Vol.2014-SLP-104 No.8
2014/12/15

劣モジュラ最適化を用いた文部分集合選択によるコーパス構築法

篠原 雄介[1,a)]

概要：所与の文集合から所望の単位分布（音素分布や単語分布）を持つ文部分集合を選択する方法を提案する．所望の単位分布を持つ文部分集合の選択は，音声認識・合成や自然言語処理の研究開発に用いるコーパスを構築する上で基礎的な手続きである．従来はヒューリスティックな方法が用いられてきたが，理論的な保証がなく悪い文部分集合を選択してしまう恐れがあった．本稿では所望の単位分布を持つ文部分集合を選択する問題が劣モジュラ最適化の問題として定式化出来ることを示す．この定式化を用いることで劣モジュラ最適化の強力なアルゴリズムの適用が可能となり準最適な文部分集合を高速に選択することが可能となる．提案法を用いて所望の音素分布を有する音声コーパスを構築する実験を行ったところ良好な結果を得たので報告する．

キーワード：劣モジュラ最適化，部分集合選択，音素分布，Kullback-Leibler ダイバージェンス，コーパス構築，音声認識，音声合成

1. はじめに

所与の文集合から所望の単位分布を持つ文部分集合を選択する問題は音声言語処理の多くの場面で現れる重要な問題である．ここでいう単位とは音素や単語，n-gram などであり，文はこのような単位の集まりとして表現される（例：bag of words）．音声認識や音声合成においてモデルの学習に用いる音声コーパスは種々の音素をバランス良く含むと良いと言われており，音声コーパス収録において話者に読み上げてもらう文集合を構築する方法が長く研究されてきた．言語処理においても言語モデルのドメイン適応のために所望の n-gram 分布を持つ文部分集合を選択する方法が研究されている．本稿では音声認識のために所望の音素分布を持つ文部分集合を選択する問題を題材として議論を進めていくが，提案する方法は音声認識に限らず幅広い分野で応用可能である．

種々の音素をバランスよく含む文集合を構築する数々の方法が提案されてきた．与えられた文集合から予算内（例：所定の文数内）で最良の部分集合を選択する問題として定式化するのが一般的である．最も古くから知られた手法のひとつはエントロピー最大化法である [1]．文部分集合の音素バランスの良さを音素分布のエントロピー（つまり一様性）で評価する．与えられた文集合から無作為に初期の部分集合を選択した後，無作為に選択した文の対（母集合と部分集合から一文ずつ）の交換を繰り返すことでエントロピーを最大化する．所望の音素分布として一様分布しか使えないという限定はあるものの，この方法で構築された音声コーパスは音声認識・合成の分野で現在でも広く使われている．より最近の手法として Kullback-Leibler (KL) ダイバージェンス最小化法がある [2]．所望の音素分布と文集合が与えられた時，音素分布が所与の音素分布に（KL ダイバージェンスの意味で）最も近くなるように部分集合を選択する．類似の手法を Siohan も用いている [3]．所望の i-vector 分布を有する文部分集合を選択するために KL ダイバージェンスを最小化する方法が研究された．Siohan により用いられたヒューリスティックなアルゴリズムは，元々は言語モデル適応のために n-gram 分布間の KL ダイバージェンスを最小化する方法として開発されたものである [4]．

これらのアルゴリズムはそれぞれの目的関数を最大化（あるいは最小化）する方法として必ずしも最適とは言えない．例えば，繰り返し交換法 [1] は十分な回数の交換が実行されれば最適であるが，大規模なコーパスを構築する際には収束するまで交換を繰り返すのは膨大な時間が掛かり非現実的である．Siohan [3] および Sethy [4] に用いら

[1] 東芝研究開発センター
Corporate Research and Development Center, Toshiba Corporation, Kawasaki, Kanagawa 212–8582, Japan
[a)] yusuke.shinohara@toshiba.co.jp

れたアルゴリズムは，無作為な順序で所与の文集合内の各文を選択し，KL ダイバージェンスが小さくなればその文を部分集合に加え，そうでなければ破棄するという単純なヒューリスティックに基づくものである．このアルゴリズムは簡潔で高速であるが明らかに最適ではない．

