多次元時系列データに対する類似部分シーケンス問合わせの高速化

安田　裕真†　　塩川　浩昭††

† 筑波大学情報学群情報科学類　〒 305–8577 茨城県つくば市天王台 1-1-1
†† 筑波大学計算科学研究センター　〒 305–8577 茨城県つくば市天王台 1-1-1
E-mail: †s1911386@s.tsukuba.ac.jp, ††shiokawa@cs.tsukuba.ac.jp

あらまし　多次元時系列データに対する類似部分シーケンス問合せとは，クエリとして入力された多次元時系列データに対して類似度の高い部分シーケンスをデータベースから検索する問題であり，医学やスポーツ科学の分野などで広く利用されている．類似部分シーケンス問合せではデータベース内に存在する全ての部分シーケンスが検索の対象となることから，多次元時系列データのシーケンス長が増加すると膨大な計算時間を要する問題がある．そこで本研究では多次元時系列データにハッシュ処理を行うことで次元削減をする手法を提案する．本研究では人工データと実データを用いた性能評価を行い，従来手法と比較して提案手法は高速かつ正確に類似部分シーケンスを検索できることを示す．

キーワード　時系列データ処理，データ構造・索引，問合せ処理

1　序　　論

多次元時系列データとはある現象の時間的な変化を連続的に観測して得られた多次元の実数値からなる系列（シーケンス）であり，行動分析やスポーツデータ分析などの分野において，重要な要素技術となっている [1]．近年，多次元時系列データ解析において，クエリと類似した部分シーケンスを検索する類似部分シーケンス問合せ処理技術が注目を集めている．例として，図 1 に示すような多次元時系列データとクエリが与えられたとき，類似部分シーケンス問合せ処理はクエリと多次元時系列データの取りうる全ての部分シーケンスとの間で類似度を計算する．その後，類似度が高くなった図中の赤色で示した部分シーケンスを類似部分シーケンスとして出力する．しかしながら，類似部分シーケンス問合せでは多次元時系列データ内に存在する全ての部分シーケンスを検索の対象とするため，長いシーケンスに対して膨大な問合せ処理時間を要する問題がある．

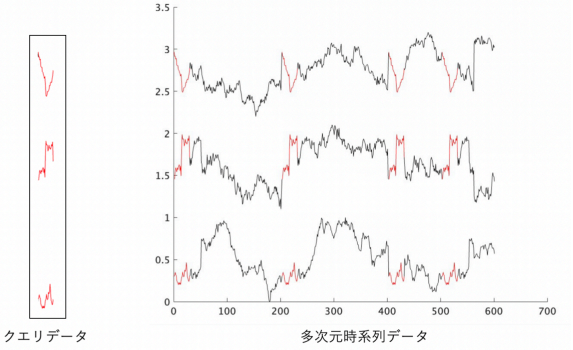

図 1　類似部分シーケンス問合せの例

これまで，時系列データからクエリに類似する部分シーケンスを検索する手法の様々な研究がなされており，Piecewise Aggregate Approximation（PAA）[2] や離散フーリエ変換（DFT）を用いて次元削減を行う手法 [3] などが開発されてきた．しかし，これらの手法は 1 次元の時系列データに対する手法であり，多次元への応用はなされていなかった．また，多次元時系列データから類似部分シーケンスを検出する手法の研究として，Matrix Profile [4] や MD-DTW [5] がある．しかし，これらの手法は多次元時系列データから最も類似した部分シーケンスのペアを検出するという手法であり，クエリに類似する部分シーケンスを検索することはできなかった．

そこで，本研究では多次元時系列データに対する高速な類似部分シーケンス問合せ手法を提案する．多次元時系列データとクエリが与えられたとき，提案手法はクエリとのピアソン相関が閾値 θ 以上の部分シーケンスを全て出力する．提案手法はこの問合せ処理を高速化するために，天方らによる先行研究 [6] を多次元時系列データへと拡張し，多次元時系列データとクエリを事前に次元削減する．その後，次元削減した時系列データを用いて問合せ処理を行うことで，全ての部分シーケンスを探索すること無く問合せ処理を行う．より具体的には，まず提案手法は時系列データの各部分シーケンスに対して局所性鋭敏型ハッシュ関数（LSH）[6] [7] を用いてハッシュ値の計算を行う．このとき，ハッシュ値が類似したデータを同じグループに格納することで，問合せ処理時において全ての部分シーケンスを探索することを回避し，計算時間の削減を図る．本論文では人工データと実データを用いた提案手法の評価実験を行い，提案手法が従来手法と比較して最大 4 倍高速かつ正確に類似部分シーケンスを検出できることを確認した．

