

多地域 Crowdsensing におけるワーカー最適サンプリング

松浦 千紘[†] 上山 憲昭[†]

[†] 立命館大学 情報理工学部 情報理工学科

〒525-8577 滋賀県草津市野路東1丁目1-1

E-mail: †is0524ee@ed.ritsumeai.ac.jp, ††kamiaki@fc.ritsumeai.ac.jp

あらまし 高性能のセンシング能力を搭載した携帯端末で計測したセンシングデータを、様々なワーカーから収集して真値を推定するモバイルクラウドセンシング (MCS: mobile crowdsensing) の利用が拡大している。2010年代にスマートフォンが広く普及してから、優れたセンシング機能を持ち合わせていることや広範囲から大量にデータを取得できること、インフラの構築が不要なため低コストであることなどからモバイル端末の IoT デバイスとしての活用が期待されており、特にセンサデバイスとしての MCS が注目されている。しかし、実際にはセンサで計測した値は誤差が生じていると予想され、MCS を用いたサービス事業者にとって誤差を抑制する方法の検討は必要不可欠である。既存の研究では、単一地域を対象としていた DPA 法を複数地域に拡張・適用し、推定誤差が最大となるように攻撃者を配置する方法などが提案されているが、複数地域における推定誤差の抑制法については未検討である。また、各エリアに存在している全てのワーカーからデータを収集することを前提とした研究が行われているが、ワーカーにはインセンティブの提供が必要であり、MCS の予算制約から、実際には一定の確率でサンプルしたワーカーからのみデータを収集することが予想される。そこで本研究では、総サンプルワーカー数が固定であるという条件のもと複数エリアのワーカーからデータを収集する際の、各エリアの最適サンプル人数設定法を提案する。サービス事業者はデータを取得するワーカーの総数を設定し、サンプルワーカー数に対する平均推定誤差を格納したデータベースから値を取得、サンプルワーカー数を変化させる操作を行った後の推定誤差の変化量から各エリアのサンプルワーカー数を決定する。このような動作のシミュレーション実験を行った結果、誤差の大きいエリアから多くのデータを収集することで、誤差の抑制効果を確認した。

キーワード モバイルクラウドセンシング, サンプリング

Optimum Worker Sampling in Crowdsecsing with Multiple Areas

Chihiro MATSUURA[†] and Noriaki KAMIYAMA[†]

[†] College of Information Science and Engineering, Ritsumeikan University

1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 JAPAN

E-mail: †is0524ee@ed.ritsumeai.ac.jp, ††kamiaki@fc.ritsumeai.ac.jp

Abstract The use of mobile crowdsensing (MCS), in which sensing data measured by mobile devices equipped with high-performance sensing capabilities are collected from various workers to estimate true values, is expanding. Since the widespread use of smartphones in the 2010s, MCS has attracted particular attention as a sensor device because of its excellent sensing capabilities, its ability to acquire large amounts of data from a wide range of locations, and its low cost because it does not require the construction of infrastructure. However, in reality, it is expected that the values measured by sensors have errors, and it is essential for service providers using MCSs to consider how to control errors. Existing studies have proposed methods such as extending and applying the DPA method from a single area to multiple areas and deploying attackers so that the estimation error is maximized, but methods for controlling estimation error in multiple areas have not been studied. In addition, although studies have been conducted based on the assumption that data is collected from all workers existing in each area, it is necessary to provide incentives to workers, and due to budget constraints of the MCS, it is expected that data will actually be collected only from workers sampled with a certain probability. Therefore, this study proposes a method for setting the optimal number of workers to sample when collecting data from workers in multiple areas under the condition that the total number of sampled workers is fixed. The service provider sets the total number of workers to collect data, obtains values from a database that stores the average estimation error for the number of sampled workers, and determines the number of sampled workers in each area based on the amount of change in the estimation error after changing the number of sampled workers. Simulation experiments of this type of operation confirmed the effectiveness of error suppression by collecting more data from areas with large errors.

Key words mobile crowdsensing, sampling

1. はじめに

高性能のセンシング能力を搭載したモバイル端末で計測したセンシングデータを、様々なワーカーから収集して真値を推定するモバイルクラウドセンシング (MCS: mobile crowdsensing) の利用が拡大している。モバイル端末は今や生活に欠かせないものとなり、高い普及率や扱うデータの幅の広さ、データ量の多さからデータの収集元として注目されている。IoT 機器を街中に設置し各地のデータを収集するという従来の方法では、設置にかかる費用や維持費、台数制限などのコストを考慮する必要があったが、普及率が高く新しくインフラを設置する必要のないモバイル端末を利用したセンシングの活用はこれを解決する。広範囲に偏在しているというモバイル端末の特徴から、実際には MCS サービス事業者は複数のエリアからデータの収集を行う。