本稿では所望の音素分布を持つ文部分集合を選択するための準最適かつ高速なアルゴリズムを提案する．まず文部分集合の良さを評価する新しい評価関数を提案する．この評価関数は文部分集合が大きいほど（多くの音素を含むほど），また音素分布が所望のものに近いほど評価値が高くなる性質を持っており，”there is no data like more and balanced data” という意味合いを持つものである．次にこの評価関数が劣モジュラ性という特別な性質を有することを示す．劣モジュラ性を有する評価関数を用いることにより劣モジュラ最適化のアルゴリズムの適用が可能となり，理論的に準最適な解を高速に得ることが可能となる．具体的には，厳密解（これを求める問題は NP 困難である）の評価値に対して最悪でも約 63 %（サイズ制約の場合）または 約 32 %（ナップザック制約の場合）の評価値を持つ解が得られることが理論的に保証される．この約 63 % という下限は多項式時間アルゴリズムで達成可能な最良の下限であることが知られている．また評価関数の劣モジュラ性を活用することでアルゴリズムの大幅な高速化が可能で，提案法は大規模な問題に対しても適用することが出来る．このアルゴリズムは準最適かつ高速であるだけでなく簡潔で容易に実装できるという利点も有する．さらに，ある緩やかな条件を仮定したとき，提案する評価関数の最大化と KL ダイバージェンスの最小化が等価であることを示す．最後に提案法の妥当性を示す簡単な実験結果について紹介する．

2. 関連研究

音声合成のための音声コーパスの構築を目的とする場合には，所与の音素セット（例えばダイフォン）をカバーするような最小サイズの文部分集合を求めるアルゴリズムが用いられてきた．この問題は集合被覆問題と呼ばれ，簡単な貪欲アルゴリズムで効率良く解けることが知られている．集合被覆問題において貪欲アルゴリズムは準最適な解法であることが知られている．特に有名な方法として例えば [5], [6] などがある．音声認識の研究開発などより大きなコーパスが求められる場合には，集合被覆問題として扱うのは一般的ではなく，所望の音素分布を持つ文部分集合を選択する問題として扱われることが多い．例えば [1], [2], [7], [8], [9] などの方法が知られている．本稿では後者の場合について検討する．

劣モジュラ最適化 [10], [11] は近年大きな注目を集めている．ICML2013 や NIPS2013 など主要な国際会議で多くのチュートリアルが開催されており，幅広い問題へ応用可能であることが知られている [12]．離散最適化における劣モジュラ性は連続最適化における凸性と類似しており，目的関数がこの性質を持つ場合には強力な最適化手法を適用することが可能となる．機械学習の分野では盛んに研究されている技術であるが，音声言語情報処理の分野においてはまだ広く認知されるには至っていない．我々の知る範囲においては，所望の単位分布を有する文部分集合を選択する問題へ劣モジュラ最適化を適用するのは本研究が初めてである．

3. 方法

3.1 問題の定式化

文集合 U と予算 B が与えられた時，S に関する予算の制約のもとで，S の効用が最大になるように U の部分集合 S を選択する問題を考える．$J(S)$ を文部分集合 S の効用を評価する集合関数とし，文部分集合選択問題を次のように定式化する．

$$S^* = \arg\max_{S \subseteq U} J(S) \ \text{ subject to } \ \sum_{s \in S} c(s) \leq B, \qquad (1)$$

ただし $c(s)$ は文 s のコストである．例えば全ての文に対して単位コストを用いる，すなわち $c(s) = 1$ の場合，最大で B 個の文を選択出来るという制約条件になる．具体的には，問題は次の形に単純化される．

$$S^* = \arg\max_{S \subseteq U} J(S) \ \text{ subject to } \ |S| \leq B. \qquad (2)$$

このような制約のことを「サイズ制約（cardinality constraint）」と呼ぶ．より一般的には各文に異なるコストを設定することも出来る．例えば長い文ほど高いコストを設定するのは合理的であろう．このような制約のことを「ナップザック制約（knapsack constraint）」と呼ぶ．