本論文の構成は以下の通りである．２節で前提となる知識と本研究で取り扱う問題について説明し，３節で提案手法について述べ，４節で性能評価結果を示す．５節で関連研究について述べ，6 節で本研究のまとめを行う．

2 事前準備

多次元時系列データ $\boldsymbol{T}$ とは z 正規化された時系列データの集合であり $\boldsymbol{T}=[T^{(1)}, T^{(2)},...,T^{(d)}]$ と表記する．d は多次元時系列データ $\boldsymbol{T}$ の次元数，$T^{(i)}$ は長さ n の 1 次元時系列データとであり，$\boldsymbol{T}\in\mathcal{R}^{(d\times n)}$ である．ここで，$T^{(j)}$ において，先頭から i 番目の長さ m である部分シーケンスを $T^{(j)}_{i,m}=[t^{(j)}_i, t^{(j)}_{i+1}, \ldots, t^{(j)}_{i+m-1}]$ と表す．ただし，$t^{(j)}_{i+k}$ は $T^{(j)}_{i,m}$ の $i+k$ 番目の実数値を表す．また，$\boldsymbol{T_{i,m}}$ は $\boldsymbol{T}$ 内に含まれる先頭から i 番目の長さ m である部分シーケンス集合であり $\boldsymbol{T_{i,m}}=[T^{(1)}_{i,m}, T^{(2)}_{i,m}, \ldots, T^{(d)}_{i,m}]$ である．また，$\boldsymbol{T_{i,m}}$ を $(d\times m)$ 行列とみなしたとき，k 列目の列ベクトルを $\mathbf{t}^k_{i,m}$ とすると $\mathbf{t}^k_{i,m}=(t^{(1)}_{i+k-1}, t^{(2)}_{i+k-1}, \ldots, t^{(d)}_{i+k-1})^T$ と表すことができる．クエリデータ q とは，次元数 d，長さ m の実数値から構成される多次元時系列データであり，$q\in\mathcal{R}^{(d\times m)}$ である．

多次元時系列データに対する類似部分シーケンス問合せとは $\boldsymbol{T}$ とクエリデータ q が与えられたとき，クエリと類似した部分シーケンスを $\boldsymbol{T}$ から検出することである．本研究では部分シーケンス間の類似度としてピアソン相関を採用する．定義は以下の通りである．

定義 1 (ピアソン相関)．2 つの部分シーケンス $\boldsymbol{T_{a,m}}$ と $\boldsymbol{T_{b,m}}$ の間のピアソン相関は以下のように求める．

$$\rho(\boldsymbol{T_{a,m}}, \boldsymbol{T_{b,m}}) = 1 - \frac{|dist(\boldsymbol{T_{a,m}}, \boldsymbol{T_{b,m}})|^2}{2m}$$

ただし，$dist(\mathbf{T}_{a,m}, \mathbf{T}_{b,m})$ は $\mathbf{T}_{a,m}$・$\mathbf{T}_{b,m}$ 間のユークリッド距離であり，以下のように計算する．

$$dist(\boldsymbol{T_{a,m}}, \boldsymbol{T_{b,m}}) = \sqrt{\sum_{k=1}^{m}(t^k_{a,m} - t^k_{b,m})^2}$$

本研究で対象とする類似部分シーケンス問合せ問題を以下のように定義する．

定義 2（類似部分シーケンス問合せ問題）．多次元時系列データ $\boldsymbol{T}$，クエリデータ q，閾値 $\theta\in[0, 1]$ が与えられたとき，ピアソン相関 $\rho(\boldsymbol{T_{i,m}}, q)\geqq\theta$ を満たすような全ての部分シーケンス $\boldsymbol{T_{i,m}}$ を $\boldsymbol{T}$ から検索する．

