽出して実験を行った．ただし，論文が 1 件しかない省略名はクラスタ数が 1 であることが自明なので，用いていない．それぞれの省略名を著者として含む論文の集合を取得し，表 1 の説明変数の値を求めた．2000 件の省略名をランダムに 10 個の部分集合に分割し，90%のデータでモデルを作成し，残り 10%のデータで評価する 10 分割交差検定を行った．モデル作成においては，サポートベクトル回帰を実装している SVM^{light}[Joachims 99] を用いた．評価指標としては，正解クラスタ数と予測クラスタ数の自乗誤差の平均を以下の式により計算して用いた．

$$\frac{1}{M}\sum_{m=1}^{M}\left(y^m_{\text{answer}} - \lfloor y^m_{\text{prediction}} + 0.5 \rfloor\right)^2$$

ここで，M はテスト集合における問題（省略名）の数，y^m_{answer} は省略名 m の正解クラスタ数，$y^m_{\text{prediction}}$ は省略名 m のモデルによる予測クラスタ数である．式 (1) による予測値は実数値であるが，クラスタ数は自然数でなければ意味がないため，値を四捨五入で整数値にしている．

表 3 に線形モデルによる予測誤差を示す．各試行ごとの誤差のばらつきが大きいことが分かる．2 回目の試行で特に誤差が大きい理由として，この試行のテスト集合に 55 という大きな正解クラスタ数を持つ省略名があり，この問題のクラスタ数を低く予測したことによる大きな誤差が全体に影響を与えていることが考えられる．

試行	誤差	試行	誤差
1	2.275	6	1.185
2	11.445	7	1.510
3	1.920	8	2.315
4	3.040	9	2.760
5	1.425	10	5.220
平均	3.310		

表 3: 線形モデルによる予測誤差

全体的な予測精度を向上させるには，大きなクラスタ数を持つ問題に対する予測を正確に行う必要があるが，訓練集合において多くの例題がクラスタ数 1 であり，クラスタ数が 2 以上の例題は少ない．モデルがクラスタ数 1 の例題に過適合すると，全体的に低いクラスタ数を予測するようになり，多クラスタの問題の予測精度が悪くなることが考えられる．このような場合，テスト集合でクラスタ数の大きな問題が与えられると大きな誤差を出し，全体の誤差に影響を与えることになる．

5. クラスタ数予測のための 2 段階モデル

1 クラスタの例題数と多クラスタの例題数の不均衡による，多クラスタデータ集合に対する予測精度低下の問題を解決するため，以下のような 2 段階のモデルを考える．

1. 与えられたデータ集合が 1 クラスタからなるか，複数クラスタからなるかを判別するモデル
2. 複数クラスタと判別されたデータ集合に対して，クラスタ数の予測を行うモデル

このような 2 段階のモデル化は，生態学の分野で生物種の生息数を予測するモデルを構築する際に用いられている [Fletcher 05]．生物種の生息数は多くの地理領域では 0 であるが，少数の領域では数多くの個体が生息するという歪んだ分布をしている．前述のように，オブジェクト識別においてもクラスタ数は冪乗分布という歪んだ分布をしており，クラスタ数 1 の問題と 2 以上の問題を別々にモデル化することで，予測の精度を上げることができると考えられる．

我々は，元の訓練集合から，クラスタ数が 1 か 2 以上かをラベルとする 2 値ラベルの訓練集合と，クラスタ数 2 以上のデータだけを取り出した整数値ラベルの訓練集合の 2 つの訓

省略名	K. Tanaka	J. Lee	V. Zissimopoulos	M. Kitsuregawa
クラスタ数	28	316	1	1
NumPapers	183	910	10	110
NumCoauthors	216	982	7	66
NumTitleWords	637	2249	51	370
NumJournals	98	395	6	59
YearRange	34	31	11	21
YearSTD	7.59	4.24	2.97	5.35
LastFreq	16	24	1	1
InitialFreq	4.15	8.73	2.01	9.99