3.2 目的関数

P を全音素の集合とする．例えば，文脈非依存音素を用いる場合，P は約 50 個の音素からなる．あるいは，トライフォンを用いる場合，P は約 5,000 個の（文脈依存）音素からなる．π_i を i 番目の音素の所望の確率，$\pi = (\pi_1, \ldots, \pi_{|P|})$ を所望の音素分布，$f_i(S)$ を S における i 番目の音素の頻度とする．このとき，文部分集合 S の効用を次式により定義する．

$$J(S) = \sum_{i=1}^{|\mathcal{P}|} \pi_i \log f_i(S). \qquad (3)$$

この効用の設計思想は次のとおりである．まず，効用が音素の対数頻度の線形結合となるようにした．これは，データの効用はその量の対数値に比例するという経験則に基づくものである．線形結合の重みには各音素の所望の確率を用いた．これは所望の音素分布のもとでの効用の期待値を表現することを意図したものである．

3.3 目的関数の解釈

この目的関数 $J(S)$ は "there is no data like more and balanced data" と唱えているものと解釈することが出来る．以下ではこの解釈の詳細について述べる．まず，文部分集合のバランスの良さを所望の音素分布からの KL ダイバージェンスで測るものと定義する．すなわち，

$$D_{\mathrm{KL}}(\pi \parallel p(S)) = \sum_{i=1}^{|\mathcal{P}|} \pi_i \log \frac{\pi_i}{p_i(S)} \quad (4)$$

$$= -\sum_{i=1}^{|\mathcal{P}|} \pi_i \log p_i(S) + \mathrm{Const.} \quad (5)$$

ただし，$p_i(S)$ は S における i 番目の音素の確率である．

$$p_i(S) = \frac{f_i(S)}{f_T(S)}. \quad (6)$$

ここで，$f_T(S)$ は $\sum_i f_i(S)$ により定義される総頻度である．すると次式が得られる．

$$D_{\mathrm{KL}}(\pi \parallel p(S)) = -\sum_i \pi_i \log \frac{f_i(S)}{f_T(S)} + \mathrm{Const} \quad (7)$$

$$= -\sum_i \pi_i \log f_i(S) + \log f_{\mathrm{T}}(S) + \mathrm{Const.} \quad (8)$$

両辺整理して，

$$J(S) = \log f_{\mathrm{T}}(S) - D_{\mathrm{KL}}(\pi \parallel p(S)) + \mathrm{Const.} \quad (9)$$

各項の意味するところは次の通りである．

- $J(S)$: データの効用
- $\log f_{\mathrm{T}}(S)$: データの量（対数尺度）
- $D_{\mathrm{KL}}(\pi \parallel p(S))$: データのバランスの悪さ

したがって，データの効用はその量とバランスの良さの和である，というのが前式の意味するところである．換言すれば "There is no data like more and balanced data" となる．

3.4 劣モジュラ性

任意の $S \subseteq T \subseteq U$ と $s \in U \backslash T$ に対して，集合関数 $J(\cdot)$ が次の性質を持つとき，その集合関数は劣モジュラであるという．

$$J(S \cup \{s\}) - J(S) \geq J(T \cup \{s\}) - J(T). \quad (10)$$

文部分集合選択の文脈でいうと，ある文 s を小さな文部分集合 S に加える場合と，同じ文を大きな文部分集合 T に加える場合とでは，前者のほうが効用の増分は大きい，というのがこの不等式の意味するところである．別の言葉でいえば，この不等式は効用逓減の法則を表現している．

次に，式 (3) で定義した目的関数 $J(S)$ が劣モジュラであることを示す．ここで $0 < x \leq y$ と $0 \leq d$ に対して成り立つ次の不等式を用いる．

$$\log(x+d) - \log(x) \geq \log(y+d) - \log(y). \quad (11)$$

よって，全ての i に対して次の不等式が成り立つ．

$$\log f_i(S \cup \{s\}) - \log f_i(S) \geq \log f_i(T \cup \{s\}) - \log f_i(T). \quad (12)$$