3 提案手法

本節では提案手法について説明する．本研究の目的は，多次元時系列データに対して高速かつ正確に類似部分シーケンス問合せを行うことである．この目的のために，提案手法では局所性鋭敏型ハッシュ関数（LSH）[6] に基づく問合せ処理の高速化手法を提案する．本節の構成は次の通りである．3.1 節では提案手法の概要について説明する．3.2 節から 3.4 節では提案手法の詳細について説明する．

3.1 提案手法の概要

提案手法の基本アイデアは，多次元時系列データの各部分シーケンスと局所性鋭敏型ハッシュ関数（LSH）を用いて多次元時系列データの次元削減を行うことである．LSH は，類似したデータが高確率で同じハッシュ値をとるような局所鋭敏なハッシュ関数を用いることで，ハッシュ関数を介してデータ検索の高速化を図るものである．天方らは LSH を 1 次元時系列データに適用することで，類似検索の高速化に成功している [6]．提案手法では天方らの手法を多次元時系列データへと拡張する．提案手法は事前に LSH を用いて多次元時系列データ $\boldsymbol{T}$ の次元削減を行うとともに，類似したハッシュ値を持つ部分シーケンス同士をグループ化する．問合せ処理時にクエリデータに対しても LSH を適用し，クエリデータとのピアソン相関が大きい部分シーケンスグループを特定する．より具体的には，提案手法は以下の 3 つの処理で構成される．

(1) ハッシュ処理：$\boldsymbol{T}$ の各部分シーケンス $\boldsymbol{T_{i,m}}$ に L 個のハッシュ関数 $h_1, h_2, \ldots, h_L$ を用いて，ハッシュ値系列 $\boldsymbol{H}=[h_1(\boldsymbol{T_{i,m}}), h_2(\boldsymbol{T_{i,m}}), ..., h_L(\boldsymbol{T_{i,m}})]$ $(1\leqq i\leqq n-m+1)$ を得る．

(2) グループ化：$\boldsymbol{H}$ の各部分シーケンス同士のピアソン相関を求め，相関が強い部分シーケンスを同じグループに格納する．このグループの集合を $\boldsymbol{H}$-*group* と呼ぶ．

(3) 検索：クエリ q に時系列データと同様のハッシュ処理を行い，$\boldsymbol{H}_q$ とする．$T^{(j)}$ において，$\boldsymbol{H}_q=[h_1(q), h_2(q), ..., h_L(q)]$ と $\boldsymbol{H}$-*group* の各グループとのピアソン相関を求め，閾値 θ よりも大きいグループを検索する．検索されたグループ内の部分シーケンスのハッシュ値を元データに復元し，閾値 θ よりも大きいものを再度検索して結果を出力する．

上記の処理のうち，(1),(2) は多次元時系列データが与えられた時点で事前処理し，(3) はクエリデータが与えられた時点で処理を行う．以降の節では (1)〜(3) の詳細について述べる．

3.2 ハッシュ処理

ハッシュ処理では多次元時系列データ $\boldsymbol{T}$ 内の各部分シーケンス $\boldsymbol{T_{i,m}}$ に LSH を適用し，ハッシュ値を得る．本研究では，先行研究 [6] で提案された LSH を拡張し，以下のハッシュ関数を用いる．

定義 3（LSH）．部分シーケンス $\boldsymbol{T_{i,m}}\in\mathcal{R}^{(d\times m)}$ に対して，LSH $h(\boldsymbol{T_{i,m}})$ を以下のように定義する．

$$h(\boldsymbol{T_{i,m}}) = \frac{(\boldsymbol{T_{i,m}}\cdot\mathbf{a})^T\cdot\mathbf{b}+cw}{w}$$

ただし，$\mathbf{a}\in\mathcal{R}^{(d\times 1)}$，$\mathbf{b}\in\mathcal{R}^{(m\times 1)}$ はそれぞれ各要素が正規分布に従うランダムな値を持つ列ベクトルである．c は $[0, w)$ から選ばれたランダムな実数であり，w は定数である．上式の通り，LSH $h(\boldsymbol{T_{i,m}})$ は最終的に，1 つの実数値を出力する．