表 2: クラスタ数と説明変数の値の例

試行	誤差	試行	誤差
1	1.500	6	0.980
2	6.560	7	1.365
3	1.650	8	1.665
4	1.775	9	1.575
5	1.265	10	3.875
平均	2.221		

表 4: 2 段階モデルによる予測誤差

練集合を作成した．前者の訓練集合をサポートベクトルマシンへの入力としてクラスタ数が 1 か 2 以上かを判別するモデルを構築し，後者の訓練集合をサポートベクトル回帰への入力としてクラスタ数の予測モデルを構築した．

構築した 2 つのモデルをテスト集合に対して 2 段階に適用することでクラスタ数の予測を行った．線形モデルと同様の手法で誤差を計算した結果を表 4 に示す．線形モデルの場合と比べて，誤差が軽減されていることが分かる．

6. おわりに

本稿では，オブジェクト識別におけるクラスタ数決定問題について述べた．例として文献データベースにおける省略された著者名の曖昧性解消問題を取り上げ，省略名あたりのクラスタ数が冪乗分布をすることを示した．冪乗分布においては平均値で全体を代表させることは適切ではなく，平均的なクラスタ数をパラメータとして事前に設定するのではなく，個々のデータ集合に応じてクラスタ数を決定する必要がある．

我々は与えられたデータ集合のクラスタ数を予測するモデルを構築することを提案し，著者名の曖昧性解消問題において，クラスタ数の予測に有効と予想されるいくつかの統計量をモデルの説明変数として導入した．線形回帰モデルを用いた実験の結果，訓練集合が多数の 1 クラスタの例題と少数の複数クラスタの例題からなることが，モデルの精度を下げる結果になっていることが考えられた．そこで，データ集合が複数のクラスタからなるか否かを判別するモデルと，複数クラスタの場合にそのクラスタ数を予測するモデルを別々に構築する 2 段階のモデル化を提案し，実験によりその有効性を示した．

今後の研究課題として，各説明変数の予測精度への寄与の分析，他の有効な説明変数の考察，およびモデルを用いないヒューリスティックなクラスタ数決定方式との精度の比較を行うことを考えている．

謝辞

本研究の一部は，文部科学省科学研究費補助金（課題番号 18049041，19700091）および文部科学省「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」プロジェクトによる．ここに記して謝意を表します．

参考文献

[Bansal 04] Bansal, N., Blum, A., and Chawla, S.: Correlation Clustering, *Machine Learning*, Vol. 56, No. 1–3, pp. 89–113 (2004)

[Drucker 97] Drucker, H., Burges, C. J., Kaufman, L., Smola, A., and Vapnik, V.: Support Vector Regression Machines, in *Advances in Neural Information Processing Systems*, Vol. 9, pp. 155–161 (1997)

[Fletcher 05] Fletcher, D., MacKenzie, D., and Villouta, E.: Modelling Skewed Data with Many Zeros: A Simple Approach Combining Ordinary and Logistic Regression, *Environmental and Ecological Statistics*, Vol. 12, No. 1, pp. 45–54 (2005)

[Joachims 99] Joachims, T.: Making Large-Scale SVM Learning Practical, in Schölkopf, B., Burges, C., and Smola, A. eds., *Advances in Kernel Methods : Support Vector Learning*, pp. 169–184, MIT Press (1999)

[Milligan 85] Milligan, G. W. and Cooper, M. C.: An Examination of Procedures for Determining the Number of Clusters in a Data Set, *Psychometrika*, Vol. 50, No. 2, pp. 159–179 (1985)

[Oyama 06] Oyama, S. and Tanaka, K.: Learning a Distance Metric for Object Identification without Human Supervision, in *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD2006)*, pp. 609–616 (2006)

[Tibshirani 01] Tibshirani, R., Walther, G., and Hastie, T.: Estimating the Number of Clusters in a Dataset via the Gap. Statistic, *Journal of the Royal Statistical Society* **B**, Vol. 63, No. 2, pp. 411–423 (2001)