π_i はすべての i に対して非負値なので，

$$J(S \cup \{s\}) - J(S) \geq J(T \cup \{s\}) - J(T), \quad (13)$$

となる．すなわち，我々の目的関数 $J(S)$ は劣モジュラである．

3.5 貪欲アルゴリズム

3.5.1 サイズ制約の場合

効用を最大化する文部分集合を厳密に探索する組み合わせ最適化問題は NP 困難であることが知られている．よって厳密な解を求めることはあきらめ，何らかの近似アルゴリズムを用いる必要がある．サイズ制約 $c(s) = 1$ のもとで劣モジュラ関数を最大化する問題に対して，貪欲アルゴリズムが準最適であることが知られている．すなわち，貪欲アルゴリズムよりも良い下限を持つ多項式時間アルゴリズムは存在しない．より厳密にいうと，次の定理が知られている．

定理 [15] $J(\cdot)$ が非負かつ単調な劣モジュラ関数のとき，貪欲アルゴリズムで求めた文部分集合 S^* の効用は次の下限を持つ．$f(S^*) \geq (1 - 1/e) \max_{|S| \leq B} f(S)$.

Algorithm 1 に文部分集合選択の貪欲アルゴリズムを示す．空集合として初期化した後，各繰り返しにおいて効用を最大化する一文を選択し，所定の文数に達した時点で停止する．

Algorithm 1 Sentence subset selection with a greedy algorithm (cardinality constraint)

Input: U, B, π
$S \leftarrow \phi$
repeat
  $s^* \leftarrow \arg\max_{s \in U \backslash S} J(S \cup \{s\}) - J(S)$
  $S \leftarrow S \cup \{s^*\}$
until $|S| = B$
Output: S

3.5.2 ナップザック制約の場合

サイズ制約を使った場合には貪欲アルゴリズムは長い文を選びやすくなる．しかし長い文を読み上げるのは短い文と比べてより高いコストが掛かるのが通常である．いろいろな長さの文からなる文集合 U が与えられたとき，貪欲アルゴリズムによって得られる文部分集合は高い音声収録コストにつながる恐れがある．このような状況においては各文に異なるコストを設定するのが合理的であろう．ナップザック制約の場合には，前記の貪欲アルゴリズムを改良

したもの [16], [17] を用いることが出来る．このアルゴリズムは次の性能下限を持つことが知られている．

$$f(S^*) \geq \frac{1}{2}\left(1-\frac{1}{e}\right) \max_{S\subseteq U:\sum_{s\in S} c(s)\leq B} f(S). \qquad (14)$$

なおこのアルゴリズムよりも良い下限，具体的には $1-1/e$ の下限を持つアルゴリズム [18] もあるが，より複雑度が高く大規模な問題へのスケールが困難であるため実用化は難しいと思われる．

Algorithm 2 Sentence subset selection with a greedy algorithm (knapsack constraint)

Input: U, B, π

```
//Unit-cost
```
$S_{uc} \leftarrow \phi$
while $\sum_{s\in S_{uc}} c(s) \leq B$ **do**
  $s^* \leftarrow \arg\max_{s\in U\setminus S_{uc}} J(S_{uc}\cup\{s\}) - J(S_{uc})$
  $S_{uc} \leftarrow S_{uc}\cup\{s^*\}$
end while

```
//Cost-benefit
```
$S_{cb} \leftarrow \phi$
while $\sum_{s\in S_{cb}} c(s) \leq B$ **do**
  $s^* \leftarrow \arg\max_{s\in U\setminus S_{cb}} \frac{J(S_{cb}\cup\{s\})-J(S_{cb})}{c(s)}$
  $S_{cb} \leftarrow S_{cb}\cup\{s^*\}$
end while

Output: $\arg\max\{J(S_{uc}), J(S_{cb})\}$

3.6 高速化

ここまで説明してきた文部分集合選択のための貪欲アルゴリズムはさほど高速ではない．代わりに大幅に高速な改良版 [16], [17] を用いることが出来る．目的関数の劣モジュラ性を活かして関数評価の回数を大幅に削減出来ることを利用した高速化法である．この高速化法はサイズ制約とナップザック制約のどちらの場合にも適用することが出来る．この高速化法をセンサー配置問題へ適用した例 [17] を紹介する．センサーを設置出来る場所の集合が与えられたとき，最適な部分集合を選択する戦略が検討された．実験ではこの高速化法を適用することで標準的な貪欲アルゴリズムと比べて約 700 倍の高速化が実現された．この例と同様にして，本稿で検討している文部分集合選択問題も非常に高速に解くことが可能である．後述する実験ではナップザック制約のもとでこの高速化法を用いた．