しかし、MCS の不特定多数からデータを収集するという特徴によって、収集データに矛盾が発生するといった可能性がある。データを記録したり送信する過程で機器の誤作動などセンサの不具合や、ヒューマンエラーによる誤った測定値の送信が起きることで収集先の推定値に影響を及ぼすといった問題が発生している。特に、意図的に真の値から外れた値を送信し推定値を歪ませる悪意のあるワーカーが存在し、この攻撃は一般的にデータポイズニング攻撃と呼ばれている。

このようなデータポイズニング攻撃に対して推定誤差を抑制するために、ワーカーごとに報告値と推定値との差異に基づく信頼性を推定し、信頼性を重みとした報告値の加重平均を推定値に用いる CRH (Conflict Resolution on Heterogeneous data) 法が提案されている [1]。また、この CRH 法を活用して、悪意のあるワーカーがより推定誤差を大きくするように報告値を設定する DPA (Data Poisoning Attack) 法が提案されている [2]。この他にも攻撃ワーカー、サービス事業者それぞれの視点から攻撃法や防御法が提案されている。これらの研究は、単一のエリアに存在するワーカーからデータを収集して真値を予測することを想定している。しかし実際には、図 1 に示すように、データ収集領域は複数の K 個のエリアから構成され、各エリアにおいてセンシングデータをモバイル端末のワーカーから収集し、エリアごとに測定値を推定するサービスが想定される。そこで筆者らはこれまでに、複数エリアの MCS を対象とした、DPA の最適攻撃者配置法を提案した [7]。

しかしこれらの研究では、収集するエリアに存在している全てのワーカーからデータを収集するモデルを想定している。しかしワーカーにはモバイル端末を操作してデータを収集する負担や、モバイル端末の電力コストが生じることから、ワーカーにデータを提供してもらうインセンティブの提供が不可欠である。MCS を用いたサービス事業者は予算に制約があり、そのため収集するエリアに存在している全てのワーカーからデータを収集することはできず、実際には一定の確率でサンプルしたワーカーからのみデータを収集すると考えられる。

そこで本研究では、総サンプリング数が一定であるという制約条件のもと、複数エリアからデータを収集しエリアごとに真値を推定する場合に、全てのエリアの総推定誤差を最小化するように、各エリアの最適サンプル数を設定するアルゴリズムを提案する。本稿の貢献を以下にまとめる。

- 単一エリアにおいてワーカー数を変化させたときの CRH 法の推定精度を評価し、ワーカーの測定値の分散が CRH 法の推定精度に与える影響を明らかにする。
- 複数エリアの MCS において、各エリアの推定精度の総和が最大化するよう、各エリアにおいてサンプルするワーカー数を最適設計するアルゴリズムを提案する。

- 計算機シミュレーションにより、提案方式の有効性を明らかにする。またワーカーの測定値の分散が最適サンプル数や推定精度に与える影響を明らかにする。

- さらに攻撃者が DPA 法により攻撃を行った場合においても、提案方式を適用することで推定精度の劣化を回避できることを明らかにする。

以後、2. 節で関連研究をまとめ、3. 節で既存の CPA 法と複数エリアを対象として拡張した DPA 法の概要をまとめ、4. 節で CRH 法を単一エリアに適用した場合の結果について考察する。そして 5. 節で本稿で新たに提案する複数エリアの MCS における最適サンプルワーカー数設定法を述べ、6. 節で数値評価結果を示し、最後に 7. 節で全体をまとめる。

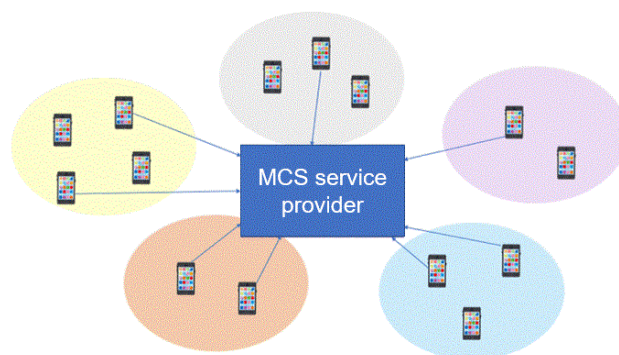


図 1 複数エリアでの MCS サービス

2. 関連研究

これまでに、MCS サービスへのデータポイズニング攻撃方法や、それに対する防御法などが様々な観点から提案されている。

気温や湿度などの連続量を扱う MCS サービスを想定した研究に、MCS で送信される正常ワーカーの報告データを観測し、攻撃者のデータパターンと似たパターンの正常ワーカーデータに malicious データを混在させることで、malicious ワーカーが検知される確率を下げる Poisoning 攻撃が知られている [3]。また Miao らは、複数の各オブジェクトに対し、カテゴリ選択型のデータを複数のワーカーから集め、各オブジェクトの真のカテゴリを推測する場合に、ワーカーの信頼度に応じた重みを考慮する Dawid-Skene モデルをクラウドが用いている場合の攻撃者の最適攻撃法を提案している [4]。さらに Miao らは複数の各オブジェクトに対し、カテゴリ選択型のデータを複数のワーカーから集め、各オブジェクトの真のカテゴリを推測する場合に、(1) 推測結果の誤差を最大化、(2) 設定したターゲットに推測結果を誘導、の二つのタイプの Poisoning 攻撃の最適攻撃法を提案している [5]。

