無線センサネットワークを用いた環境モニタリングのためのデータ圧縮手法

柳沢 豊[1] 村松 純[1] 岸野 泰恵[1] 須山 敬之[2] 納谷 太[1]

概要：本稿では，温度，湿度，照度などの環境データを複数の地点で継続的にモニタリングする，無線センサネットワークに適したデータ圧縮方式について提案する．観測地に複数の無線センサノードを設置して環境データを測定するとき，それぞれのセンサノードで観測される値の間には相関があることが多い．この相関を用いたデータの圧縮は，LDPC 符号を用いることで実現可能であることが知られている．しかし，LDPC 符号を用いる場合，符号化されたデータを復号するよりも前に，データ系列内に含まれるシンボルの出現頻度を正確に推定する必要があるという問題があった．そこで本稿では，頻度計算に必要となる一部のデータのみを符号化せず復号器に送信しておき，そのデータを用いて頻度計算を行うことで，出現頻度を高い精度で推定する方法を提案する．

Data Gathering on Wireless Sensor Networks for Environmental Monitoring

Yutaka Yanagisawa[1] Jun Muramatsu[1] Yasue Kishino[1] Takayuki Suyama[2] Futoshi Naya[1]

1. はじめに

本稿では，温度や湿度などの環境データを観測する無線センサネットワークに適した，データの圧縮方式について提案する．対象とする無線センサネットワークは，多数の観測地点に配置された無線通信機能をもつセンサが，観測したデータをマルチホップ通信によって，特定のノード (シンクノード) に集約するネットワークのことである．こうしたセンサネットワークは，環境のモニタリングや，動植物の生態調査などに用いられている [1], [2]．一般に，こうした無線センサネットワークを構成するセンサノードは，バッテリの搭載量が少ない．各センサノードの消費電力を可能な限り抑え，ネットワークのライフタイムを伸ばすことは，無線ネットワークの研究分野における重要な課題の 1 つとなっている．

消費電力を抑える最も効果的な方法は，無線通信にかかる消費電力，すなわち通信コストを抑えることである [3]．無線を用いる場合の通信コストは，一般に送受信データ量に比例し，ノード間の距離の二乗に比例する．つまり，通信コストを下げるためには，通信経路を設計するときに各センサノードがデータの送受信を行う距離を短くすることと，データの符号化技術や集約技術を用いて送受信データ量を削減するという 2 つのアプローチが必要になる．

この課題に対して，Cristescu ら [4]，Liu ら [5] は Networked Slepian-Wolf Coding (NSWC) と呼ばれる 1 つの解決法を提示している．NSWC は，Slepian-Wolf Coding (SWC)[6] と Shortest Path Tree (SPT) を組み合わせた通信路の設計手法である．各センサノードが観測するデータに相関性があるときに，NSWC を用いることで，ネットワーク全体の通信コストを最小化できることが Liu らにより示されている．ここでいうデータの相関性とは，各ノードで観測するデータが一定の確率分布に従うとき，ある二つ以上のノード v_1, v_2 が観測するそれぞれのシンボル（計測値）x_1, x_2 の出現する確率 $p(x_1), p(x_2)$ とこれらの結合確率 $p(x_1, x_2)$ について $p(x_1) \cdot p(x_2) \neq p(x_1, x_2)$ が成り立つ場合があることを言う．本稿で扱う相関度は，確率変数 X_1, X_2 の相互情報量 $I(X_1; X_2)$ と同じ意味である．相関性があるということは，すなわち $I(X_1; X_2) > 0$ である

1 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
2 国際電気通信基礎技術研究所 認知機構研究所

ことを意味する．例えば，屋外で雨量や温度，湿度といった気象データを収集する場合，各センサノードで観測されるデータ系列間には相関性が存在することが多い．NSWC を用いると，こうした相関性のあるデータを効率的に集約することができる．

NSWC で設計した通信路の通信レートを実現するデータの圧縮手法として，低密度パリティ検査 (LDPC) 符号 [7], [8] がある．LDPC 符号を使うことで，NSWC で与えられる理論的な下限に近い通信レートまでデータを可逆圧縮できることが知られている [9]．しかし，実際の無線センサネットワークで LDPC 符号を用いたデータ圧縮手法を用いて NSWC を実装するためには解決すべき問題がある．

LDPC 符号では復号する時にデータ系列に出現するシンボル (数値) の出現確率を使用する t ため，復号前に復号器においてシンボルの出現確率が既知である必要がある．復号時に用いる出現確率の情報が実際のデータの出現確率と異なる場合，復号に失敗することがある．復号に失敗した場合，データを再度符号化して送信するという再送プロセスが必要になるため，総合的に見た圧縮率は大きく低下することになる．仮に出現確率が時間によらず一定であるようなデータ系列であれば，過去のデータに含まれるシンボルの出現確率を調べる方法で出現確率を求め．これを復号時に使用することで良好な復号率を得ることができる．しかし温度や湿度などのセンサデータでは，シンボルの出現確率は時間によって大きく変動することが普通である．そのため，この方法で求めた出現確率は，復号しようとするデータに含まれるシンボルの出現確率と一致することは稀であり，結果的に復号に失敗する確率が上昇することが多い．