3.7 KL 最小化との等価性

ある緩やかな条件のもとで，上記の貪欲アルゴリズムで求めた文部分集合は，所望の音素分布からの KL ダイバージェンスが最小となる文部分集合と等価であることを示す．

まずサイズ制約の場合について考える．提案法によって求められた文部分集合は，サイズ制約 $|S| = B$ を満たすあらゆる文部分集合 $S \subseteq U$ の中で（近似的に）$J(S)$ を最大化するものである．ここで U の各文の長さ（i.e. その文を表現する音素系列の長さ）が概ね等しいと仮定する．この仮定は多くの状況において現実的である．例えば音声コーパス収録に用いる文集合を選択する際には，母集団となる文集合に含まれる文をおおよそ等しい長さ（読み上げやすい長さ）にしておくのが一般的である．この一定の長さを L とおく．すると $|S| = B$ なる制約から $f_{\mathrm{T}}(S) = BL$ が導かれる．すなわちこの仮定のもとではすべての S について $\log f_{\mathrm{T}}(S)$ が一定となる．よって $J(S)$ を最大化する文部分集合 S^* は，与えられたサイズ B を持つすべての文部分集合 $S \subseteq U$ の中で，最小の KL ダイバージェンス $D_{\mathrm{KL}}(\pi \,\|\, p(S))$ を持つ文部分集合と等価であることが分かる．ナップザック制約の場合についても，$\log f_T(S) = B$ なる仮定を置いた場合，これと同様の議論をすることが出来る．

4. 実験結果

提案する文部分集合選択アルゴリズムと無作為選択アルゴリズムの比較を実施した．文を抽出する元の文集合 U として，新聞や小説など多様な情報源から収集した日本語の 248,530 文を用いた．一文あたりの平均の音素数は約 21 であった．U に登場した 4,911 個のトライフォンを音素セット P として用いた．総音素数 400,000 個以下という予算（制約条件）のもとで提案した貪欲アルゴリズムにより文部分集合 S を選択した．すなわち，文の長さ（文を構成する音素の数）をコスト $c(s)$ として用いた．本実験では，所望の分布 π として一様分布を用いた．これは，実験結果の解釈を容易にするためにとった措置（音素分布が平坦であるほど所望の分布に近い）であるが，実用の際には任意の分布を用いることが出来る．比較のため，同じ予算のもとで無作為に文を選択した結果，18,939 文からなる文部分集合を得た．図 1 に 2 つのアルゴリズムで得られた文部分集合について音素（トライフォン）の頻度（対数尺度）を示す．提案するアルゴリズムで作成した文部分集合のほうがより平坦な音素分布を持っていること，すなわち所望の分布に近い分布を持っていることが分かる．提案するアルゴリズムで選択された文部分集合のほうが，（特に低頻度のトライフォンにおいて）より高い頻度を持っている．音声の統計モデルを構築する際に用いる音声コーパスとして，このような性質は好適といえる．なぜならば，そのような音声コーパスを用いて作成した音声認識（合成）システムは，低頻度な音素の認識（合成）において良好な性能を得ると期待されるからである．我々の C++による実装では貪欲アルゴリズムは僅か 237 秒で 18,989 文の選択を完了した．