また、これらの攻撃に対する防御方法の研究としては、Huang らの、複数の各オブジェクトに対し真値を推測する場合に、誤差最大化を目的とした Poisoning 攻撃者に対し、推測値からの誤差が閾値を超えたワーカーを除外する提案がある [2]。複数のカテゴリから該当するものを選択する MCS を対象としたものとして、例えば Varshney は複数の報告値から多数決でカテゴリを選択する場合を対象とした Worker 間の結託に対する防御法を提案している [6]。

しかし、これらは単一のエリアにおける MCS サービスを想定しており、複数のエリアにおける MCS を対象とした研究ではない。これに対し我々は、複数エリアでの MCS サービスを想定し、全てのエリアの誤差の最大化と、最も誤差の小さなエリアの誤差の最大化を目的とした攻撃ワーカーの配置方法アルゴ

リズムを提案[7]。しかし本研究では全ワーカーからデータを収集することを想定しており、各エリアのワーカーからランダムサンプルすることを想定していない。

3. MCSに関する既存方式

3.1 CRH 法

CRH 法は、複数の情報源から真実を推測することを目的としたアルゴリズムである[1]。推定値と入力データとの間の全体的な加重偏差を最小化する。その際にワーカーごとの信頼性を推定し、信頼性を重みとした測定値の加重平均を推定値に用いる。MCS に適用した場合、測定値の推定値と報告値との差異が小さいワーカーの信頼性は高く、大きいワーカーの信頼性は低くなるようにワーカーの信頼性を設定する。推定値とワーカーごとの信頼性を交互に更新することで、推定値と真値の差を最小化する。以下に CRH 法のアルゴリズムを示す。ただし正常ワーカーの集合を N 、攻撃ワーカーの集合を A とする。

- (1) 各ワーカー k の信頼性 (重み) w_k を 1 に初期化
- (2) 式 (1) で、各ワーカー k の報告値 v_k と w_k から、推定値 $\tilde{v}$ を計算

$$\tilde{v} = \frac{\sum_{k \in N \cup A} v_k w_k}{\sum_{k \in N \cup A} w_k} \quad (1)$$

- (3) 式 (2) でワーカーごとの信頼性 w_k を更新

$$w_k = -\log \frac{(v_k - \tilde{v})^2}{\sum_{k \in N \cup A} (v_k - \tilde{v})^2} \quad (2)$$

- (4) $\tilde{v}$ 及び w_k が収束するまで step 2, step 3 を反復

3.2 DPA 法

DPA 法は、CRH 法による推定値及びワーカーごとの信頼性の更新ステップ、攻撃ワーカーの報告値の更新ステップの 2 ステップを繰り返すことで、推定誤差を最大化する攻撃ワーカーの最適報告値設定アルゴリズムである[2]。攻撃者の最適化目標は、正常ワーカーのみで推定した値 $\tilde{v}$ と、攻撃ワーカーを含めた全ワーカーで推定した値 $\hat{v}$ の絶対差を最大化することである。単一エリアを対象とした DPA 法のアルゴリズムを以下に示す。

- (1) 各攻撃者 k の報告値 v_k の初期化
- (2) 正常ワーカーのみで CRH 法を用いて推定値 $\tilde{v}$ を算出
- (3) 全ワーカーを対象に CRH 法を用いて推定値 $\hat{v}$ を算出
- (4) 各攻撃ワーカー k に対し、式 (3) で報告値 v_k を更新

$$v_k = v_k + 2 \times (\hat{v} - \tilde{v}) \times \frac{w_k}{\sum_{k \in N \cup A} w_k} \quad (3)$$

- (5) v_k が収束するまで step 3, step 4 を反復

3.3 誤差を最大化する最適ボイズニング攻撃

複数エリアの MCS サービスを対象とした最適データボイズニング攻撃法について、DPA 法を複数エリアに拡張したアルゴリズムに誤差の上限値と下限値を設定することで、総推定誤差の最大化と攻撃ワーカーの分散配置の両立を実現する[7]。

動作について示す。全攻撃ワーカーを未配置状態に初期化し、未配置攻撃ワーカーの中から仮配置攻撃ワーカー m を決定する。エリア s に仮配置し、DPA 法を用いて正常ワーカーのみを用いて算出した推定値 $\tilde{v}_s$ と、全ワーカーを対象に算出した推定値 $\hat{v}_s$ を算出。その差分の絶対値をエリア s の誤差 e_s とする。複数エリアを対象とした攻撃者配置のアルゴリズムを以下に示す。

