2値を出力するマルコフモデル推定のための重なりのない頻度

高本　綺架†　　吉田　光男††　　梅村　恭司††

† 豊橋技術科学大学　情報・知能工学専攻　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1–1
†† 豊橋技術科学大学　情報・知能工学系　〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1–1
E-mail: †a153350@edu.tut.ac.jp, ††yoshida@cs.tut.ac.jp, †††umemura@tut.jp

あらまし　センサーデータのような 2 値の文字からなる長い文字列において，部分文字列の頻度を計算する場合，重複して計算されることが普通である．このような方法は，プログラムがあるパターンの頻度を過大評価してしまうことがあると考えた．本研究では，文字列における頻度計算において特定のパターンが重複して計算されることでラベル推定に与える影響を調べ，その重複を避けることによる変化を検討する．本研究では，分類手法として文字列の頻度を用いた情報量計算を使用する．情報量計算は，対象となる文字列を可能な分割方法全てに分割し出現した各文字列の頻度を計算することで行う．この情報量計算時の頻度計算において，重複して計算される可能性のあるものは重複を回避して計算を行うことを提案する．頻度計算の実験はマルコフ過程によって生成された文字列を使用し，情報量計算を行う．実験の結果，文字列の頻度を計算する際に重複してカウントすることを回避することで，モデル推定のバイアスが軽減されることがわかった．

キーワード　情報量計算，頻度．マルコフモデル

1. はじめに

センサーデータのような数値からなる文字列を用いて対象データのラベルを推定する研究がある [1]．このような研究の中で，対象となるデータに含まれる文字列の頻度を計算する場合がある．文字列の頻度を計算する際，同一文字が連続している場合，重複してカウントされる場合がある．文字列 “0” は文字列 “001” の中に 2 回出現する．それぞれ文字が 3 回繰り返すとすると重複を許して数えた場合文字列 “000” は文字列 “000000111” に 4 回出現し，文字列 “111” は 1 回と考えられる．重複のない頻度で計算すると文字列 “000”’ は文字列 “000000111” の中に 2 回，文字列 “111” は 1 回となる．

特定のパターンにおいて重複して出現する場合，そのパターンの頻度のが過大評価される可能性がある．そこで本研究では，パターンが重複してカウントされることによって生じる影響とその解決方を検討する．

文字列の頻度計算において，重複して計算されることの影響を調べるために 2 値の文字列のラベル推定実験をおこなった．テストデータの文字列は遷移確率が 20%であるものと 1%であるマルコフ過程に基づき生成する．同様に頻度計算に用いるデータベースとなる文字列も同じマルコフ過程に基づいて生成する．データベースとなるデータ群は各確率ごとに 3 セトづつ，計 6 セット作成する．このデータ群を用いてテストデータの情報量を計算した．ここで情報量はデータ群中におけるテストデータの出現確率を反映したものとなる．実験の結果，遷移確率が 1%であるデータ群を用いた場合の情報量が著しく小さくなり，推定が困難となることがわかった．そこで我々は，文字列の頻度計算において重複が発生する場合，それを回避するという手法を提案する．提案手法を用いて実験を行った結果，推定の精度には変化がないものの，情報量が著しく小さくなるという問題を解決することができた．以上の結果から，文字列の頻度を計算し推定を行う際，重なりのない頻度を用いるほうが，より有用であることを報告する．

2. 情報量計算

2.1 情報量計算

この節では，情報理論に基づきある文字列が持つ情報量を計算する手法について述べる．一般に，文字列の情報量を計算する際は一文字あたりの情報量を加算することで計算する．ある文字 c が持つ情報量 $I(c)$ はその文字が生起する確率 $P(c)$ を用いることで以下のように計算できる [2], [3].

$$I(c) = -\log_2 P(c) \tag{1}$$

長さが N の文字列 S において i 番目の文字を c_i とする．文字 c_i は互いに独立であり，その生起確率は $P(c_i)$ である．文字列全体の情報量は全ての文字が持つ情報量の合計になるため，以下のようになる．

$$\begin{aligned} I_c(S) &= I_c(\prod_{i=1}^{N} P(c_i)) \\ &= \sum_{i=1}^{N} I(c_i) \\ &= -\sum_{i=1}^{N} \log P(c_i) \end{aligned} \tag{2}$$