5. おわりに

所望の単位分布を持つ文部分集合を選択する問題が劣モ

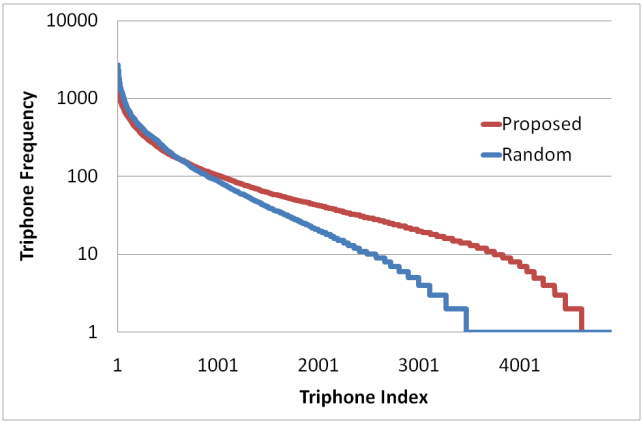

図 1 提案法と無作為選択法の比較

ジュラ最適化の問題として定式化出来ることを初めて示した．この定式化に従って所望の単位分布を持つ準最適な文部分集合を高速に選択するアルゴリズムを提案した．またある緩やかな条件下で提案する評価関数の最大化と KL ダイバージェンスの最小化が等価になることを示した．所望の単位分布を持つ文部分集合を選択する問題は音声認識に限らず幅広い問題に応用可能であると考える．今後さらに多くの問題への提案法の適用を検討していきたい．劣モジュラ最適化は音声言語処理の分野ではまだ十分に認知されているとは言い難いが，本稿でも示したとおり幅広い問題に適用可能な強力な手法である．本稿が劣モジュラ最適化の魅力を伝える一助になったならば幸いである．

参考文献

[1] Iso, K. and Watanabe, T.: Design of a Japanese Sentence List for a Speech Database, *Proceedings of the Acoustical Society of Japan Spring Meeting*, 2–2–19 (1988).
[2] Cui, Xiaodong and Alwan, Abeer: Efficient adaptation text design based on the Kullback-Leibler measure, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2002).
[3] Olivier Siohan and Michiel Bacchiani: iVector-based acoustic data selection, *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (2013).
[4] A. Sethy, P. G. Georgiou, B. Ramabhadran, and S. S. Narayanan: An iterative relative entropy minimization-based data selection approach for n-gram model adaptation: *IEEE Transactions on Audio, Speech and Language Processing* (2009).
[5] Jan P.H. van Santen and Adam L. Buchsbaum: Methods for optimal text selection: *Proceedings of the Eurospeech* (1997).
[6] Helene Francois and Olivier Boeffard: Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem, *Proceedings of the Eurospeech* (2001).
[7] Jia-lin Shen, Hsin-min Wang, Ren-yuan Lyu, and Lin-shan Lee: Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary mandarin speech recognition, *Computer Speech and Language* (1999).
[8] Yi Wu, Rong Zhang, and Alexander Rudnicky: Data selection for speech recognition, *Proceedings of ASRU* (2007).
[9] Evandro Gouvea and Marelie H. Davel: Kullback-Leibler divergence-based ASR training data selection, *Proceedings of INTERSPEECH* (2011).
[10] Satoru Fujishige: *Submodular Functions and Optimization* (2005).
[11] Andreas Krause and Daniel Golovin: *Submodular Function Maximization*, chapter in Tractability: Practical Approaches to Hard Problems, Cambridge University Press (2012).
[12] Andreas Krause and Carlos Guestrin: Optimizing sensing: From water to the web, *IEEE Computer* (2009).
[13] H. Lin and J. Bilmes: How to select a good training-data subset for transcription: Submodular active selection for sequences, *Proceedings of the Interspeech* (2009).
[14] H. Lin and J. Bilmes: Optimal selection of limited vocabulary speech corpora, *Proceedings of the Interspeech* (2011).
[15] G. Nemhauser, L. Wolsey, and M. Fisher: An analysis of the approximations for maximizing submodular set functions, *Mathematical Programming*, vol. 14 (1978).
[16] Minoux, M.: Accelerated greedy algorithms for maximizing submodular set functions, *Optimization Techniques, LNCS*, 234–243 (1978).
[17] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance: Cost-effective outbreak detection in networks, *Proceedings of the ACM International Conference on KDD* (2007).
[18] M. Sviridenko: A note on maximizing a submodular set function subject to knapsack constraint, *Operations Research Letters*, vol. 32, pp. 41–43 (2004).