- (1) 全攻撃者を未配置状態に初期化

- (2) 未配置攻撃者の中から一人、攻撃者 m を選択
- (3) 各エリア s を選択し、 m を仮配置
- (4) エリア s に DPA 法を用いて、正常ユーザのみを用いて算出した推定値 $\tilde{v}_s$ と、全ユーザを対象に算出した推定値 $\hat{v}_s$ を算出
- (5) エリア s に攻撃者 m を配置したときのエリア s の誤差 e_s を算出
- (6) e_s が最大となるエリア s に攻撃者 m を配置
- (7) 全攻撃者を配置し終わるまで step 2 ~ step 6 を反復

ここで、各エリア s に誤差の上限値 η_u と下限値 η_l を設定する。誤差上限法 ULM では、 $e_s < \eta_u$ のエリアに攻撃ワーカーを仮配置し、それを上回るエリアは攻撃ユーザ仮配置対象から除外する。 $e_s < \eta_u$ のエリアの中で、 e_s の増加量が最大となるエリアに攻撃ワーカーを追加配置する。この処理を全てのエリアが上限値に達するまで反復する。全エリアが上限値に達し、未配置の攻撃ワーカーが存在する場合、 e_s の増加量が最大となるエリアに追加配置する処理を行う。誤差下限法 LLM も同様に、 $e_s < \eta_l$ のエリアに攻撃ワーカーを仮配置し、それを上回るエリアは攻撃ワーカー仮配置対象から除外する。LLM では、 $e_s < \eta_l$ のエリアの中で、 e_s が最小となるエリアに攻撃ワーカーを追加配置する。この処理を全エリアが下限値に達するまで反復して行い、下限値に全エリアが達し、かつ未配置攻撃ワーカーが存在する場合、 e_s の増加量が最大となるエリアに追加配置する処理を行う。このようにして全攻撃ワーカーを配置し終わると、処理を終了する。

図 2 に、上限値 η_u と下限値 η_l に対して、各手法の総誤差 E と最小誤差 $e_{\min}$ をプロットする。誤差上限法 (ULM) においては、 η_u が小さい領域では、すぐに全エリアの e_s が η_u に達するため、攻撃ワーカーは e_s の増加量が大きなエリアに集中的に配置され、 E が大きく $e_{\min}$ が小さい。また η_u の増加に伴い、より e_s の増加量が小さいエリアに攻撃ワーカーが多く配分される結果、 E は減少していく半面 $e_{\min}$ は増加し、攻撃ワーカーの偏りは減少する。さらに η_u を増加させていくと、再度、より多くの攻撃ワーカーを e_s の増加量が大きなエリアに割り当て可能となり、 η_u の増加に伴い E が増加するが、 $e_{\min}$ は減少し、配置攻撃ワーカー数の偏りがより大きくなるといった傾向が見られた。

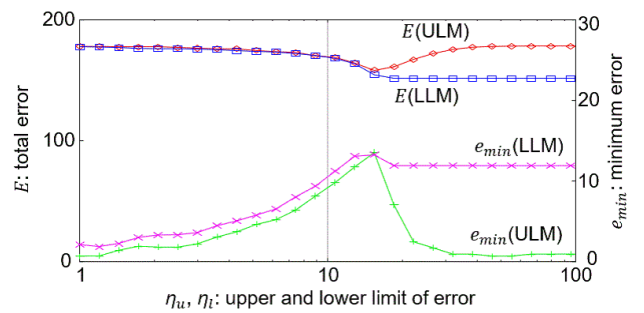


図 2 全エリア同一条件

一方、誤差下限法 (LLM) においては、 η_l が小さい領域では、すぐに全エリアの e_s が η_l を上回るため、攻撃者は e_s の増加量が大きなエリアに集中的に配置され、 E が大きく $e_{\min}$ が小さい。また η_l の増加に伴い、より e_s の増加量が小さいエリアに攻撃者が多く配分される結果、 E は減少する半面 $e_{\min}$ は増加し、攻撃ワーカーの偏りは減少する。さらに η_l を増加すると、より多くの攻撃ワーカーが必要な e_s の増加量が小さなエリアに多くの攻撃ワーカーを割り当てる。そのため η_l の増加に伴

い E が減少するが、 e_{min} が増加し、配置攻撃ワーカー数の偏りも増加する。しかし、さらに η_l を増加させると E と e_{min} は η_l が変化しても一定となるといった傾向があった。これは、どのエリアも η_l の値に届かなくなり、攻撃ワーカーの配置状況が変わらなくなるためである。

正常ワーカー数が各エリアで異なる場合や、正常ワーカーの報告値の標準偏差がエリア毎に異なる場合なども同様の実験を行ったところ、いずれも η_l を 7~12 程度に設定した LLM が攻撃者にとって望ましい攻撃者配置法と考えられる。

