自然言語処理 117－15
（1997. 1. 21）

最良パス探索を用いた凝集型クラスタリングアルゴリズム

本田 岳夫, 奥村 学
Email:{honda,oku}@jaist.ac.jp
北陸先端科学技術大学院大学 情報科学研究科

[概要]
従来の凝集型クラスタリングでは、クラスの併合の順序により、より適切と思われる分類があるにも関わらず、異なった分類がなされることがあるため、併合の順序という文脈を考慮してクラスタリングする必要がある。また、クラスタリングをする時、外部からアイテム間の類似性の情報などの文脈が与えられた時、その文脈を考慮したクラスタリングを行なう必要がある。

本稿では、クラスタリングをクラスの併合の状態の系列ととらえ、併合の状態系列をViterbiアルゴリズムを用いて推定する手法を提案する。本手法では、ある併合のステップで、それまでの併合の系列により尤もらしい併合を割り当てることによって併合の順序の文脈を考慮する。また、状態の連接コストに外部からの文脈を導入することが可能である。
[キーワード] クラスタリング, シソーラス, Viterbiアルゴリズム

Bottom-up Clustering Algorithm using Best First Search

HONDA Takeo,OKUMURA Manabu
School of Information Science, Japan Advanced Institute of Science and Technology
(Tatsunokuchi Ishikawa 923-12 Japan)

Abstract

In this paper, we present a new bottom up clustering algorithm using a kind of best first search, Viterbi algorithm. Our algorithm regards the step in which two classes are merged into a class as one state. Clustering corresponds to selecting the most likely sequence of merging state. As each merging state is assign according as the sequence before current state, our algorithm can consider the context that order of merging steps condition on quality of hierarcy. We can add the context ,from external clustering criteria, that a certain class is similar to other classes into connection cost.

Key Words clustering, thesaurus, Viterbi Algorithm

1 はじめに

クラスタリング [2] は、類似した対象をグループに分けることであり、自然言語処理では、シソーラスの獲得 [10, 9, 8, 6]、テキスト分類 [3], 括弧つきコーパスからの文脈自由文法の非終端記号獲得 [5] などに用いられている。

クラスの階層を獲得する階層型クラスタリングでは、アイテム1つ1つが1つのクラスである状態からはじめて、最終的に1つのクラスにまとめあげる凝集型クラスタリングが多く用いられている。凝集型クラスタリングでは、その時点で一番類似度の高い2つのクラスを1つにまとめる操作を繰り返し、デンドログラム (樹状図) を作る。アイテムは一度クラスにまとめられたら別のクラスに併合されることがないので、デンドログラム全体を眺めた時に併合の順序によっては、より適切と思われるクラスがあるにも関わらず、異なったクラスに分類されてしまうことがある。

また例えば、単語「ワープロ」「鉛筆」「テレビ」「万年筆」「ビデオデッキ」をクラスタリングすることを考える。この時、「ワープロと鉛筆は類似している (書くための道具である)」という情報がクラスタリングする時の文脈として外から与えられていれば、「ワープロ」「鉛筆」「万年筆」が含まれるクラスがクラス階層の下位の方で生成されることが望ましく、「ワープロとテレビは類似している (電化製品である)」という文脈が与えられれば、「ワープロ」「テレビ」「ビデオデッキ」が含まれるクラスが早期に生成されることが望ましい。Kojima らは単語間の距離を重みつきユークリッド距離で計算し、文脈が与えられた時の重みを動的に変化させる手法を提案している [4]。

このようにクラスタリングを行なう時には、クラスを併合する順序としての文脈と、外から与えられる文脈を考慮できることが望まれる。本稿では、これら2つの文脈を考慮したクラスタリングアルゴリズムを提案する。本アルゴリズムは、2つのクラスを併合するステップの系列と見ることで、可能な系列のパスの中から最良なパスを推定する。最良なパスの探索には Viterbi アルゴリズム [7] を用いる。系列中の1つの状態のコストは併合しようとしているクラス間の距離がそのまま使える。Viterbi アルゴリズムはラティス状のパス集合の中から最適なパスを推定するアルゴリズムであるので、望ましい順序でクラスを併合するようなクラス階層が生成することが可能になる。また、状態間の連接のコストに、外から与えられる文脈を反映させることができるので、外からの文脈を導入することが可能である。