ここで，ピアソン相関の閾値を θ とおくと，本研究では $\rho(\boldsymbol{T_{a,m}}, \boldsymbol{T_{b,m}})\geqq\theta$ となるような類似部分シーケンスに注目し，これらのハッシュ値が類似するようにしたい．そこで先行研究 [6] に従い，$w=\sqrt{2m(1-\theta)}$ とする．

提案手法は定義 3 に示したハッシュ関数を L 個用意して，各部分シーケンスに対して L 個のハッシュ値 $\boldsymbol{H}=[h_1(\boldsymbol{T_{i,m}}), h_2(\boldsymbol{T_{i,m}}), ..., h_L(\boldsymbol{T_{i,m}})]$ $(1\leqq i\leqq n-m+1)$ を生成する．これにより，サイズ $L\times(n-m+1)$ の行列であるハッシュ値系列 $\boldsymbol{H}$ を得る．

3.3 グループ化

グループ化では，ハッシュ値系列 $\boldsymbol{H}=[h_1(\boldsymbol{T}_{i,m}),\ h_2(\boldsymbol{T}_{i,m}),\ ...,\ h_L(\boldsymbol{T}_{i,m})]$ $(1 \leqq i \leqq n-m+1)$ の各部分シーケンス同士のピアソン相関を求め，相関が強い部分シーケンスを同じグループに格納する．具体的には $\boldsymbol{H}$ の各列間のピアソン相関を求める．このとき，2 つの部分シーケンスが $\rho \geqq \theta$ となる場合，同じグループに格納していく．この操作を全ての部分シーケンスの組合せに対して実行し，ピアソン相関の大きな部分シーケンス同士を同じグループに格納する．グループ化によって作成されたグループの集合を $\boldsymbol{H}$-*group* と呼ぶ．

3.4 検　　索

クエリデータ q が到着したら，時系列データと同様のハッシュ処理を q に行い，$\boldsymbol{H}_q=[h_1(q),\ h_2(q),\ ...,\ h_L(q)]$ とする．$\boldsymbol{H}_q$ と $\boldsymbol{H}$-*group* の各グループとのピアソン相関を求め，閾値 θ よりも大きいグループを検索．検索されたグループの集合を $\boldsymbol{H}_q$-*group* とする．ここで $\boldsymbol{H}_q$-*group* 内の各部分シーケンスと $\boldsymbol{H}_q$ を元のデータに復元する．同様に復元したクエリと元の部分シーケンスのピアソン相関を計算し，ピアソン相関が閾値 θ よりも大きい部分シーケンスを出力する．

4 評 価 実 験

提案手法の検索時間と処理精度の評価を行うために，比較手法として 2 つの手法を用意する．1 つ目の手法は，クエリデータと多次元時系列データの全ての部分シーケンスとのピアソン相関を計算するという手法である．これをベースライン手法と呼ぶ．2 つ目の手法は，ハッシュ処理したクエリデータ $\boldsymbol{H}_q$ とハッシュ値系列 $\boldsymbol{H}$ の全ての部分シーケンスとのピアソン相関を計算するという手法である．これを，ハッシュ検索手法と呼ぶ．検索時間と処理精度を評価するために，乱数を用いて生成した人工データと実データをそれぞれ利用する．各データセットの詳細はそれぞれ 4.1 節と 4.3 節に述べる．また，処理精度として適合率と再現率を用いる．ベースライン手法によって検索された正解データ集合を G，提案手法によって検索された部分シーケンス集合を R とすると，適合率と再現率はそれぞれ以下のように求める．

$$\text{適合率} = \frac{|G \cap R|}{|R|},\quad \text{再現率} = \frac{|G \cap R|}{|G|}$$

提案手法は事前計算部分である **(1) ハッシュ処理**と **(2) グループ化**と問合せ処理部分である **(3) 検索**の 2 つに分けられる．4.1 節では事前処理のコストに関する評価を，4.2 節以降は提案手法の問合せ処理の評価実験の結果を示す．