4. ワーカー数が CRH 法の精度に与える影響

本節では、サンプルワーカー数が CRH 法の推定精度に与える影響を明らかにするために、単一エリアを対象に CRH 法を適用したときの、正常ワーカー数が推定誤差に与える影響を分析する。実験条件として、エリア数を $K = 1$ 、正常ワーカーの測定値の平均を 50 に設定する。また、正常ワーカー数を $2 \leq n \leq 100$ の範囲で変化させ、推定誤差の推移を観察する。正常ワーカーの測定値の標準偏差を 20, 40, 60 のそれぞれに設定したときの、推定誤差 E をサンプルワーカー数に対して、図 3-5 に各々示す。ただし 1,000 回の試行における平均値を示す。

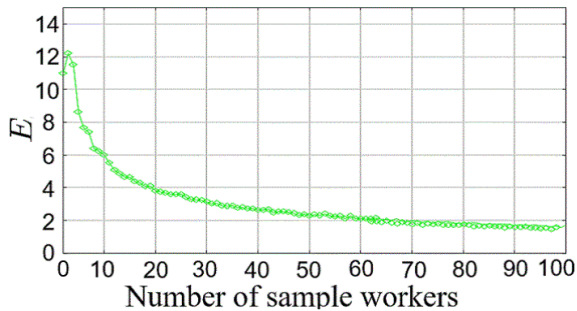


図 3 正常ワーカーの測定値の標準偏差が 20 の場合の推定誤差

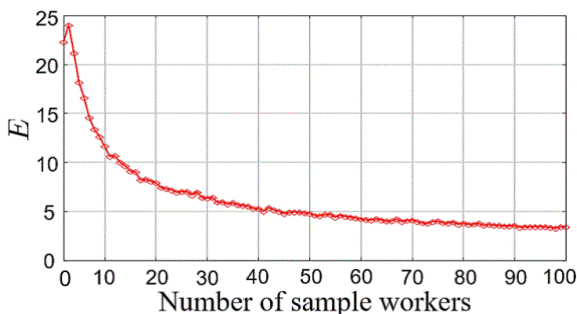


図 4 正常ワーカーの測定値の標準偏差が 40 の場合の推定誤差

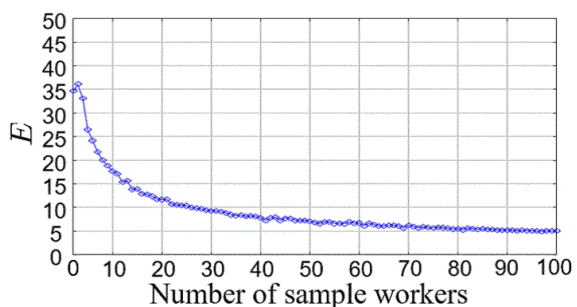


図 5 正常ワーカーの測定値の標準偏差が 60 の場合の推定誤差

どの条件においても、正常ワーカー数の増加に伴い総誤差 E は単調に減少する傾向が見られる。縦軸の推定誤差の値を標準偏

差別に観察すると、標準偏差の小さい、すなわち正常ワーカーの測定値のばらつきの小さい場合ほど、推定精度が高いことがわかる。また、サンプルワーカー数の変化に注目すると、標準偏差の大きいエリアほどサンプリング数を増加させたときの推定誤差の減少量は大きい。

このように CRH 法を単一エリアに適用した場合、正常ワーカーの数が多く、測定値のばらつきが小さいほど推定精度が高くなる。この結果を利用して、複数エリアの MCS サービスにおいて総サンプルワーカー数が固定である場合の、総誤差の最小化を目的とした各エリアのサンプルワーカー数の最適化アルゴリズムを提案する。

5. 複数エリアの MCS におけるサンプルワーカー数の最適化

MCS の運用者は、ワーカーからデータの提供を受けるにはインセンティブをワーカーに支払う必要があるので、総サンプルワーカー数の上限 N を制約条件として考慮し、本条件下で総誤差 E が最小となるよう各エリア i のサンプルワーカー数 u_i を最適設計する。

4 節では、単一エリアで CRH 法を用いて正常ワーカーの人数と推定誤差との関係性を明らかにした。CRH 法を特性の異なる複数の各エリアで独立に適用することを想定する。

5.1 提案方式の概要

総サンプルワーカー数を N 人とし、各エリアのサンプルワーカー数を N/K として初期化する。各エリアのサンプルワーカー数をインクリメント、デクリメントした際の総推定誤差の変化量から、最も総推定誤差が小さくなるように各エリアのサンプルワーカー数を更新する。文献 [1] で提案されている CRH アルゴリズムを収集する全てのエリアに適用し、サービス事業者は一定の間隔でデータを収集、推定値計算を行う。総推定誤差の変化量の近似解を、各エリアのサンプルワーカー数に対する平均推定誤差を格納したデータベースを参照することで得る。サービス事業者の最適化目標として、全てのエリアの誤差の最小化を考える。