この問題の解決法として本稿では，符号化しようとするデータ系列を一定のルールに従って二分割し，分割された一方のデータを先に復号器に送信してシンボルの出現頻度の推定計算に使用し，他方のデータを LDPC 符号を用いて符号化して送信する方法について述べる．この方法では，符号化しようとするデータ系列の一部を使用してシンボルの出現確率を計算するため，過去のデータを用いる方法に比べて出現確率の推定精度を大きく高めることができる

一方で，一部のデータは LDPC 符号で圧縮が行われないため，全てのデータを利用するよりは圧縮率が低下する可能性がある．ただし，出現確率の推定精度の向上によって復号率は上昇するため，推定精度の低い出現確率情報を用いて全てのデータを符号化した場合より，総合的な圧縮率は高めることができる．この出現確率の推定精度と復号の失敗率の関係を調べるため，筆者らは実際にと屋内で 10 地点において 1 か月間に観測した温度データを用いてシミュレーション評価実験を行った．その結果，過去のデータを用いて出現頻度を推定する場合にくらべ，提案手法ではシンボルの出現確率の推定精度が大きく向上することが分かった．これにより，復号の成功率も 1.5% 程度から 40 % 以上に改善することが分かった．

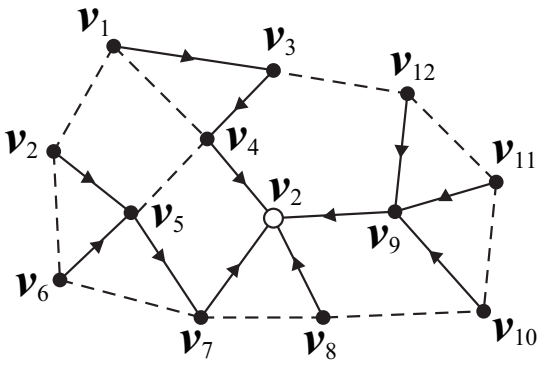

図 1　対象とする無線センサネットワーク
Fig. 1　Wireless Sensor Network

2.　相関性を利用したデータ集約

本章では，観測されるデータに相関性がある無線ネットワークにおいて，通信コストを最小化できるネットワークの設計方法である Networked Slepian-Wolf Coding について説明する．まず 2.1 節 で本稿が対象とする無線センサネットワークについて説明する．次に 2.2 節でデータ集約コストを最小にできるネットワークの設計法である NSWC について述べる．そして 2.3 節では本稿に関する関連研究について触れる．

2.1　無線センサネットワーク

図 1 に示すに s 台のセンサノード (source node) $v_1, \ldots, v_s$ とセンサデータの集約拠点である 1 台のシンクノード (sink node) v_t という $s+1$ 個のノードで構成される無線センサネットワークを考える．s 個のセンサノードとシンクノードの集合を $V = \{v_1, \ldots, v_s, v_t\}$，通信可能なノードの対 v_i, v_j を結ぶ枝 (通信経路) を $e(i,j) \in E$ とすると，この無線センサネットワークは $s+1$ 個のノードをもつグラフ $G(V,E)$ とみなせる．シンクノードもセンサノードのひとつであるため，以下特に断らない限り，センサノードと称するときはシンクノードも含むものとする．各センサノードは，センサを用いて降水量や温度などのセンサデータを取得する機能，そのデータを符号化する機能，および近隣のノードと無線通信する機能をもち，定期的にデータを観測する．各ノードの無線通信機能に性能差はなく，通信可能な距離の上限は同じであると想定する．各センサノードからシンクノードへ直接通信が行えない場合は，他のセンサノードを中継してデータを送る．このとき，無線センサネットワークは分断されておらず，シンクノードから全てのセンサノードに到達できる通信経路があるものとする．ノードの位置は移動しないものとし，通信にかかる時間 (遅延) やスループット，データ送受信の失敗については本

稿では考慮しない．

各ノード　$v_i \in V$　で観測されるセンサデータは，一定時間毎にサンプリングされた離散的な数値列として表す．この数値列は，一定の確率分布に従う無記憶で離散的な情報源 X_i が発生させるものと見なす．このとき，$X_1, \ldots, X_s$ の結合ベクトル $X = \{X_1, \ldots, X_s\}$ は，同時確率分布 $p(X_1 = x_1, \ldots, X_s = x_s) = p(x_1, \ldots, x_s)$ と表せる．時刻を τ，$X(\tau)$ を時刻 τ にすべてのノードで同時に観測された観測値の結合ベクトルとする．なお，本稿ではデータの空間的な（ノード間の）相関性のみに議論の焦点を絞るため，$X_1, \ldots, X_s$ は独立同分布 (i.i.d) であるとする．

次に，データ集約時の通信コストについて定義する．本稿で扱う通信コストとは，通信にかかる電力量のことである．一般に，無線通信によってデータを送受信するために必要となる電力量は，通信距離の 2 乗に比例し，通信するデータの長さに比例する．

二つのノード v_i, v_j を結ぶ通信経路 $e(i,j)$ を流れるデータの, 1 秒あたりの平均情報量(bit) を $b(i,j)$, 通信経路 $e(i,j)$ の距離を $\mathrm{len}(e(i,j))$，通信経路 $e(i,j)$ で要する通信コストを $r(i,j)$ (以降，簡略化のため，$e(i,j), b(i,j), r(i,j)$ をそれぞれ e, b, r と表す) と表記する．このとき，通信経路 e でデータの送受信に要する通信コスト r は $r = \alpha \times b \times \mathrm{len}^2(e)$ で表せる．ここで α はノードの使用する通信手法および，ノードを構成するハードウェア等に依存する定数である．ネットワークを構成するノードのハードウェアおよび通信手法がすべて同種の場合，この定数 α はどのノード間においても同じ値であるとみなせる．このため，本稿では記述の簡素化のために通信コスト r は単に $r = b \times \mathrm{len}^2(e)$ で与えられるものとする．