第2節で、従来の階層クラスタリングについて説明し、第3節で、我々の提案するアルゴリズムについて説明する。第4節で、本アルゴリズムの使用例として、動詞-名詞の共起関係から名詞の概念階層を生成する。

2 階層クラスタリングアルゴリズム

階層型クラスタリングアルゴリズムはアイテムを階層的なクラスに分類するアルゴリズムで、分割型と凝集型に分かれる。分割型クラスタリングは、すべてのアイテムを含む1つのクラスからはじめて、順次分割していく手法である。凝集型クラスタリングは、1つ1つのアイテムからはじめて最終的に1つのクラスにまとめる手法である。凝集型クラスタリングアルゴリズムは、次のようになる。

1. それぞれのアイテムに対してその1つのアイテムからなるクラスを作る。
2. クラスの数が1つになるまで以下を繰り返す。
   - **2-1** それぞれのクラス間の類似度 (もしくは距離) を計算する。
   - **2-2** もっとも類似度の高い (もしくはもっとも距離の小さい) クラスペアを併合して1つのクラスとする

ここでクラス間の類似度もしくは距離が必要となる。クラス間の類似度 (距離) は、2つのアイテム間の類似度 (距離) から計算する方法と、直接クラス間の類似度を計算する方法がある。

2つのアイテム間の類似度 (距離) には、次のものが良く用いられる。ここで、2つのアイテムを $I_r = (I_1^r, \ldots, I_n^r), I_s = (I_1^s, \ldots, I_n^s)$ とし、2つのアイテム間の類似度を $\mathrm{sim}(I_r, I_s)$, 距離を $\mathrm{dis}(I_r, I_s)$ とする。

$$\mathrm{dis}(I_r, I_s) = \sum_{i=1}^{n} (I_i^r - I_i^s)^2$$

$$\mathrm{dis}(I_r, I_s) = \sum_{i=1}^{n} |I_i^r - I_i^s|$$

$$\mathrm{sim}(I_r, I_s) = \sum_{i=1}^{n} I_i^r I_i^s$$

$$\mathrm{sim}(I_r, I_s) = \frac{\sum_{i=1}^{n} I_i^r I_i^s}{|I_i^r||I_i^s|}$$

クラス間の類似度 (距離) は、次の手法が良く用いられる。

最近隣法 クラス Xに含まれるアイテム x とクラス Yに含まれるアイテム yのうち最も類似度の高い (距離の小さい) 値をクラス間の類似度 (距離) とする。

最近隣法 クラス Xに含まれるアイテム x とクラス Yに含まれるアイテム yのうち最も類似度の低い (距離の大きい) 値をクラス間の類似度 (距離) とする。

重心法 クラス Xの重心とクラス Yの重心の類似度 (距離) をクラス間の類似度 (距離) とする。

グループ間平均連結法 クラス Xに含まれるアイテムとクラス Yのクラス Yに含まれるアイテムのすべての組合せで類似度 (距離) を計算し、その平均をクラス間の類似度とする。

アイテムが共起情報からシソーラスを構築する時のように単語でベクトルの要素が共起頻度である場合や、テキスト分類をする時のようにアイテムがテキストでベクトルが単語の出現頻度である場合などでは、クラス間の類似度を直接計算する手法が提案されている [3, 5, 10]。

凝集型クラスタリングでは、1 つの併合のステップで最も類似度の高いクラスを併合し、併合されたクラスは他のクラスとの併合の可能性を捨象している。このため、生成される階層が最良の階層でない可能性がある。

3 最良パス探索を用いたクラスタリング

ここでアイテム数を Nとすると、凝集型クラスタリングは N個のクラス ($C = \{c_1, \ldots, c_N\}$) を併合して最終的に 1 つのクラスにすることになる。ある併合のステップで $c_i, c_j (i < j)$ を併合したとき、新しいクラスを番号の小さい c_iになるとする。このように定義すると、クラスタリングは、$N * (N-1)/2$ 個のクラスとクラスを併合する状態の最適なパスを探索する問題と見ることができる。t 回めの併合のステップを状態