5.2 誤差を最小化する最適化問題

エリアの数を K 、エリア i の推定値を v_i 、真の値を p_i とすると、エリア i の推定誤差は $|v_i - p_i|$ となるので、全エリアの誤差の総和 E の最小化を目的とする目的関数は (4) 式で得られる。

$$\min E(u_1, u_2, \dots, u_K) = \sum_{i=1}^K (v_i - p_i)^2 \quad (4)$$

ただし制約条件を次式で与える。

$$\sum_{i=1}^K u_i = N \quad (5)$$

5.3 近似計算アルゴリズム

エリア数や総サンプルワーカー数が多い場合、上記最適化問題の厳密回を現実的な時間で得ることは困難である。そこで、各エリア i の測定値の平均を μ_i 、標準偏差を σ_i とし、以下のアルゴリズムで本最適化問題の近似解を得る。

(1) データ収集する総ワーカー数を N 、各エリアの正常ワーカー測定値の平均を μ_i 、正常ワーカー測定値の標準偏差を σ_i 、エリア数を K とし、これらを入力値として与える

(2) 各エリアのサンプルワーカー数を $u_i = N/K$ とし、これを初期配置状態 S_0 とし、このときの総誤差 E_{ini} を算出

(3) 何回かランダムにサンプルワーカー数を与えたときの平

均推定誤差を算出することで、各エリアのサンプルワーカー数に対する平均推定誤差 e_{i,u_i} のデータベースを作成

(4) ワーカー数調整前の総誤差を E_{pre} とし、各エリア i のサンプル人数をインクリメントしたときの平均誤差 e_{i,u_i+1} をデータベースから取得

(5) 各エリア i のサンプル人数更新後の推定誤差の減少量 $e_{dec} = e_{i,u_i} - e_{i,u_i+1}$ を算出

(6) 各エリア i のサンプル人数をデクリメントしたときの平均誤差 e_{i,u_i-1} をデータベースから取得

(7) 各エリア i のサンプル人数更新後の推定誤差の増加量 $e_{inc} = e_{i,u_i-1} - e_{i,u_i}$ を算出

(8) e_{dec} が最大となるエリアの u_i をインクリメントし、 e_{inc} が最小となるエリアの u_i をデクリメントし、このときの総誤差 E_{post} を算出

(9) 総誤差の変化量 $|E_{post} - E_{pre}|$ が閾値 η を下回るまで、step 4 ~ step 8 を反復

(10) step 9 の条件を満たしたとき (状態 S_1) の各エリアのサンプルワーカー数を最適サンプル人数とし、このときの総誤差 E_{conv} を算出

6. 性能評価

6.1 複数エリアのサンプリング数調整に伴う総誤差の変化

各エリアの測定値の平均 μ_i を 50 とし、各エリア測定値の標準偏差 σ_i を最小 2, 最大 20 の 2 刻みで設定する。各エリアの σ_i の設定値を表 1 に示す。エリア数 10 個、総サンプルワーカー数 400 人より、提案方式を適用すると各エリアの初期配置ワーカー数は 40 人となる。初期配置状態を S_0 と表し、 S_0 の際の総誤差を E_{ini} 、ワーカー調整終了時の総誤差を E_{conv} とする。本節では、上記のパラメータを入力値として、総誤差 E_{ini} から E_{conv} にかけての推移を観察する。ひとつ前の調整時に算出した総誤差を E_{pre} と表し、 E_{post} と E_{pre} の差分の閾値判定に $\eta = 10^{-5}$ を用いる。

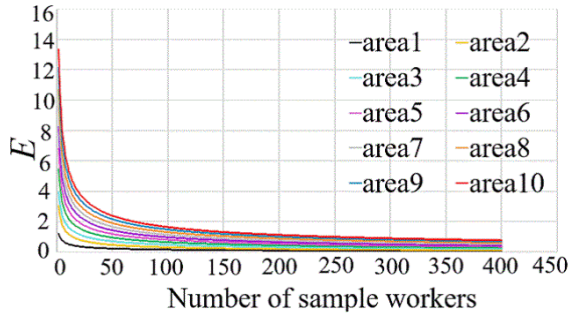


図 6 各エリアのサンプルワーカー数に対する平均推定誤差

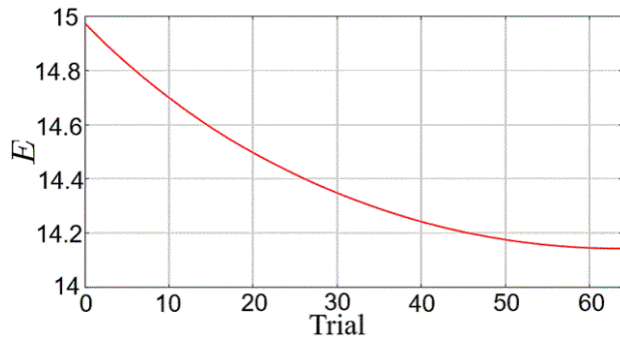


図 7 サンプリング数調整回数に伴う総誤差

図 6 は、提案方式のアルゴリズムの step 3 における、各エリアのサンプルワーカー数に対する平均推定誤差 e_{i,u_i} を示している。各エリアのサンプルワーカー数が同一である場合に u_i をインクリメントすると、減少量の大きい area10 が選択される。また、同様の条件のもと u_i をデクリメントすると、増加量の小さい area1 が選択されることが読み取れる。このように処理を進めると、ある時点で総誤差の値が減少しない、収束状態となる。このときサンプルワーカー数のエリア間の調整を終了する。