次に，ネットワーク全体の通信コストについて定義する．まず，グラフ $G(V,E)$ の各通信経路 $e \in E$ に関する通信コスト r の集合を W とする．シンクノード v_t との間に直接通信経路を持たないセンサノード v_i は，他のセンサノードを経由してデータを送信する必要がある．このとき，ノード v_i からシンクノードまで中継される経路の距離の二乗和を d_i と記述する．例えば，v_i が v_t と直接通信する場合は，$d_i = \mathrm{len}^2(e(i,t))$ である．ノード v_j を経由する場合は，$d_i = \mathrm{len}^2(e(i,j)) + \mathrm{len}^2(e(j,t))$ である．このとき，s 個のセンサノードで観測したデータ系列をシンクノードに送るために必要な通信コストの総計 W_G は次式で表せる．

$$W_G = \sum_{i=1}^{s} b \cdot d_i \tag{1}$$

この W_G を最小化できるようにネットワークを設計することで，無線センサネットワーク全体の消費電力を最小化することができる．

2.2 Networked Slepian-Wolf Coding

Networked Slepian-Wolf Coding (NSWC)[4], [5] は，相関性のあるデータを観測するネットワーク $G(V,E)$ において，データ集約時の通信コスト W_G を最小化できるネットワークの設計手法である．通信コスト W_G を小さくするには，次の 2 つのアプローチが必要となる．

- 観測されたデータを符号化してデータの長さを圧縮し，通信経路を流れるデータの平均ビット数 b を小さくする．
- 各センサノードからシンクノードまでの通信経路を決める際に，通信経路の距離 len の二乗の総和を可能な限り小さくする．

NSWC は，通信経路として Shortest Path Tree (SPT)，通信路として　SWC を併用するネットワークの設計方式である．この方法を用いてネットワークを設計することで，通信コスト W_G を最小化できる．この証明は，Liu らによって示されている [5] ため，以下では NSWC の概略についてのみ述べる．

まず, Slepian-Wolf Coding (SWC)[6] について説明する．SWC とは，図 2 のように 2 つの無記憶情報源 X_1，X_2 に対して，それぞれ個別の符号器で符号化し，それら符号を 1 つの復号器で元の 2 つのデータ列に復元する伝送システムのことである．この伝送路では，情報源 X_1, X_2 に対して最適な符号化を行うことで，伝送路の帯域の下限は結合エントロピーと等しくなることが知られている．つまり，情報源 X_1 のエントロピー (平均情報量) を $H(X_1)$，X_1, X_2 の結合エントロピーを $H(X_1, X_2)$　とし，X_1, X_2 に対する伝送路の帯域を R_1, R_2 としたとき，$R_1 + R_2 \geq H(X_1, X_2)$ となるように帯域を設計することで，理論的にはすべてのデータをロスなく伝送できる．情報源 X_1, X_2 の間に相関性がある場合，$H(X_1) + H(X_2) > H(X_1, X_2)$ が成り立つので，X_1, X_2 に対して独立した伝送路を用意する場合（すなわち $R_1 \geq H(X_1), R_2 = H(X_2)$ とする場合）よりも，少ない帯域でデータ列を伝送できる．

Cristescu ら [4] は，情報源のデータ間の相関性があらかじめ既知であれば, 他の情報源のデータを使用せずに各ノードで観測した情報だけを用いて符号化できる (符号化の際にノード間で通信する必要がない) こと，s 個の無記憶情報源 $X_1, X_2, \ldots, X_s$ に対しても最適な符号化を行う場合，エントロピーの総和が結合エントロピー $H(X_1, X_2, \ldots, X_s)$ と等しくできることを利用したネットワークの設計法として，NSWC を提案した．さらに Liu らは NSWC を用いることで, 通信コストの総和を最小化できることを示した．NSWC では，最短経路木を通信経路とする無線通信ネットワーク上において, 相関性を利用して各ノードで生成した符号をシンクノードまで送信する．つまり，グラフ G のシンクノード v_t を根とする SPT をなす部分グラフを $G_{SPT}(V, E_{SPT})$ とし，n 個センサノード $v_1, \ldots, v_s$ が，G_{SPT} 上で v_t への

通信経路の距離が短い順にソートされている場合，各センサノード v_i において，$R_i \geq H(X_i|X_{i-1},\ldots,X_1)$(図 3) となるように帯域を設計する．このように設計すると，通信コストの総和 W_G の下限は次式で与えられる．

$$W_G \geq \sum_{i=1}^{s} H(X_i|X_{i-1},\ldots,X_1) \times d_i) \tag{2}$$

この式で与えられる W_G が，ネットワーク G における通信コストの理論的な最小値である．すなわち，通信経路として SPT を用い，情報源間の既知の相関性を利用して最適な符号化を行えば，通信コストは W_G まで低減できる．