例えば，文字列 “ababc” であれば，この文字列は “a”，“b”，“a”．“b”，“c” で構成される系列であると考えることができる．出現する文字，“a”，“b”，“c” それぞれの文字の生起確率は 1/4，1/2，1/4 であるとする．各文字が互いに独立であれば，文字列 “ababc” の情報量は各文字が持つ情報量の合計と等しい．従って，その情報量は $-\log 1/4$，$-\log 1/2$，$-\log 1/4$，

$-\log 1/2$，$-\log 1/4$ の合計となる．

しかし，実際の文字列においては単語のように特定の文字列が繰り返し出現することが考えられる．そのような文字列に対しては 1 文字づつ情報量を加算する手法は適切ではない．先ほどの例であれば，“ab” という文字列が繰り返し出現するため，この “ab” を一つの塊と考える方がより自然である．そこで，本研究では文字列を可能な全ての部分文字列に分割し，各文字列の出現頻度に基づいて計算する手法を用いる [4]．文字列を部分文字列から構成される系列であると考えると，文字列が持つ情報量は次のように表せる．

$$I_s(S) = \min_{\pi_k \in \pi(S)} \left(- \sum_{t \in \pi_k} \log_2 P(t) \right) \tag{3}$$

式 (3) において，$\pi(S)$ は文字列 S の分割の集合であり，その数は 2^{N-1} である．文字列 {“abc”} であればその分割方法は {“abc”}，{“ab”，“c”}，{“a”，“bc”}，{“a”，“b”，“c”} の 4 通りである．またこれらの分割方法で出現する部分文字列は “abc”，“ab”，“bc”，“a”，“b”，“c” の 6 種類であり，これらは式 (3) における pi_k である．これら全ての部分文字列について頻度を計算し，その頻度 $P(t)$ を用いて情報量を計算する．式 (3) を用いて各部分文字列の持つ情報量を計算し，最小のものを加算することで文字列の情報量とする．

この情報量計算手法は計算量が膨大になるという問題点があるため，計算量の削減を行う．文字列 “abd” の情報量を計算する場合を考える．本計算手法では一文字づつ情報量を計算し，加算していく．0 文字目の情報量は，文字が含まれていないため，0 である．1 文字目の文字 ‘a’ が持つ情報量は $-logP(“a''$) であり，これは 0 文字目の情報量に 1 文字目の情報量を加算することで得られる．2 文字目までの文字列 “ab” で得られる部分文字列は “ab”，“a”，“b” であるが，文字 “a” の情報量はすでに計算しているためその値で置き換えることができる．従って，文字列 “ab” で計算するのは文字 “b” の情報量と文字列 “ab” のみである．以降の文字列が持つ情報量も同様に計算できる．3 文字目までの文字列 “abd” の部分文字列は “abd”，“a”，“b”，“d”，“ab”，“bd” である．これらの部分文字列のうち，“a” は 1 文字目の情報量計算で，“ab” と “b” は 2 文字目までの情報量計算ですでに計算されているため，それらの情報量で置き換えることができる．従って 3 文字目までの情報量を求めるために必要な情報量計算は “abd” と “bd”，“d” となる．このように計算する文字列の情報量を既に計算した情報量に置き換えることで計算量を削減できる．

2.2 情報量計算に基づくラベル推定

この節では情報量計算を用いてラベルを推定する際の推定方法について述べる．式 (3) を用いて文字列の情報量計算を行うには，各部分文字列の頻度を計算する必要がある．情報量計算においてはこの頻度をどのように計算するかが重要となる．本研究では，先行研究と同様に同ラベルのデータ群をまとめデータベースを作成しそのデータ群から頻度を計算する [5]．計算対象の組み合わせを図 1 に示す．図 1 において，矢印は起点のデータ群における頻度から確率を推定し，終点のデータがそこから取り出されるとしたときの情報量を計算することを意味する．また，x は対象となる文字列データである．また，データ群 A からデータ群 C はデータベースとなる文字列群のラベルである．対象となるデータに含まれる文字列を部分文字列に分割し，任意の部分文字列の出現頻度をデータ群から計算する．もし，データ x に含まれる部分文字列がデータ群 A に多く出現した場合，データ x のデータ群 A に対する情報量は他のデータ群に対する情報量より小さくなる．従ってデータ x のラベルはデータ群 A と同様に A である可能性が高いと考えられる．このようにデータのラベル推定を行う．