図 7 は、提案方式を用いたときの総誤差 E_{ini} から E_{conv} にかけての推移のグラフである。調整回数 (Trial) を重ねるにつれ総推定誤差は単調減少していることがわかり、初期配置状態 S_0 と比較して、提案方式によるサンプルワーカー数のエリア間の調整後 S_1 の総推定誤差は減少することが確認できる。表 1 に状態 S_0 と状態 S_1 の各エリアのサンプルワーカー数を示す。

表 1 提案方式の適用前後の各エリアのサンプルワーカー数

area	σ_i	S_0	S_1
1	2.0	40	13
2	4.0	40	23
3	6.0	40	28
4	8.0	40	34
5	10.0	40	39
6	12.0	40	44
7	14.0	40	48
8	16.0	40	53
9	18.0	40	57
10	20.0	40	61

提案方式を適用すると (S_1)、各エリアのサンプルワーカー数は標準偏差が大きいエリアから順に多くのワーカーが割り当てられるという結果になった。標準偏差が小さいエリアは推定誤差が小さく、また正常ユーザの数が多いほど誤差は小さくなる。一方でサンプルワーカー数の増加に伴い推定誤差は減少するため、誤差の大きなエリアに多くのサンプルワーカーが割り当てられる。このように総サンプルワーカー数を固定した場合、提案方式を用いてサンプルワーカー数を設定すると、ワーカーの測定値の標準偏差が大きいエリアほどより多くのワーカーが配置される。

6.2 総サンプルワーカー数を変化させた場合

図 8 に、6.1 節と同様の数値条件で、総サンプルワーカー数を最小が 100, 最大が 1,000 の 100 刻みで設定したときの、総推定誤差 E_{ini} と E_{conv} の推移を観察したものである。総サンプルワーカー数を増加すると総推定誤差は単調減少し、推定精度は向上することがわかる。

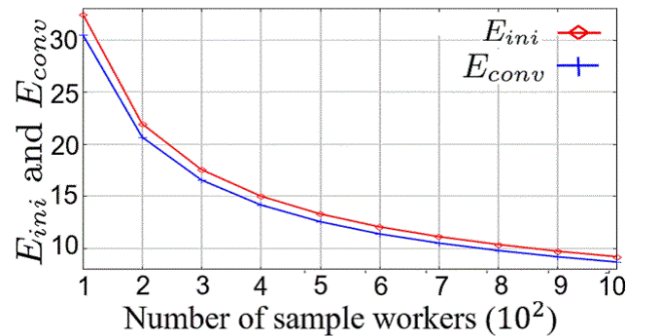


図 8 総サンプルワーカー数を変化させた場合の総推定誤差の推移

また、推定誤差の減少幅 $E_{ini} - E_{conv}$ はサンプルワーカー数の増加に伴い小さくなる。

6.3 攻撃ワーカーがサンプルされた場合

DPA 法を用いて攻撃者が攻撃ワーカーを各エリアに混在させた場合の評価を行う。各エリアの測定値の平均 μ_i を 50，各エリアの標準偏差を最小 2，最大 20 の 2 刻みで設定し，エリア数を 10 個，総サンプルワーカー数を 400 人とする。攻撃ワーカーの測定値の初期値を 50 とし，サンプルワーカー数に対する各エリアの攻撃ワーカーの割合を 0.3 とする。MCS の運用者は正常ワーカーと攻撃ワーカーを区別できないので，各エリアにおいて，各正常ワーカーと各攻撃ワーカーは同じ確率で各々，サンプルされるとする。攻撃ワーカーの測定値の設定には，誤差の最大化を目的としたアルゴリズムである DPA アルゴリズム [2] を用いる。