2.3 関連研究

本稿では，NSWC に基づきデータ集約を行う手法について述べるが，NSWC 以外にもデータ集約時の通信コストを低減しようとする研究は，これまでに多数行われている．

例えば，Wavelet 変換等の信号処理に用いられる手法を利用してデータを圧縮する方式 [10], [11]，一定観測されるデータが疎であることを利用してデータの圧縮を行う Compressive Sensing Theory [12] に基づく集約方式 [13], [14] がある．これらは，単独のノードで観測したデータの時間的な相関性を利用したり，シンクノードまでの通信経路上にある他のノードとのデータの相関性を利用してデータを圧縮している．一方で，NSWC は時間的な相関性を考慮せず，ネットワーク全体で観測するデータの空間的な相関性を利用してデータの集約を行う．そのため，これら従来の方法と併用することで，通信コストをさらに削減することが可能であると考えられる．

また，未来の観測データを予測し，その予測の範囲内でデータが観測されている間はデータを送信しないことで送信データ量を削減する方式 [1], [2] もある．これらの手法は，観測するデータが疎であったり，変化の予測が容易なデータであるときには，通信コストの削減に大きな効果がある．しかし，変化の頻度が高い場合には削減効果が薄くなる．NSWC は，こうしたデータの特性に依存することなく，常に低い通信コストでデータの集約を行うことができる．

さらに，ネットワーク全体の通信コストを平滑化することで，ネットワークのライフタイムを最大化することを目的とした研究もある [15]．中でも Luo ら [16] は，データを集約する経路全体に通信コストを分散させることで，通信コストを可能な限り削減しながらコストの平滑化行う方法を示している．これらの手法は通信コストを平滑化できる一方で，ネットワーク全体の通信コストの総和は NSWC より高くなる．このため，ネットワークの運用者が，用途に応じて NSWC とこれらの手法とを選択的に使うことが，実用上は想定される．

3. 解決すべき問題

LDPC 符号を用いることで理論的な最小の符号化レートまでデータを圧縮し，歪みなく復元することが可能である．しかし，これはデータ系列に含まれるシンボルの出現確率に時間的な定常性，および同時性が仮定できる場合である，実際のデータではこの仮定が成り立たないため，理論的な最小の符号化レートまで圧縮することは難しい．以下，この問題について説明するために，まず LDPC 符号について概略を説明する．次に定常性，同時性が実際のセンサデータで仮定できないことにより生じる問題についてそれぞれ説明する．

3.1 低密度パリティ検査符号 (LDPC)

NSWC は，理論的に符号化レートを最小化できるネットワークの設計法を与えるものである．しかし，NSWC によって与えられる通信コスト W_G でデータを伝送するための，データの符号化手段については NSWC では与えられない．この通信レートを実現できる符号化方式としては，低密度パリティ検査 (LDPC: Low Density Parity-Check) 符号が知られている．LDPC については文献 [7] が詳しい．ここでは，LDPC について以降の議論に必要な最小限の説明を行う．

以下ではまず，二元線形符号としての LDPC 符号を用いて，ひとつのセンサノード X が観測するデータの系列 $\vec{X}$ を符号化する場合について説明を行う．A を，各要素が 1 もしくは 0 である $m \times n (0 < m < n)$ 行列とする [*1]．$\vec{x}$ を長さ n のデータの系列とする．行列 A に含まれる 1 の数が，A の要素数に比べて極めて少ないものとする．$\vec{X}$ は $\vec{X} = (x_1, x_2, \ldots, x_n)$ で与えられ，各 x_i の取りうる値は 0 または 1 である．すなわち $x_i \in \{0, 1\}$ である．このとき，データ系列 $\vec{X}$ に対する LDPC 符号の符号語 $c(\vec{X})$ は次の式で与えられる．

$$c(\vec{X}) = A\vec{X}^T$$

符号語 $c(\vec{X})$ の長さは m であり，各要素の取りうる値は 0 もしくは 1 である．符号語 $c(\vec{x})$ の復号には，符号語と行列 A，データ系列の各シンボルの出現確率 $p(x=0), p(x=1)$ の情報が必要になる．これらの情報を入力情報とし，sum-product アルゴリズム [7] を利用することで復号できる．このアルゴリズムでは復号時に行列 A の情報を用いる．このとき，LDPC 符号によるデータの圧縮比率は行列式 A のサイズ $n.m$ により m/n で与えられる

[*1] LDPC は一般にはエラー訂正符号として用いられるが，本稿ではデータの可逆圧縮のための符号として用いる．このため，検査行列にはエラー訂正ビットを含まないものとして扱う

