

複数地域の Crowdsensing におけるワーカーの最適サンプリング

Optimum Sampling of Workers in Crowdsensing with Multiple Locations

松浦 千紘

上山 憲昭

Chihiro Matsuura

Noriaki Kamiyama

立命館大学 情報理工学部

College of Information Science and Engineering, Ritsumeikan University

1. はじめに

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。ワーカー推定誤差を最小化するように各ワーカーの測定値を重みづけした重みづけ平均で推定する CRH (Conflict Resolution on Heterogeneous data) 法が提案されている [1]。また筆者らは、複数エリアごとに複数ワーカーから測定値を推定する MCS において、攻撃ワーカーが推定誤差を最大化するように各エリアの配置攻撃者数を最適化する方式を提案した [2]。しかし [1][2] では各エリアに存在する全ワーカーからデータを収集するが、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リングしたワーカーのみデータを収集することが予想される。そこで本稿では、複数エリア MCS において、全エリアの誤差総和の最小化を目的とした各エリアの最適サンプル数設定法を提案する。

2. CRH 法

CRH 法は、複数の測定値から真の値を推測することを目的としたアルゴリズムである [1]。推定値と測定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各ワーカーの信頼性を設定し、信頼性を重みとした測定値の加重平均を推定値として用いる。推定値と各ワーカーの重みを交互に更新し、収束した値を推定値として扱う。ワーカーの k の重みを w_k 、正常ワーカーの集合を N 、攻撃ワーカーの集合を A とすると、各ワーカーの重み w_k と推定値 $\tilde{v}$ は以下の式で導出される。

$$w_k = -\log \frac{(v_k - \tilde{v})^2}{\sum_{k \in N \cup A} (v_k - \tilde{v})^2} \quad (1)$$

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

3. 提案方式

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ため、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。ただし総エリア数を K とし、各エリアの推定値を CRH 法を利用して算出する。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を (3) 式で与える。また制約条件を (4) 式で与える。

$$\min E(u_1, u_2, \dots, u_K) = \sum_{i=1}^K (v_i - p_i)^2 \quad (3)$$

$$\sum_{i=1}^K u_i = N \quad (4)$$

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。

- 各エリアのサンプルワーカー数の初期値を $u_i = N/K$ 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
- 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
- ワーカー数調整前の総誤差を E_{pre} とする。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
- 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

- 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
- 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
- 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
- 総誤差の変化量 $|E_{post} - E_{pre}|$ が閾値 η を下回るまで、step 3 ~ step 7 を反復
- step 8 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

4. 性能評価

エリア数 K を 10、各エリアの測定値の平均 μ_i を 50 とし、各エリアの測定値の標準偏差を最小 2、最大 47 の 5 刻みで設定した。

図 1 に、提案方式を適用し総サンプルワーカー数を変化させた場合の、総推定誤差を示す。サンプルワーカー数調整前の総誤差である E_{ini} と調整終了後の E_{conv} を比較すると E_{conv} が全てのサンプルワーカー数の場合で E_{ini} を下回り、提案方式による総誤差抑制効果を確認できる。また、総サンプルワーカー数の増加に伴い単調に総誤差は減少することが確認できる。同サンプルワーカー数における総誤差の減少量 ($E_{ini} - E_{conv}$) に着目すると、サンプルワーカー数が多いほど減少量は小さくなる。

また、全体を通して、初期配置状態 S_0 と調整終了後 S_1 の各エリアのサンプルワーカー数の配置状態を比較すると、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た。

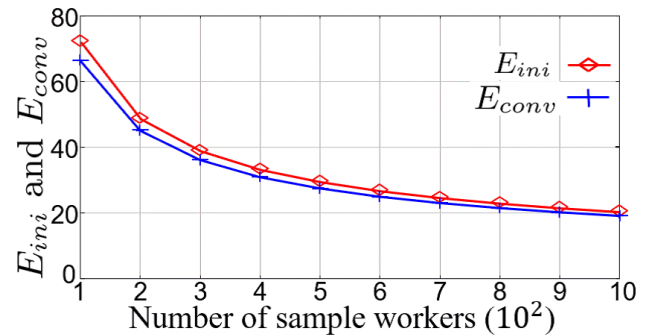


図 1: 総サンプルワーカー数を変化させたときの総誤差の推移

次に各エリアの設計サンプルワーカー数について考察する。例えば $N = 400$ とすると S_0 のとき $u_i = 40$ であるが、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、標準偏差が小さいエリアから順に、7, 20, 26, 33, 40, 45, 50, 55, 60, 64 となった。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。

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。

参考文献

- [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] R. Fujimoto and N. Kamiyama, Poisoning Attacks in Crowdsensing Over Multiple Areas, IEEE GLOBECOM 2022