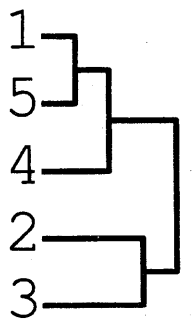

図 1: 5 アイテムのデンドログラム

$q(t)$ で、クラス c_iと c_jを併合する状態を $q(i,j)$ で表すとすると、t 回めの併合のステップで、クラス c_i, c_j を併合した状態は、$q(t) = q(i,j)$ となる。よって、N 個のアイテムをクラスタリングすることは状態の系列 $q(1), q(2), \ldots, q(N-1)$ を推定することになる。図 1 は、5 つのアイテムのデンドログラムである。最初に 1 と 5 を併合してクラス 1 にし、次に 1 と 4 を併合してクラス 1 に、2 と 3 を併合して 2 に、最後に 1 と 2 を併合して、1 にしている。これを状態の系列で表すと図 2のようになる。

1 回併合を行なうと次の併合で考慮する 1 つクラスが減る。例えば、最初にクラス 1 とクラス 2 を併合した ($q(1) = q(1,2)$) とすると、次の併合のステップでは、クラス 1, 3, 4, 5 の 4 つの中で併合するクラスを探すことになる。クラスタリングを状態系列で見た時には、状態 $q(1) = q(1,2)$ から状態 $q(2)$ へ接続できるかどうかの判定をする必要がある。この場合、$q(2) = (1,3), (1,4), (1,5)$ などには連接できるが、クラス 2 がすでにないため $q(2) = (1,2), (2,3)$ などには連接できない。状態間で連接できるものを実線の矢印で、連接できないものをアミカケの矢印で表すと、図 3のようになる。また、図 4の例では、$q(1) = q(1,5)$, $q(2) = q(1,4)$ から、$q(3) = q(2,5)$ には連接できないが、$q(3) = (2,3)$ には連接可能である。つまり、$q(t) = q(i,j)$ には、$q(1)$ から $q(t-1)$ までの

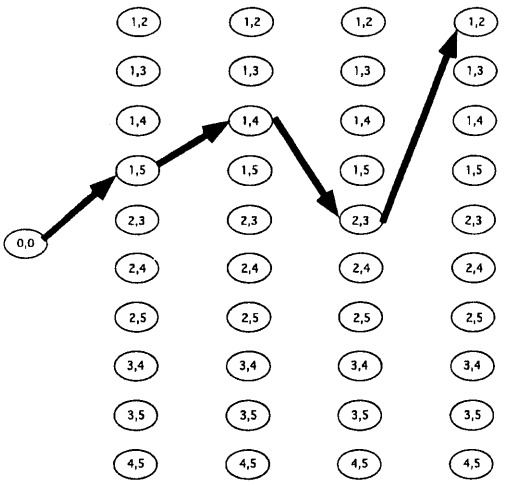

図 2: 併合の状態系列

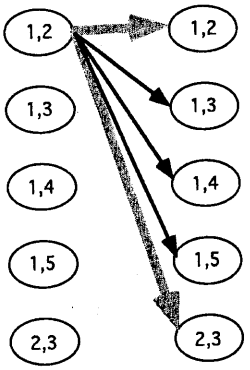

図 3: $q(1,2)$ からの連接可能性

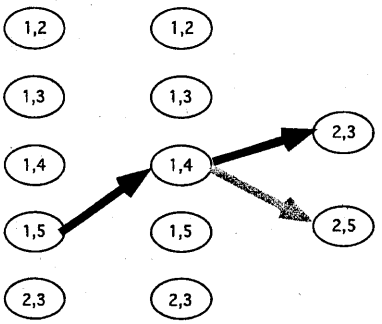

図 4: $q(1,5), q(1,4)$ からの連接可能性

間に、$q(*,j)$ の状態と $q(*,i)$ の状態がないときに連接できる。これは、$q(t-1)=q(k,l)$ に到達する最適なパスを記録して置けばそれまでの状態がわかり、連接可能性が判定できる。

また、同じ $q(t)=q(i,j)$ でもそれまでのパスが異なっていれば、i,jに含まれるアイテムが異なってくる。この例では、$q(1,2), q(1,4)$ から $q(1,3)$ に到達し