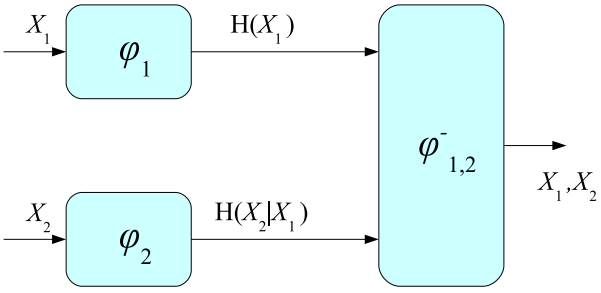

図 2 ノードが 2 点の場合の Slepian-Wolf Coding
Fig. 2 Slepian-Wolf Coding for two nodes

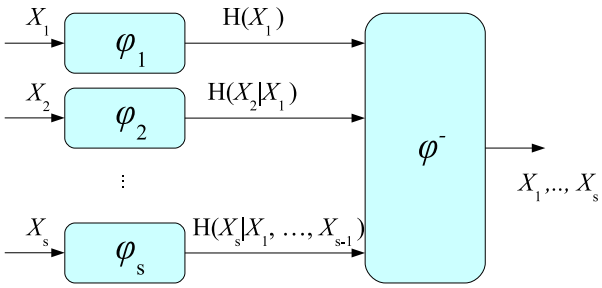

図 3 ノードが s 点の場合の Slepian-Wolf Coding
Fig. 3 Slepian-Wolf Coding for s nodes

3.2 データの非定常性に起因する問題

まず，データ系列が p 個存在し，それぞれのデータ系列を $X_1, \ldots, X_p$ とする．データ系列 X_i が出力する，時区間 $T = [t, t+1, t+2, \ldots t+n-1]$ のデジタル信号の系列を $X_i(T)$ とする．以降，記述の簡易化のために $X_i(T)$ を単に X_i と表記し，デジタル信号の系列を単に「デジタル信号」を表記する．LDPC 符号を用いて $X_i(T)$ を符号化して得た符号語を $\phi(X_i)$ と表記する．データ系列 X_i において LDPC の符号化に用いる検査行列 A のサイズを $m \times n$ (m 行 n 列) とする．n デジタル信号の長さであり，固定の値である．データ系列 X_i の検査行列の行の数 m_i は次の条件を満たすように決める．

$$m_i = nR_i \geq nH(X_i|X_{i-1}, X_{i-2}, \ldots, X_1) \qquad (3)$$

この式 3 で与えられる符号化レートは，データ系列に対して一定である場合，すなわちデジタル信号中に出現するシンボルの生起確率の定常性が仮定できる場合においてのみ，歪みなく復号ができる下限値を示している．しかし，現実のセンサから観測されるデータに含まれるシンボルの生起確率について，定常性を仮定することは難しい．定常性が保証されないデジタル信号に対しては，この式で与えられる符号化レート R_i で信号を送信した場合であっても，歪みなく復号できることは保証されない．

データ系列を歪みなく復号できるようにするためには，条件付きエントロピー $H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$ の変動に対応できるように，符号化レートを式 3 で与えられる R_i よりも高く設定する必要がある．具体的には，符号化レートが $H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$ よりも小さく設定された場合は，復号器が X_i を誤って復号する確率が極めて大きくなる．符号化レートの増大を防ぐためには，$H(X_i|X_{i-1}, X_{i-2}, \ldots, X1)$ が小さくなることが好ましい．

3.3 データ間の同時性が保証されない問題

LDPC 符号では，符号化されたデジタル信号 $\phi(X_i(T))$ の復号時に，補助情報として $X_{i-1}(T), X_{i-2}(T), \ldots X_1(T)$ を用いる．しかし，同じ時区間 T に取得したそれぞれのデジタル信号 $X_i(T)$, $X_{i-1}(T)$, $X_{i-2}(T)$, $\ldots X_1(T)$ について，必ずしも相関性が高い，すなわち条件付きエントロピー $H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$ が低い値となるとは限らない．これは，たとえばデジタル信号の観測時に機械的な時差が生じる場合や，観測する信号が空間上をゆるやかに伝播するため，相関性が時差を伴って現れる場合などがある．このような，データの同時性が保証されない場合は，同じ時区間のデジタル信号を補助情報として用いても，小さな符号化レートが得られないことがある．

4. 提案手法

定常性，同時性が仮定できないデータ系列について，条件付きエントロピー $H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$ を小さくするために，本稿では新たに下記の二つの手段を導入する．

(1) データ系列を，エントロピーを計算するために従来法で圧縮して送信するデータ系列と，LDPC 符号を用いて圧縮して送信するデータ系列とに二分割する方法
(2) LDPC 符号で符号化された符号語を復号するときに，従来法で圧縮して送信されたデータを利用して，過去に受信したデータ系列からもっとも条件付きエントロピーが小さくなるデータ系列を検索し，そのデータ列を補助情報に用いて復号する方法

非定常性が高いデータ系列について，定常性が最大限保たれるようにデータ系列をの長さを決定した上でデータ系列を二分し，その一方のデータ系列を使ってあらかじめ条件付きエントロピーを計算する方法を用いることで，データ系列のもつ非定常性が符号化レートに与える影響を抑制することができる．

また，過去のデータ系列からもっとも条件付きエントロピーが低い系列を選択し，それを補助情報として復号を行うことで，符号化レートを小さく保つことができる．さらに，過去に受信したデータ系列が多ければ多いほど，より条件付きエントロピーが低い系列を見つけることができる確率が上昇するため，多数の情報源が存在し，なおかつ長期にわたってデータ系列を受信し続けるほど，符号化レートを小さくできる可能性が増大する．