4.1 事前処理について

本節では提案手法とハッシュ検索手法における事前処理に要する実行時間の測定と比較を行う．ハッシュ検索手法の事前処理部分は，提案手法の事前処理部分の **(1) ハッシュ処理**のみである．実験には人工データを使用する．人工データとは，Matlab の randn 関数を用いて生成した乱数データである．実験として，時系列データの長さを変化させた場合と次元数を変化させた場合，用いるハッシュ関数の個数を変化させた場合の実行時間を比較する．実験結果を図 3 から図 5 に示す．図 3 から図 5 より，特に時系列データの長さが事前処理の実行時間に大きな影響を与えることがわかる．

4.2 実行時間の比較

提案手法と比較手法との実行時間の比較実験を行う．実験には人工データを使用する．データの次元数を 3 で固定し，データの長さを 1000, 10,000, 100,000 の 3 種類として検索時間の比較を行なった．実験結果を図 2 に示す．図 2 より，データの長さが 1000, 10,000 と短いときは比較手法と大きな差はないが，100,000 と長いときは．ベースライン手法の約 4 倍，ハッシュ検索手法の約 1.5 倍高速で検索できることがわかる．これはグループ化によって時系列データの次元削減をしたことによるものと考えられる．

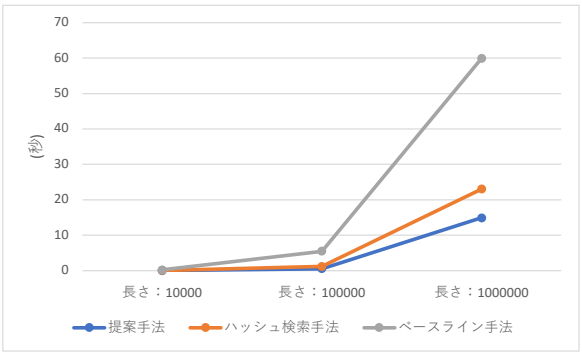

図 2　問合せ処理時間の比較

4.3 問合せ処理の処理精度

提案手法の問合せ処理の処理精度の測定を行う．本実験で用いる実データの詳細を表 1 に示す．また，実験準備として実データの 1 番目のサブシーケンスをクエリデータ q として抜き出す．1 番目のサブシーケンスを抜き出した実データに対して，ベースライン手法を用いてクエリとのピアソン相関を求め，ピアソン相関係数が閾値 θ 以上の部分シーケンスをあらかじめ検索しておく．この実データにおいて提案手法を用いて類似データの検索を行い，各データセットにおいて $L = 10, 25, 50$ についての処理精度の測定を行う．

表 1　データセットの詳細

データセット	n	d	詳細
Appliances Dataset	19,736	7	家電製品のエネルギー使用の予測データセット [8]
Room Dataset	10,000	2	部屋の占有率推定のためのデータセット [9]

実験結果として表 2 から表 5 に示す．表 2 から表 5 より，L=5 では適合率と再現率が下がってしまう場合があったが，L=10, 25 では全てのデータセットで 100%の適合率，再現率を出すことができた．

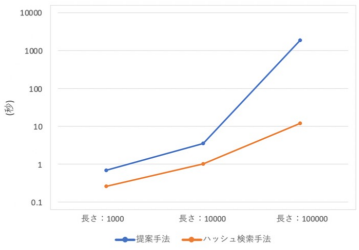

図 3　長さによる影響

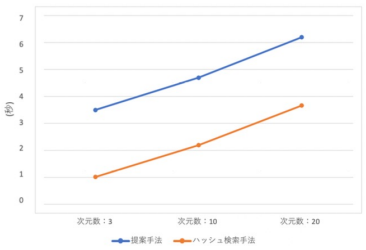

図 4　次元数による影響

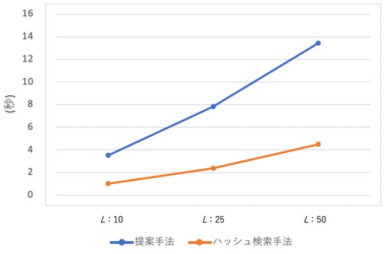

図 5　ハッシュ関数の個数による影響

表 2　Appliance Dataset に対する提案手法の適合率

閾値	L=5	L=10	L=25
θ=0.4	100%	100%	100%
θ=0.3	100%	100%	100%
θ=0.2	100%	100%	100%
θ=0.1	100%	100%	100%