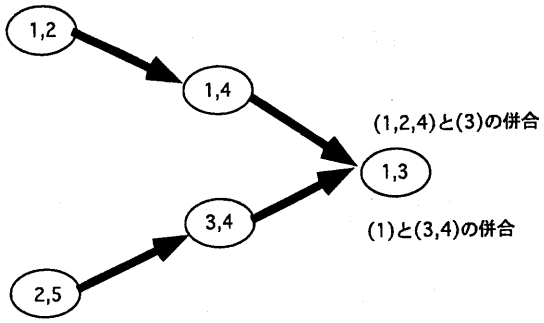

図 5: 1 と 3 の併合

た時には、クラス 1 にはアイテム 1,2,4 が、クラス 3 には、アイテム 3 が含まれており、アイテム 1,2,3,4 の併合をすることになり、$q(2,4), q(3,4)$ から $q(1,3)$ に到達した時は、クラス 1 にはアイテム 1 が、クラス 3 には、アイテム 3,4 が含まれており、アイテム 1,3,4 の併合をすることになる。

すべての可能な連接に対して、すべてのパスを計算し最もコストの小さいパスを選択すれば良いが、組合せ爆発を起こすので Viterbi アルゴリズム [7, 1] を応用して直前の状態までの最も有効なパスと次の状態の連接のみを考慮して、最終的にコストの小さいパスを見つけることにする。

本手法のアルゴリズムを次に示す。

SEQSCORE(t,i,j) は、$q(t)$ まで最もコストが低くなるパスを通って $q(i,j)$ まで到達した時のコストで、BACKPTR(t,i,j) はその SEQSCORE のときの $q(t-1)$ の状態を示すポインタであるとする。また、$q(t)$ が、$q(i,j)$ であるときのことを $q(t,i,j)$ で表している。クラス i,j 間のクラス間距離は、$dis(C_i,C_j)$ で表している[1]。

[1]それまでのパスによってクラスが含むアイテムが異なるため、それぞれのクラスが含むアイテムを BACKPTR をたどることに

initial step 1. それぞれのアイテムに対して、その1つのアイテムからなるクラスを作る。

2. $BACKPTR(1,i,j) = q(0,0,0)$

3. $\mathrm{SEQSCORE}(1,i,j) = \mathrm{dis}(C_i, C_j)$

iteration step $t = 2$ から、$t = N-1$ まで以下を繰り返す

1. $q(t) = q(i,j)$ に対して、到達可能な $t-1$ の状態 $(q(t-1) = q(k,l))$ の集合 $Q(t)$ を求める。

2. $\mathrm{SEQSCORE}(t,i,j) = min_{k,l}(\mathrm{SEQSCORE}(t-1,k,l) + dis(C_i, C_j))$

3. $\mathrm{BACKPTR}(t,i,j) =$ その時の $t-1$ の状態 $q(t-1,k,l)$

identification step $\mathrm{SEQSCORE}(N-1,1,j)$ が最小になる状態 $q(N-1,1,j)$ の $\mathrm{BACKPTR}(N-1,1,j)$ から状態を $q(1,k,l)$ までパスたどる

このアルゴリズムでは、$q(t-1), q(t)$ のパスを候補として最大状態数 $(N*(N-1)/2)$ 個分考慮しており、また、それまでのパスを含めたコスト SEQSOCRE を最小にするパスを残していくので、併合の順序としての文脈を考慮していることになる。

外から与えられる文脈をコストの中に導入するには、連接コストとして導入し、SEQSCORE に連接コストを加えて、計算するようにする。例えば、アイテム 1,3 がクラスになり易いということがあらかじめ文脈として、与えられたとすると、できるだけ早い段階でアイテム 1 と 3 がクラスになるような併合の系列が残って欲しいので、1,3 が 1 つのクラスになるような併合をする状態への連接コストを、併合の系列のはじめの方では、小さく、後の方では大きく与えるようにすれば良い。

4 名詞のクラスタリング

本節では、第 3節で述べたアルゴリズムを、名詞集合のクラスタリングに適用した例について述べる。クラスタリングの対象となる名詞は、コーパスから採集し、格関係で動詞と共起する頻度を「格+動詞」を 1 つの次元としたベクトルで表現する。本節での例では、次の 10 語[2]をクラスタリングする。括弧内の番号はアイテムの番号である。