4.1 符号化および復号の手順

LDPC 符号を用いて，センサノードが観測したデータ系列を符号化して符号語を生成し，シンクノードに無線通信により送信した後，シンクノードの復号器で符号語を元のデータ系列に歪みなく復号する方法について説明する．なお，シンクノードには予め長さ n の時区間 T の複数のデータ系列 $\boldsymbol{X} = \{X_1(T), X_2(T), \ldots X_{i-1}(T)\}$ が保存されているものとする．センサノードでデータ系列 $X_i(T)$ を観測するものとする．以下，混乱のない限り $X_i(T)$ を X_i と記述する．X_i の長さを n (n は自然数) とする．センサノード上のシンクノードは X_i を二分割し，それぞれ $X_i(U), X_i(V)$ ($U, V \subset T$ かつ $U \cap V = \emptyset$ である) とする．$X_i(U)$ の長さを n_U とし，$X_i(V)$ の長さを n_V とする．$n_U + n_V = n$ である．n_U, n_V の値は，あらかじめ決めておき，センサノードとシンクノードで共有しておく．このとき，下記のプロセスにおいて符号化および復号を行う．センサノードとシンクノードは，あらかじめ各 m, n_V に対応する $m \times n_V$ LDPC 行列をひとつずつ保持しているものとする．この $m \times n_V$ 行列を A_m と記す．

(1) センサノードは無歪圧縮器 ψ で $X_i(U)$ を符号化し，符号語 $\psi(X_i(U))$ を得る．

(2) センサノードは $\psi(X_i(U))$ を無線通信によりシンクノードに送信する．

(3) シンクノードは受信した $\psi(X_i(U))$ を復号して $X_i(U)$ を得る．

(4) シンクノード上の復号器は $\boldsymbol{X}$ の中から $H(X_i(U)|X_j(U'))$ を最小とする長さ n_U のデータ系列 $X_j(U')$ を探す．$X_j(U')$ は，次の手順で生成する．

  (a) $\boldsymbol{X}$ の中から時区間の長さが T と等しい時区間 $T' = [t', t'+1, t'+2, \ldots t'+n-1]$ であるようなデータ系列 $X_j(T')$ をひとつ選択する．

  (b) $X_j(T')$ を，それぞれ長さが U, V と等しい時区間 U', V' に分割し，$X_j(U'), X_j(V')$ を得る．分割の方法は，$X_i(U), X_i(V)$ の分割方法に準ずる．

(5) シンクノードは $m/n_V \geq H(X_i(U)|X_j(U'))$ となるように m の値を選択する．

(6) シンクノードはセンサノードに，無線通信により m の値を伝送する．

(7) センサノードの符号化器は m, n_V の値に対応する $m \times n_V$ 行列 A_m を選択し，この行列を使って $X_i(V)$ を符号化する．符号語は $\phi(X_i(V)) = A_m X_i(V)$ で与えられる．$R = m/n_V$ が信号の符号化レートとなる．

(8) センサノードは $\phi(X_i(V))$ を無線通信によりシンクノードに送る．ここで，j は $X_j(U')$ と対応する j である．

(9) シンクノードは符号語 $\phi(X_i(V))$ を A_m と補助情報 $X_j(T')$ を使って復号データ $X'_i(V)$ を得る．

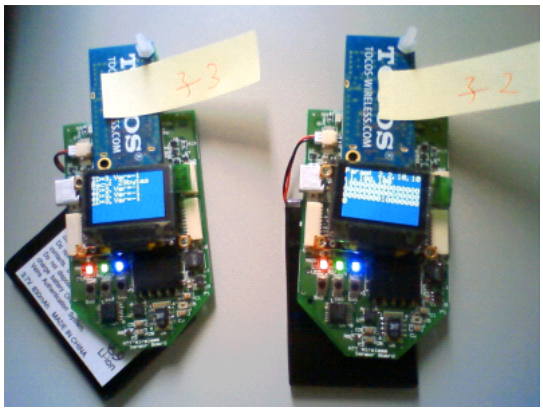

図 4 データ取得に使用したセンサノード
Fig. 4 Sensor Nodes

(10) シンクノードは $A_m X'_i(V) = \phi(X'_i(V))$ が成り立つかを検査する．成り立つ場合，復号に成功したと判断し，$X'_i(V)$ を復元データ $X_i(V)$ とする．成り立たない場合，復号に失敗したと判断し，センサノードに可逆圧縮ができる符号化器 ψ で $X_i(V)$ を再送するように要求する．センサノードはその要求を受けた場合は，ψ で $X_i(V)$ を符号化して得た符号語 $\psi(X_i(V))$ をシンクノードに送り，シンクノードはその符号語を復号して，復元データ $X_i(V)$ を得る．

(11) シンクノードは $X_i(V)$ と $X_i(U)$ を連結して $X_i(T)$ を得る．

以上のプロセスにより，LDPC 符号による可逆なデータ圧縮システムを構築することができる．前項の過程 3 において，データ系列 $X_i(T)$ を $X_i(U)$, $X_i(V)$ に分割する方法は以下の通りである．

予め自然数 k を予め決めておく．たとえば $T = \{1, 2, \ldots, n\}$ であり，$U = \{k, 2k, 3k, \ldots, lk\}$ であるとき，V を T から U に含まれていない要素をすべて含む集合とする．たとえば $k = 2$ であれば，X の要素のうち $2, 4, 6, 8, \ldots$ 番目の要素を集めたものが U であり，$X_i(U) = (x_2, x_4, x_6, \ldots)$ である．このとき $X_i(V) = (x_1, x_3, \ldots)$ である．また $k = 5$ であれば，T の要素のうち 5, 10, 15, 20, ... 番目の要素を集めたものを U とし，それ以外の T 中の要素の集合を V とする．なお，$V = \{x_k, x_{2k}, \ldots\}$ とし，U を T の要素から V の要素をその除いた集合として使用することも可能である．