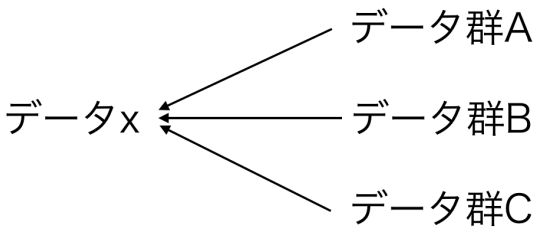

図 1　情報量計算手法

表 1　データセットの詳細

ラベル	遷移確率	長さ
テストデータ 1	20%	1760
テストデータ 2	1%	1760
データ群 1	20%	8000
データ群 2	20%	8000
データ群 3	20%	8000
データ群 4	1%	8000
データ群 5	1%	8000
データ群 6	1%	8000

3. 予 備 実 験

3.1 テストデータの作成

情報量計算において特定のパターンが重複して計算されることに対する影響を調査するために予備実験を行う．実験に用いるデータはマルコフ過程に基づいて作成する．使用したマルコフ過程を図 2，3 に示す．テストデータ 1 では図 2 に示すように 0 から 1 に遷移する確率が 20%であるものを用いて作成する．図 4 は作成されたテストデータ 1 の一部である．また，テストデータ 2 では図 3 に示すように 0 から 1 へ遷移する確率が 1%であるものを用いて作成する．図 5 は作成されたテストデータ 2 の一部である．図 4，5 に示すようにテストデータ 2 はテストデータ 1 より 0 または 1 が連続しやすい文字列となる．これらのマルコフ過程に基づいて 1760 文字からなる文字列を作成し，それをテストデータとする．頻度計算に用いるデータ群も同様に作成する．データ群は各確率でつ作成する．このデータ群はテストデータと同様にマルコフ過程を用いて 8000 文字からなる文字列を作成する．作成したデータについてまとめたものを表 1 に示す．表 1 では，データの名前とそのデータに用いたマルコフ過程の遷移確率，長さが示されている．

3.2 実 験 方 法

予備実験は作成した 2 つのテストデータについて，6 つの

データ群からその情報量を計算しラベルを推定する．テストデータ 1 の判定方法を図 6 に示す．矢印の意味は図 1 と同様である．テストデータ 1 の情報量は 6 つのデータ群を用いて計算される．図 6 においてデータ群 1 から 3 は遷移する確率が 20%のマルコフ過程によって生成されたものである．また，データ群 4 から 6 は遷移する確率が 1%のマルコフ過程によって生成されたものである．テストデータに出現する部分文字列の頻度を 6 つのデータ群から求め，テストデータが持つ情報量を計算する．計算された情報量の中でもっとも小さい情報量となったデータ郡のラベルがテストデータのラベルとなる．この処理をテストデータ 2 にも行い，それぞれ最も小さい情報量となったものが，テストデータのラベルとなる．

3.3 評価方法

予備実験は 20%と 1%の 2 つの遷移確率を持つマルコフ過程を用いて生成された文字列の情報量を計算し，そのラベルを推定する．計算される情報量は類似のパターンが多いほど小さくなるため，特定のパターンの頻度が重複してカウントされていても問題がなければ，各テストデータの情報量は同一遷移確率のデータ群を用いたとき小さくなると考えられる．反対に，頻度を重複して計算することに問題があれば，特定のパターンが重複しやすい文字列の情報量が小さくなると考えられる．

3.4 実験結果

予備実験の結果を表 2 に示す．表 2 の数字は情報量の小数点以下を切り捨てたものである．また各ラベルに置けるパーセンテージはその文字列を生成するために使用したマルコフ過程の遷移確率である．予備実験の結果を見ると，文字が遷移する確率が 20%であるテストデータ 1 の情報量は遷移確率が 1 %であるデータ群の方が情報量が小さい．従ってテストデータは遷移確率が 1%であると推定される．一方で，遷移確率 1%のテストデータ 2 では情報量は遷移確率 1%のデータであるデータ群 4 の情報量が最も低いため，テストデータ 2 は遷移確率 1%のデータであると推定される．遷移確率が 20%のマルコフ過程により生成された文字列であるテストデータ 1 の各データ郡に対する情報量を見ると，遷移確率が 1%であるデータ郡 4 から 6 を用いた際の情報量が小さい．直感的には，同一確率で生成された文字列を用いた方が，部分文字列の出現頻度が多くなり情報量が小さくなることが考えられる．しかし，実験の結果ではどちらの情報量も遷移確率が 1%であるデータ群を用いた情報量が小さくなっており，さらにテストデータ 1 の情報量は遷移確率が 1%のデータ群の頻度を用いた時著しく小さくなっている．これは同じ文字が連続して出現するため特定の文字列の頻度が多く計算されているためであると考えられる．