日本 (1)、中 (2)、人 (3)、問題 (4), こと (5)、米国 (6)、場合 (7)、現在 (8)、方法 (9)、今 (10)

10 語では、各併合のステップで考えられる状態は、$q(1,2), q(1,3), \ldots, q(9,10)$ で、45 通り存在する。まず、initial step として、アイテムの任意の 2 つの組合せの距離[3]を計算し、最初の状態を作る。

次に 2 回目の併合ステップに相当する $q(2)$ のそれぞれの状態で最もそれまでの系列のスコアである SEQSCORE が小さくなる $q(1)$ の状態を決定する。$q(2) = q(1,6)$ に対する $q(1)$ の状態の決定する場合、これは 2 回目の併合で「日本」を含むクラスと「米国」を含むクラスを併合するときに、1 回目にはどのような併合が行なわれたかを決めることに相当する。このとき、$q(1) = q(1,6), q(2,6), q(3,6), q(4,6), q(5,6)$ は、すでにクラス 6 が併合されて考慮する必要がなくなっているので遷移しない。$q(1,7)$ のとき SEQSCORE が最小だったとすると、$q(2) = q(1,6)$ の前の状態が $q(1,7)$ として、BACKPTR に記録する。この時の併合は「日本」「場合」のクラスと「米国」のクラスを併合したことになる。

3 回目以降の併合も同様に行なわれ、最終的に 9 回めの併合では、クラス 1 と他の 9 個のクラスの併合となる最大 9 個の系列が得られるが、途中の状態で最も距離の小さくなる併合のみを残していくため、この場合では、4 つの系列が残った。

- q(4,8) q(4,10) q(4,7) q(3,4) q(5,6) q(1,5) q(2,9) q(2,3) q(1,2)
- q(1,10) q(1,8) q(6,7) q(1,2) q(4,5) q(6,9) q(4,6) q(1,4) q(1,3)

よって決定する必要がある。

[2] この例では、EDR 共起辞書 [11] から名詞-動詞の共起を取り出した。この 10 語は EDR 共起辞書の出現頻度の高い方から 10 語である。「格+動詞」は、出現頻度の高いものから 200 語を取り出し、200 次元のベクトルにしている。

[3] アイテム間の距離は単純ユークリッド距離を用い、クラス間の距離は、グループ間平均連結法を用いているが、任意の距離の尺度を用いることができる。

- q(2,7) q(2,6) q(1,2) q(4,5) q(3,10) q(1,3) q(8,9) q(4,8) q(1,4)
- q(4,8) q(4,10) q(4,7) q(4,9) q(4,6) q(3,4) q(2,3) q(1,2) q(1,5)

このなかで、$q(9)$ のときの系列のスコアSEQSCOREが小さいものが、最終的に得られるクラスの階層になる。今回の場合は、1つめの $q(9) = q(1,2)$ の系列であった。この時のクラスの階層を図6に示す。

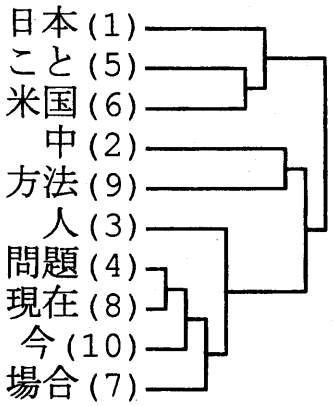

図 6: 10語のクラスタリング

ここで、例えば、「日本」と「米国」が類似しているという文脈を外から与えたとする。この場合は、アイテム1(日本) とアイテム6(米国) が系列の早い時期に併合して同じクラスになるように連接のコストを与えることによって実現する。ここでは、「$q(2)$ で 1, 6 が併合される時の連接コスト」 < 「$q(3)$ で 1, 6 が併合される時の連接コスト」 < 「$q(4)$ で 1, 6 が併合される時の連接コスト」 < ... となるようにコストを与える。最終的に次の併合の系列が得られた。

q(1,6) q(1,7) q(1,8) q(1,10) q(4,5) q(4,9) q(1,3) q(1,4) q(1,2)