5. 評価実験

本実験の目的は，提案したシンボルの出現確率の推定精度が，過去のデータを用いた出現確率の推定方法よりも高い精度で確率の推定ができることを確認することである．

5.1 実験内容

実験に用いたデータは，2012 年 7 月 1 日から 7 月 31 日

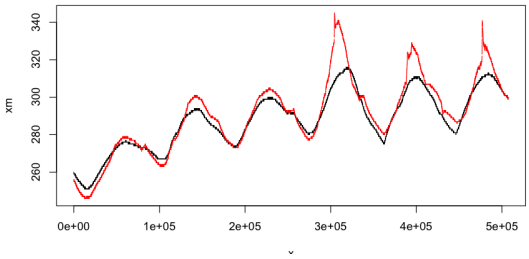

図 5　二台のセンサノードで取得した 6 日間の温度データ

Fig. 5　Temperature data obtained by two sensor node for 6 days

の間に NTT コミュニケーション科学基礎研究所の室内の 10 点にセンサノードを設置し，継続的に取得した温度データである．データを取得したセンサノードを図 4 に，観測されたデータの例を図 5 に示す．観測間隔は 1 分，測定精度は 0.1 度である．これらの取得したデータを使用し，LDPC 符号を用いて圧縮するときに必要となる，シンボルの出現確率を推定するシミュレーション実験を行った．実験では，各センサデータを 120 分に 1 回送信するものし，120 個の連続した数値を圧縮するものとする．10 個のセンサノードの中から，センサノード X と Y を対として選び，Y のデータを先にシンクノードが受け取っており，後から X のデータを LDPC 符号で符号化してシンクノードに送信し，シンクノードでは X を補助情報として用いることで，Y を復号するものとする．

まず，過去のデータを使って頻度を推定する場合について計算を行った．すなわち X を 60 個ずつに区切られた前後のデータ X_u, X_v に分割し，X_u のみを先にシンクノードに送信して $H(X_u|Y_u)$ を計算し，これを $H(X,Y$ の推定値として使用する場合についての推定誤差を調べた．推定誤差 e は下記の式で計算した．

$$e = \frac{|H(X_u|Y_u) - H(X,Y)|}{H(X,Y)} \tag{4}$$

推定が完全に正しく行えた場合は $e = 0$ となる．推定が実際の値と一致しない場合は $0 < e$ となる．

実際のデータを用いて推定した $H(X_u|Y_u)$ の値と $H(X|Y)$ の値との対応を図 6 に示す．図に示す通り，この手法では推定値と実際の値がほとんど相関していないことがわかる．この手法を用いた場合の平均の推定誤差 e は 40.9 % であった．さらに，この方法で計算した $H(X_u|Y_u)$ を用いて LDPC 符号の復号実験も行った．符号化に使用する行列のサイズを予め $120, 60$ となるように作成しておき，$H(X_u|Y_u) < 1/2$ となるデータを選択して符号化，および復号する実験を行った．このときの平均の復号の成功率は 1.5 % 以下となり，ほとんど復号に成功しなかった．これは，$H(X_u|Y_u)$ の推定精度が非常に低いことから，通信レートが適切に推定できていないことが原因であると考えられる．