表 3　Appliance Dataset に対する提案手法の再現率

閾値	L=5	L=10	L=25
θ=0.4	100%	100%	100%
θ=0.3	100%	100%	100%
θ=0.2	100%	100%	100%
θ=0.1	93%	100%	100%

表 4　Room Dataset に対する提案手法の適合率

閾値	L=5	L=10	L=25
θ=0.5	100%	100%	100%
θ=0.4	100%	100%	100%
θ=0.3	100%	100%	100%
θ=0.2	98%	100%	100%
θ=0.1	92%	100%	100%

表 5　Room Dataset に対する提案手法の再現率

閾値	L=5	L=10	L=25
θ=0.5	100%	100%	100%
θ=0.4	100%	100%	100%
θ=0.3	100%	100%	100%
θ=0.2	100%	100%	100%
θ=0.1	100%	100%	100%

5　関 連 研 究

Lin ら [2] の研究は，時系列データを等しい長さのセクションに分割し，各セクションの平均値を SAX を用いて記号化するという手法を用いることで類似部分シーケンス問合せの効率化を図った．Faloutsos ら [3] の研究は，時系列データを DFT を用いて特徴空間にマッピングし，MBR を用いて次元削減を行うことで類似部分シーケンス問合せの効率化を図っている．どちらの研究も，1 次元時系列データに対する手法である．

Yeh ら [4] の研究は，多次元時系列データから類似部分シーケンスのペアを検索するための索引構造として Matrix Profile を作成した．Gineke ら [5] の研究は，DTW を計算する際に多次元時系列データの各次元ごとに計算し，合成することで類似部分シーケンスの検出を行なった．しかし，どの手法も多次元時系列データから類似する部分シーケンスを検出する手法であった．

6　結　　論

本研究では事前に多次元時系列データに対して LSH を用いてハッシュ処理を行い次元削減することで，高速で高精度に類似部分シーケンス問合せを行える手法を提案した．評価実験により，提案手法はベースライン手法の約 4 倍ハッシュ検索手法よりも約 1.5 倍高速で高精度に検索できることが示された．

今後の課題として，事前処理方法の改善が挙げられる．提案手法では多次元時系列データが長くなると，事前処理に膨大な時間を要するという問題がある．そこで，事前処理を高速かつ効率的に行えるよう改善すべきだと考えられる．

謝　　辞

本研究の一部は，JST さきがけ (JPMJPR2033) ならびに JSPS 科研費 (JP22K17894) の支援を受けたものである．

文　　献

[1] Ryuichi Yagi and Hiroaki Shiokawa. Fast top-k similar sequence search on dna databases. In *Information Integration and Web Intelligence: 24th International Conference, iiWAS 2022, Virtual Event, November 28–30, 2022, Proceedings*, pp. 145–150. Springer, 2022.

[2] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pp. 2–11, 2003.

[3] Christos Faloutsos, Mudumbai Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. *Acm Sigmod Record*, Vol. 23, No. 2, pp. 419–429, 1994.

[4] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 1317–1322. Ieee, 2016.

[5] Gineke A Ten Holt, Marcel JT Reinders, and Emile A Hendriks. Multi-dimensional dynamic time warping for gesture recognition. In *Thirteenth annual conference of the Advanced School for Computing and Imaging*, Vol. 300, p. 1, 2007.

[6] 天方大地, 原隆浩. 相関時系列データ集合の計算のための高速アルゴリズム. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers, 2017.

[7] 古賀久志. ハッシュを用いた類似検索技術とその応用. 電子情報通信学会 基礎・境界ソサイエティ Fundamentals Review, Vol. 7, No. 3, pp. 256–268, 2014.

[8] Luis M Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, Vol. 140, pp. 81–97, 2017.

[9] Adarsh Pal Singh, Vivek Jain, Sachin Chaudhari, Frank Alexander Kraemer, Stefan Werner, and Vishal Garg. Machine learning-based occupancy estimation using multivariate sensor nodes. In *the 2018 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6. IEEE, 2018.