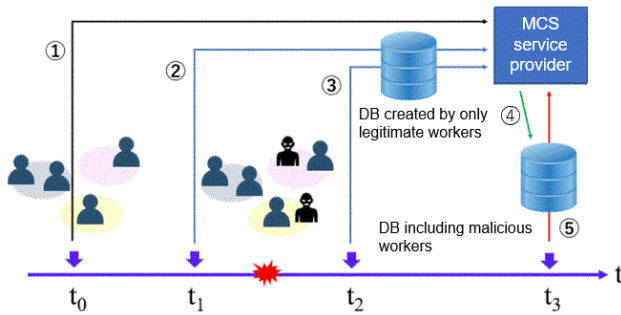


図 9 MCS の測定タイミング

図 9 に示すように，サービス事業者は時刻 $t_0 \sim t_3$ に複数エリアからデータを収集し，推定値計算を行うことを想定する。時刻 t_0 を正常ワーカーのみ，かつ初期配置状態 S_0 とし，提案方式を適用し調整を行った後の状態を時刻 t_1 とする。時刻 t_1 から t_2 の区間でサンプリング前の母集団に攻撃ワーカーが混在し，時刻 t_2 に各エリアのサンプルワーカー数 u_i に対して， $0.3u_i$ の数の攻撃ワーカーがサンプルされると仮定する。サービス事業者が時刻 t_2 に総誤差の増加を確認すると，攻撃ワーカー割合に応じてデータベースの再計算を行う。そして時刻 t_3 では攻撃ワーカー割合を考慮し再計算された値を用いてサンプルワーカー数を決定する。

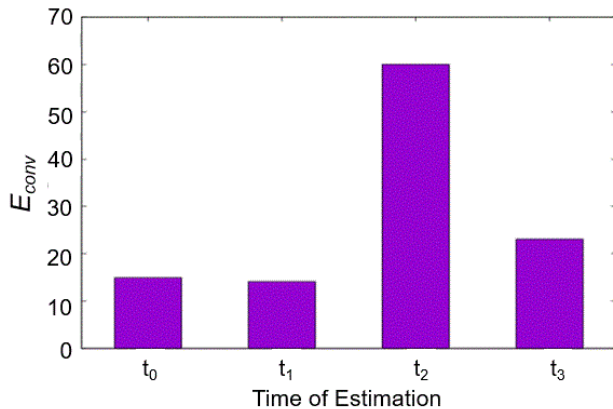


図 10 4 つの各時点における総誤差

図 10 に，4 つの各時点における総誤差をプロットする。時刻 t_0 から t_1 の区間，時刻 t_2 から t_3 の区間にかけて総誤差が減少することが確認できる。提案方式を適用し各エリアのサン

プルワーカー数を最適設計すると，総サンプルワーカー数が同一であっても推定精度を向上することが確認できる。また，サンプルワーカー数をいくつか与え，エリア毎にサンプルワーカー数に対する平均推定誤差をデータベースに格納したが，攻撃ワーカーが一定数存在していると仮定してデータベースの計算を行うと，時刻 t_1 から t_2 の数値から各エリアのサンプルワーカー数が同一である場合においても総誤差の数値が増加していることが確認できた。

7. ま と め

本稿では，総サンプリング数が一定であるという制約条件のもとで複数エリアからデータの収集を行う場合を想定し，単一エリアでワーカー数と推定誤差との関係を実験により示し，その性質を用いて全エリアの総推定誤差を最小化するサンプルワーカー数調整アルゴリズムを提案した。

単一エリアでは，サンプルワーカー数の増加に伴い推定精度は向上する。また，各エリアの特性（ワーカーの測定値の標準偏差）が異なる場合，同一サンプルワーカー数で比較すると標準偏差が小さいエリアほど推定精度が高くなる。複数エリアにおいて，総推定誤差が最小になるようワーカー数を最適化すると，推定精度の低い標準偏差が大きなエリアに多くのワーカーを配置される。提案方式は，攻撃ワーカーが含まれる場合でも有効である。

今後はサンプリングを想定した効率的な攻撃ワーカーの配置方法を提案し，さらにユーザにインセンティブを支払うモデルを考慮したサンプリングワーカーの選択による防御法の確立を目指す。

謝辞 本研究成果は，KDDI 財団研究助成寄付金 190051 の助成を受けたものである。ここに記して謝意を表す。

文 献

- [1] Q. Li, et al., Conflicts to Harmony: A Framework for Resolving Conflicts in Heterogeneous Data by Truth Discovery, IEEE Trans. Know. Data Eng., 28 (8), Aug. 2016.
- [2] Z. Huang, M. Pan, and Y. Gong, Robust Truth Discovery Against Data Poisoning in Mobile Crowdsensing, IEEE GLOBECOM 2019.
- [3] K. Quintal, E. Kara, M. Simsek, B. Kantarci, and H. Viktor, Sensory Data-driven Modeling of Adversaries in Mobile Crowd sensing Platforms, IEEE GLOBECOM 2019.
- [4] C. Miao, Q. Li, L. Su, M. Huai, W. Jiang, and J. Gao, Attack under Disguise: An Intelligent Data Poisoning Attack Mechanism in Crowdsourcing, WWW 2018.
- [5] C. Miao, Q. Li, H. Xiao, W. Jiang, M. Huai, and L. Su, Towards Data Poisoning Attacks in Crowd Sensing Systems, ACM Mobihoc 2018.
- [6] L. Varshney, Privacy and Reliability in Crowdsourcing Service Delivery, SRII 2012.
- [7] R. Fujimoto and N. Kamiyama, Poisoning Attacks in Crowdsensing Over Multiple Areas, IEEE GLOBECOM 2022.