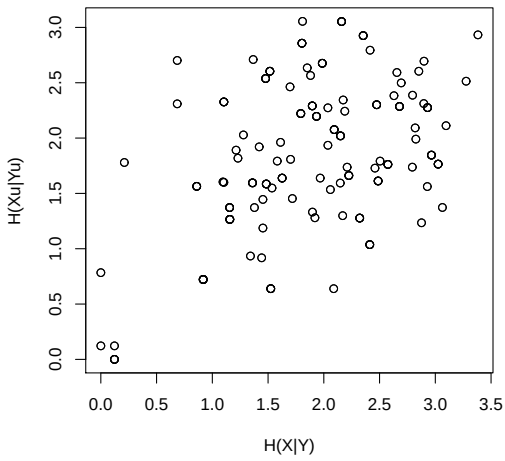

図 6　過去のデータを使用した場合の推定値と実際の値との関係

Fig. 6　Errors of by the existing method

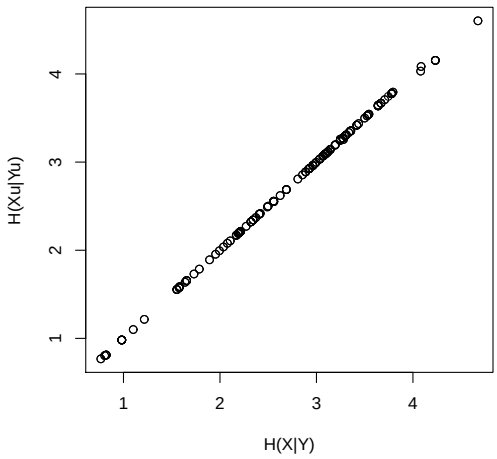

図 7　分割法を用いた場合の推定値と実際の値との関係

Fig. 7　Errors of by our proposed method

次に，提案法で $k = 2$ とした場合の推定誤差を調べた．すなわち X 内の $0, 2, 4, 6, 8, \ldots$ 番目のデータで構成される X_u とそれ以外のデータ X_v とに分割し，X_u のみを先に送信して $H(X_u|Y_v)$ を計算し，これを $H(X,Y$ の推定値として使用する場合についての推定値と，実際との関係を調べ，推定誤差を計算した．Y については，X が観測された時刻と同時刻のセンサデータだけでなく，観測したすべての期間のデータから最も $H(X_u|Y_v)$ が低くなるような Y を選んだ．このときの計算値と実際の値との関係を図 7 に示す．図に示すように，計算値と実際の値が極めて近い値となっている．平均の推定誤差は 0.4% であった．さらに長さ 120 のデータについて，$H(X_u|Y_v) < 1/2$ となるケースにおいて，この分割法により LDPC 符号で符号化，および復号す実験を行ったところ，復号の成功率は 40.3%

となり，過去のデータを用いて推定するよりも大幅に成功率が上昇することがわかった．

6. まとめ

本稿では，複数のセンサノードから観測したデータをセンサノードに集約するセンサネットワークにおいて，データの相関性を利用した圧縮法として LDPC 符号を使用する方法について述べ，符号化に必要となるシンボルの出現頻度の推定法について提案した．提案した分割法を用いることで，従来法よりも推定精度，復号率が大きく改善することが分かった．今後は，復号失敗時の再送プロセスも含めた上で，他手法と比べても高い圧縮率が得られるかどうかを調べる実験を行う予定である．

参考文献

[1] Silberstein, A., Puggioni, G., Gelfand, A., Munagala, K. and Yang, J.: Suppression and failures in sensor networks: a Bayesian approach, *VLDB '07: Proceedings of the 33rd international conference on Very large data bases*, VLDB Endowment, pp. 842–853 (2007).

[2] Leonard, N., Paley, D., Lekien, F., Fratantoni, D., Hole, W., Sepulchre, R. and Davis, R.: Collective Motion and Sensor Networks and Ocean Sampling, *Proceedings of the IEEE, special issue on the emerging technology of networked control systems*, pp. 48–74 (2007).

[3] Akkaya, K. and Younis, M.: A survey on routing protocols for wireless sensor networks, *Ad Hoc Networks*, Vol. 3, No. 3, pp. 325 – 349 (2005).

[4] Baltasar, R. C., Cristescu, R., Beferull-lozano, B. and Vetterli, M.: On Network Correlated Data Gathering, *InfoCom '04: Proceedings of The 23rd Conference of the IEEE Communications Society*, pp. 2571–2582 (2004).

[5] Liu, J., Adler, M., Towsley, D. and Zhang, C.: On optimal communication cost for gathering correlated data through wireless sensor networks, *MobiCom '06: Proceedings of the 12th annual international conference on Mobile computing and networking*, New York, NY, USA, ACM, pp. 310–321 (2006).

[6] Slepian, D. and Wolf, J. K.: Noiseless Coding of Correlated Information Sources, *IEEE Transactions on Information Theory*, Vol. 19, No. 4, pp. 471–480 (1973).

[7] 和田山正：誤り訂正技術の基礎, 森北出版株式会社 (2010).

[8] Liveris, A. D., Xiong, Z. and Georghiades, C. N.: Compression of binary sources with side information at the decoder using LDPC codes, *Communications Letters, IEEE*, Vol. 6, No. 10, pp. 440–442 (2002).

[9] Muramatsu, J., Uyematsu, T. and Wadayama, T.: Low density parity check matrices for coding of correlated sources, *IEEE Transactions on Information Theory*, Vol. 51, No. 10, pp. 3645–3653 (2005).

[10] Gupta, H., Navda, V., Das, S. and Chowdhary, V.: Efficient gathering of correlated data in sensor networks, *ACM Transactions on Sensor Networks*, Vol. 4, No. 1, pp. 1–31 (2008).

[11] Yoon, S. and Shahabi, C.: The Clustered AGgregation (CAG) technique leveraging spatial and temporal correlations in wireless sensor networks, *ACM Transactions on Sensor Networks*, Vol. 3, No. 1, p. 3 (2007).

[12] Donoho, D.: Compressed Sensing, *IEEE Transactions on Information Theory*, Vol. 52, No. 4, pp. 1289–1306 (2006).

[13] Haupt, J., Bajwa, W., Rabbat, M. and Nowak, R.: Compressive Sensing for Networked Data, *IEEE Signal Processing Magagine*, Vol. 25, No. 2, pp. 92–101 (2008).

[14] Rabbat, M., Haupt, J., Singh, A. and Nowak, R.: Decentralized Compression and Predistribution via Randomized Gossiping, *IPSN '06: Proceedings of International Conference on Information Processing in Sensor Networks*, pp. 51–59 (2006).

[15] Liang, W. and Liu, Y.: Online Data Gathering for Maximizing Network Lifetime in Sensor Networks, *IEEE Transactions on Mobile Computing*, Vol. 6, pp. 2–11 (2007).

[16] Luo, C., Wu, F., Sun, J. and Chen, C. W.: Compressive data gathering for large-scale wireless sensor networks, *MobiCom '09: Proceedings of the 15th annual international conference on Mobile computing and networking*, New York, NY, USA, ACM, pp. 145–156 (2009).