4. 提案手法

情報量計算を用いた情報量の推定において，同一文字が連続するような文字列の場合に情報量が著しく小さくなるという問題がある．この問題は部分文字列の頻度を数える際，重複して数えていることが原因であると考えられる．重複する例を以下に示す．文字列 “00001100” について部分文字列 “00” の頻度を計算する場合を考える．文字列 “00001100” において，計算される部分文字列 “00” の頻度は 4 である．これは最初の “0000” 部分で “00” が 3 回数えられているためである．しかし，直感的には “00” が重ならないように数え，頻度が 3 とする方がより自然と考えた．そこで本研究では，重なりが生じる可能性があるパターンが出現した場合，重複して数えることを回避するようにすることで情報量計算に与える影響を検討する．

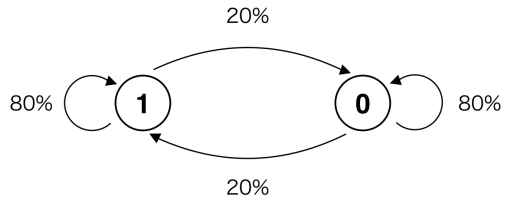

図 2　遷移確率 20%であるマルコフモデル

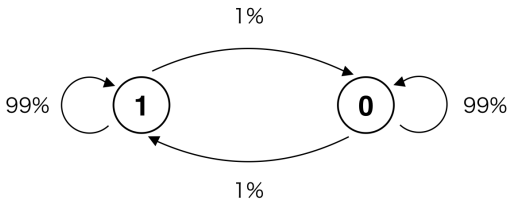

図 3　遷移確率 1%であるマルコフモデル

```
00000000001111111000001110011111111100000000111011
11111111110000011011111100101111111111100000001001
00000000000000011110011111100011000000010011111000
00111111100000111100000000000000000000011111100001
00000000111100011100111111100000011100010111111100
10000000001111110000000001011111111000000011100000
01111111111111110000011000000011111111111101111100
00000000110101100000111101111111000011111000111100
01100000000111100111000001110000000111000111001111
10011111000000110001111110001111111100001100000001
11110101111111001100000011100000111111110111111111
00000111111101110001111000000001100000000000110001
```

図 4　遷移確率 20%のマルコフモデルで生成された文字列

```
00000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000
00000000000000000000000000000000000000000000000000
00000000111111111111111111111111111111111111111111
11111111111111111100000000000000000000000000000000
00000000000000000111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
11111111111111111110000111111111111111111110000000
00000000000000000000000000000000000000000000000000
00000000000000001111111111111111111111111111111111
11111111111111111111111111111111111111111111111111
```

図 5　遷移確率 20%のマルコフモデルで生成された文字列

表 2　重複を許した場合の算出された情報量

	テストデータ 1(20%)	テストデータ 2(1%)
データ群 1(20%)	806	636
データ群 2(20%)	797	577
データ群 3(20%)	821	644
データ群 4(1%)	369	37
データ群 5(1%)	376	46
データ群 6(1%)	371	40

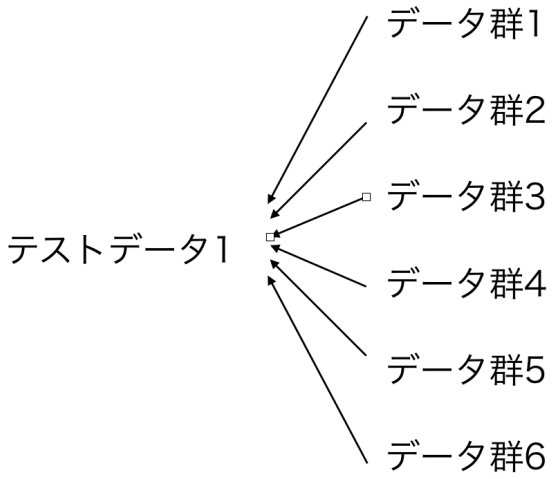

図 6　テストデータ 1 の情報量計算手法

5. 実　　験

5.1 実験方法

実験は予備実験と同様に行う．マルコフ過程を用いて生成した 2 つのデータセットに対し，2 つの遷移確率を用いて作成されたデータ群を 3 セットずつ，計 6 セットを用いてテストデータに出現する部分文字列の頻度を計算する．算出された 6 つの情報量のうち最も小さいデータ群のラベルをテストデータのラベルとする．情報量を計算する際，頻度計算において部分文字列が重複しないように計算を行う．