この時のクラスの階層を図7に示す。

同様に「現在 (8)」と「今 (10)」が類似しているという文脈を与えると、次の系列が得られる。

q(4,8) q(4,10) q(4,7) q(5,6) q(1,5) q(2,9) q(2,3) q(1,2) q(3,4)

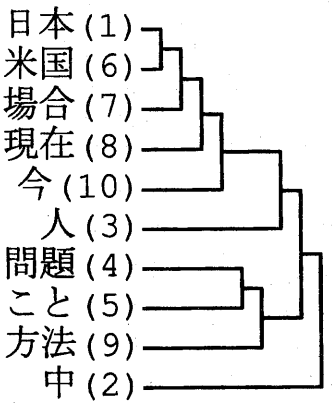

図 7: 「日本」と「米国」が似ているという文脈を与えた時のクラス階層

この時のクラスの階層を図8に示す。

5 おわりに

本稿では、階層クラスタリングをする時の2つの文脈を考慮した凝集型クラスタリングアルゴリズムを提案した。ここでいう2つの文脈とは、併合する順序に関する文脈と、外から与えられるどのアイテム同士がクラスを作りやすいかということに関する文脈である。本アルゴリズムの適用例として、10語の名詞のクラスタリングを示した。

今後の課題として、次のことがあげられる。

単語の分類やテキスト分類では、対象となるアイテムの数が莫大になる。本手法の問題として、併合する順序の文脈を考慮したことにより計算量が増えてしまっているということがある。大規模なアイテムに対応するには、ある一定数のクラスにまとめた後に1つのクラスにまとめる [10] 手法が必要になる。これは、一定数のクラスへの分類と、併合の系列の各ステップで考慮すべき併合の状態との現実的な対応をとることによって対処できる可能性がある。

外から与える文脈では、併合の状態の連接コストとして導入しているが、どのようなコストの割当がクラスタリングの対象にふさわしいかの検討も必要である。

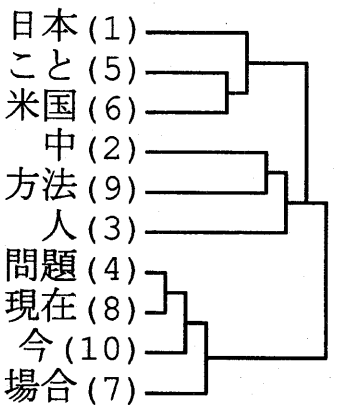

図 8: 「現在」と「今」が似ているという文脈を与えた時のクラス階層

参考文献

[1] *Natural language understanding*, chapter 7. Benjamin/Cummings Pub. Co., 2nd edition, 1995.

[2] Brian S. Everitt. *Cluster analysis 3rd ed.* Academic Press, 1994.

[3] Makoto Iwayama and Takenobu Tokunaga. Hierarchical Bayesian Clustering for Automatic Text Classification. In *IJCAI 95*, pp. 1322–1327, 1995.

[4] Hideki Kozima and Akira Ito. Context-sensitive measurement of word distance by adaptive scaling of a semantic space, 1995.

[5] Theeramunkong Thanaruk and Manabu Okumura. Learning a Grammar from a Bracketed Corpus. 情報処理学会自然言語処理研究会資料, No. 116-13, pp. 85–92, 1996.

[6] Takenobu Tokunaga, Makoto Iwayama, and Hozumi Tanaka. Automatic Thesauri Construction Based on Grammatical Relations. In *IJCAI 95*, pp. 1308–1313, 1995.

[7] A. J. Viterbi. Error bounds for convolution codes and an asymptotically optimal decoding algorithm. In *IEEEE Trans. on Information Theory*, No. 13, pp. 260–269, 1967.

[8] Takefumi Yamazaki, Michael J. Pazzani, and Christopher Merz. Learning hierarchies from ambiguous natural language data. In *ICML 95*, pp. 575–583, 1995.

[9] 平岡冠二, 松本裕治. コーパスからの動詞格フレーム獲得と名詞のクラスタリング. 情報処理学会自然言語処理研究会資料, No. 104-11, pp. 79–86, 1994.

[10] 柏岡秀紀, Ezra W.Black. 相互情報量を用いた単語の分類手法. 「自然言語処理における学習」シンポジウム, pp. 104–111, 1994.

[11] 日本電子化辞書研究所. EDR 日本語共起辞書, 1994.