1N-04

MDL基準を用いた文書集合の分割手法の高速化

村松 茂樹† 鈴木 雅実‡ 松本 一則† 橋本 和夫†
†KDD研究所 ‡通信・放送機構

1. はじめに

文書集合を単位として類似度を評価するための文書集合の特徴化手法[1]を提案している．同手法では特徴化の過程で文書集合の複数のセグメントへの分割を行っており，各分割数での記述長最小となる分割のやり方を求め，その結果にMDL基準[2]を適用して最適な分割数を決定している．このように文書集合をMDL基準を用いて複数のセグメントに分割する場合，全ての分割について記述長を計算するのは，対象文書数や分割数が増えた場合に困難である．本稿では，遺伝的アルゴリズム(以下GA)を用いて準最適な分割を決定する方法を検討したので報告する．

2. 文書集合の分割手法

2.1 MDL基準に基づいた文書集合の分割

提案している文書集合の特徴化手法では，文書集合の特徴をベクトル形式で表現するために，全文書の最小スパン木を構成し，それをMDL(Minimum Description Length)基準に基づき

$$L = L_1 + L_2 \tag{1}$$

であらわされる記述長Lを最小とするような条件でいくつかのセグメントに分割する．

ここで，

- L_1は分割モデルの記述長
- L_2は分割モデルの与えられたデータを記述するのに必要な記述長

である．

L_1は最小スパン木を分割したセグメントを記述するものであり，構成した最小スパン木からどの枝を取り除くかによって表現できる．したがって，文書の数をn，セグメント分割の数をr本としたときにセグメント分割の場合は${}_{n-1}C_{r-1}$通りあり，それを表現するための記述長は

$$L_1 = \log_2({}_{n-1}C_{r-1}) \tag{2}$$

で表現することができる．また，L_2は，各セグメントにおける文書集合中の文書の対数尤度の和であり，

$$L_2 = -\sum_i \sum_j n(S_j, seg_i) \times \log_2 \frac{n(S_j, seg_i)}{\sum_k n(S_k, seg_i)} \tag{3}$$

で表現できる．ここで，$n(S_j, seg_i)$は，セグメントseg_i中の文書集合S_jに属する文書数である．

一般に，分割数が増えるとL_1は増加し，L_2は減少する．

2.2 準最適な分割を決定するためのGA

MDL基準に基づき最適な分割を決定するため，分割のために取り除く枝を遺伝子とし，そのリストを1次元配列で表したものを染色体としてGAを用いた．

選択は，適応度上位から一定割合の親を選択するように行い，世代が進んだ際に最も適応度の高い個体の適応度が悪くならないようにするためにエリート保存戦略を用いた．また，エリート個体が急速に広まり，局所解に陥るのを防ぐため，適応度が上位一定割合の個体と同じ染色体を持つ個体の適応度にはペナルティを課した．

交叉には，1点交叉を用いた．また，各個体につき一定の割合で遺伝子の置換を行った．

1世代の個体数は1000とし，一定の世代の間に，より適応度の高い個体が生まれてこなければ終了するものとした．

また，分割数が増えた時にL_2が増加しないことを保証するために，初期状態を生成する際に，1つ少ない分割数で最も適応度の高かった個体と同じ遺伝子を含む個体を一定数生成することも行った．

3. 評価実験

「学校における複合アクセス網活用型インターネットに関する研究開発」[3]に参加している学校から8校$(S_1, S_2, \ldots, S_8)$を選択し，2.の手法に基づきそれぞれの学校のウェブページの特徴を表現するための分割を行った．対象となった文書は8校合計で259である．

"Fast Splitting Method of Document Set based on MDL Criterion",
Shigeki Muramatsu†, Masami Suzuki‡, Kazunori Matsumoto†and Kazuo Hashimoto†
†KDD R&D Laboratories Inc.
‡Telecommunications Advancement Organization of Japan

3.1 全探索との比較

表1に，全探索を行った場合とGAを用いた場合の対数尤度と計算の所要時間を示す．時間は，分割数2における全探索に要した時間を1とした場合の時間である．

表 1: 全探索と GA を用いた場合との比較

	全探索		GA	
	対数尤度	所要時間	対数尤度	所要時間
2 分割	482.434	1	—	—
3 分割	403.710	108	403.710	198
4 分割	339.811	9257	339.811	198
5 分割	—	—	314.865	199
6 分割	—	—	294.587	199

全探索を行った分割数の範囲においては，全探索の場合とGAを用いた場合で求められた解に差はなかった．また，全探索の場合は分割数が増えると計算に必要な時間が大きく増加しているのに対し，GAを用いた場合は，分割数を増やしても計算に必要な時間が一定のオーダーであった．

3.2 分割数の増加にともなう記述長の変化

図1に，分割数を変化させたときの記述長を示す．

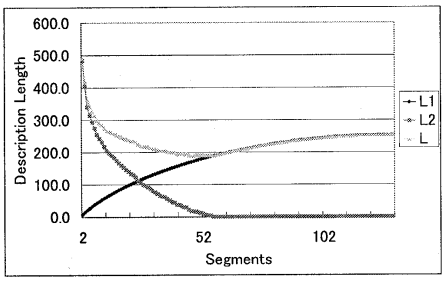
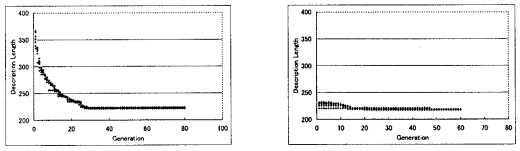


図 1: 分割数による記述長の変化

分割数が増えるに従い，L_1 が増加し，L_2 が減少していることが確認できる．また，L は分割数がある程度になるまでは減少し，その後増加することも確認できたが，MDL基準で最適となる L 最小の場合の分割数の値は大きく，分割されたセグメントあたりの文書数が少なくなっており，データに対する信頼性が低下していると考えられる．したがって，異なる文書集合に対しても評価実験を行い，提案した分割手法が有効であるような文書集合の特徴や，分割モデルとデータの記述の仕方等に関して検討を行う必要があると考えられる．

3.3 1つ少ない分割数での解の利用の検討

図2に，初期状態の生成の際に1つ少ない分割数での解の遺伝子情報を含む個体を意図的に一定数生成した場合と，初期状態をランダムに生成した場合の分割数11での世代ごとの適応度の高い個体10体の記述長の値を示す．



1つ少ない分割での解を利用しないもの　　1つ少ない分割での解を利用したもの

図 2: 世代による記述長の変化

1つ少ない分割での解を利用したものは，世代数が少ない状態でも最終的に求められた準最適解に近い値になっており，計算時間を短くしたい場合に有効であると考えられる．

4. おわりに

MDL基準を用いた文書集合の分割をGAを用いて準最適解を求めて行う手法について検討した．GAを用いることにより，全探索では確認することのできなかったが分割数の増加に伴う記述長の変化が確認できた．今後は，提案した分割手法の有効性に関して評価を行う予定である．

本研究は通信・放送機構 (TAO) の「学校における複合アクセス網活用型インターネットに関する研究開発」の一環として実施した．

参考文献

[1] 村松ら: “MDL基準を用いた文書集合の特徴化手法”, 電子情報通信学会総合大会, 2000.

[2] J. Rissanen: “Modeling by shortest data description.”, In *Automatica*, 14:465-471, 1978.

[3] 文部省/郵政省/通信・放送機構: “情報通信が創る21世紀の教育”, 先進的教育用ネットワークモデル地域事業 パンフレット, 1999.