5.2 評価方法

予備実験では，同一パターンが重複しやすい文字列を用いた場合の情報量が著しく低くなるという問題があった．そこで重複を回避して頻度を計算することでこの問題が改善されるかを検討する．提案手法により問題が改善されれば，遷移確率が 20%のマルコフ過程を用いて作成された文字列において，自身と異なる遷移確率であるデータ群 4 から 6 の情報量が増加すると考えられる．

5.3 実験結果と考察

テストデータの情報量計算をする際，出現した部分文字列の頻度を重複がないように計算した実験を行なった．実験の結果を表 3 に示す．表 2 と同様に小数点以下を切り捨てた情報量であるが，確率の推定のための頻度の数え方が異なるので表 2 と異なる値となる．予備実験の結果と比較すると，テストデータ 1 は表 2 では遷移確率が 1%のデータ群を用いて情報量を計算したものが小さかったものの，頻度計算を改善した表 3 では 2 つの情報量はほぼ等しい．最も情報量が小さいデータのラベルは遷移確率が 1%のものであるが，元の情報量に比べ情報量が増加していることから，頻度が重複してカウントされることによる影響は大きいと考えられる．同様にテストデータ 2 でも計算された情報量が，頻度の計算方法を改善する前に比べ増加していることがわかる．これらの結果から，文字列の頻度を計算する推定問題において，頻度が重複される影響は大きいため，重複を回避することが有用であると考えられる．

表 3　重複を許さない場合の算出された情報量

	テストデータ 1(20%)	テストデータ 2(1%)
データ群 1(20%)	1060	753
データ群 2(20%)	1005	690
データ群 3(20%)	1006	747
データ群 4(1%)	1009	103
データ群 5(1%)	999	105
データ群 6(1%)	1004	103

6. ま と め

本研究では，データのラベル推定において対象となるデータに出現する文字列の頻度を計算する際，頻度が重複して計算されることの影響とその解決法を検討した．予備実験では，遷移確率が 1%と 20%のマルコフ過程に基づいて，2 値の文字からなる 2 つのテストデータを作成した．2 つのテストデータに出現する全部分文字列の頻度を同様の遷移確率で生成したデータ群を用いて計算し，情報量を求めた．その結果，頻度を重複して数えた場合，同一の文字が続きやすいデータの情報量が著しく低くなることがわかった．この原因として，特定の文字列が重複してカウントされることでその文字列の頻度が高くなり，アルゴリズムが文字列の重要性を誤認したのではないかと考えた．そこで我々は，頻度を計算する際に重複する可能性がある場合，重複計算を回避して頻度を計算することを提案した．テストデータに出現する部分文字列の頻度を計算する際，重複が生じた場合はカウントを回避し情報量を計算する実験を行った．実験の結果，重複を許す頻度計算を用いた情報量に比べ情報量が増加し，著しく情報量が低下するという問題が解決された．この結果から，重複して計算される可能性がある文字列は頻度計算の重複を回避することがモデル推定に有用であると考えられる．

今後の課題として，頻度計算においてより効率的に重複したカウントを回避することが挙げられる．また，本研究で用いたテストデータはマルコフ過程によって生成されたデータである．従って，実際に取得されたセンサーデータのラベル推定においてどのような影響があるのかを検討する必要がある．

文　　献

[1] 菊地誠，阿部洋丈，岡部正幸，梅村恭司.”Compression-based Dissimilarity Measure (CDM) を用いた人感センサ情報の類似判定”, 情報学ワークショップ 2009(WINF2009) 論文集, (2009), pp.185-188.

[2] C. D. Manning, H. Schu tze et al., Foundations of statistical natural language processing. MIT Press, 1999, vol. 999, pp. 6163.

[3] 中川聖一. 情報理論の基礎と応用. 近代科学社, 1992.

[4] A. Takamoto, M. Yoshida, K. Umemura, and Y. Ichikawa, “Computing information quantity as similarity measure for music classification task,” in Proceedings of 2017 International Conference On Advanced Informatics:Concepts, Theory And Application (ICAICTA).

[5] 高本綺架, 吉田光男, 梅村恭司, 市川裕子.”作曲者判定タスクのために分析すべき楽曲の長さ”, 情報処理学会研究会報告 SIGMUS116,(2017)