Vol.2010-CSEC-50 No.49
2010/7/2

非同期秘匿分散 k-means クラスタリングの改良

青 木 良 樹[†1]　菊 池 浩 明[†1]

近年のクラウドコンピューティング技術や，携帯デバイス，ライフログの普及に伴い，多くの情報が電子化され管理されている．これらの情報をもとにして，利用者の嗜好に応じた商品を推薦する情報推薦サービスなども利用が著しい．しかし，これらのサービスはプライバシの保護という課題がある．佐久間らはユーザのプライバシを保護したまま，k-means を行う手法を提案している．本論文では，この佐久間らが提案した方式をもとに，非同期かつ平均を秘匿したままクラスタリングを行う方式を提案する．

Improvement of privacy in distributed asynchronous secure k-means clustering scheme

Yoshiki Aoki[†1] and Hiroaki Kikuchi[†1]

Information recommendation is one of the attractive applications in the age of ubiquitous computing where user's profiles and life-logs are digitally stored in cloud computing server. However, the privacy concern is current issue to study. Sakuma proposed a privacy-preserving clustering scheme according k-means algorithm without revealing privacy of hidden profile. Our paper studies the feasibility of privacy-preserving scheme due to Sakuma and reports some modified protocol with light-weight overhead.

1. はじめに

近年，クラウドコンピューティング技術や，ライフログの普及に伴い，多くの情報が電子化されて管理されてきている．これらのデータに基づいて，利用者の嗜好に応じた商品を推薦する情報推薦サービスなども利用が著しい．更に，特定の閉じたサービスの履歴だけではなく，インターネット利用者のウェブページの閲覧履歴をプロバイダ側で読み取り，ターゲットを絞った広告などに活用する DPI 技術 (Deep Packet Inspection) の検討も始まってきている[1)]．しかしながら，これらのサービスはプライバシの保護という課題がある．クラウドコンピューティング技術かプロバイダーで管理された利用者の嗜好や, 閲覧履歴が漏洩するリスクを考慮しなくてはならない．

この課題に対して，ユーザのプライバシを保護したまま，二者間でクラスタリングを行う研究が Vaidya らによって提案されている[2)]．秘匿内積評価プロトコルを用いて，各々局所的に試算したクラスタの重心を漏らさずに，正しくクラスタリングを実行する試みである．

一方，Kowalczyk らは，n 台のノードが分散する P2P ネットワークにおいて各ノードが持っている情報を中央に集中させることなく平均を計算するプロトコル “Newscast” を提案している[3)]．佐久間らはこの “newscast” を利用しプライベートな非同期平均計算プロトコル “Private Asynchronous Average Computation” を提案し[4)]，k-means アルゴリズムを，プライバシを保護したまま実行する手法を提案している[5)]．

PrivateAAC では，k-means クラスタリングを行う過程で，各クラスタの重心を秘匿して実行しているが，そこには大きな負荷がかかる．そこで，本研究ではその処理効率の向上を試みる．

本研究では，AAC に基づき，次の 2 点を改良する．(1) 各ユーザとクラスタの重心 (平均) を秘匿したまま効率良く計算するために，ユークリッド距離ではなくコサイン類似度を用いて計算を行い，各ノードのデータ，重心を秘匿したままもっとも類似度の高い重心を探索する手法について記述する．また，(2) 重心を公開，同期することなく各ノードがそれぞれクラスタリングを行う “思い込み非同期クラスタリング” について記述し，精度と効率を評価する．

2. 基本研究

2.1 k-means クラスタリング

k-means クラスタリングとは，各クラスタの平均を求め，k 個のクラスタにクラスタリングをしていく方式である．まず，ノード数を n，クラスタ数を k とし，ユーザ u_i のデータを $\boldsymbol{x}_i$，クラスタ j の重心を $\boldsymbol{\mu}_j$ とする．$\boldsymbol{x}, \boldsymbol{\mu}$ は d 次元のベクトルである．初期状態でユーザ u_i はランダムに k 個のクラスタに振り分けられる．そして，各クラスタの平均値 $\boldsymbol{\mu}_j$ を

†1 東海大学院工学研究科情報理工学専攻
Tokai University
259-1292 神奈川県平塚市北金目四丁目 1 番 1 号
ringo,kikn@cs.dm.u-tokai.ac.jp

計算する．各ユーザは自分の持っているデータ $\boldsymbol{x}_i$ と各クラスタの平均値 $\boldsymbol{\mu}_j$ との距離を計算し，最も近い (類似している) クラスタに所属を変更する．これを収束するまで回繰り返すことによってクラスタリングを行う．$n = 5$，$k = 2$ の時の初期状態と 2 世代目の状態を図 1 に示す．

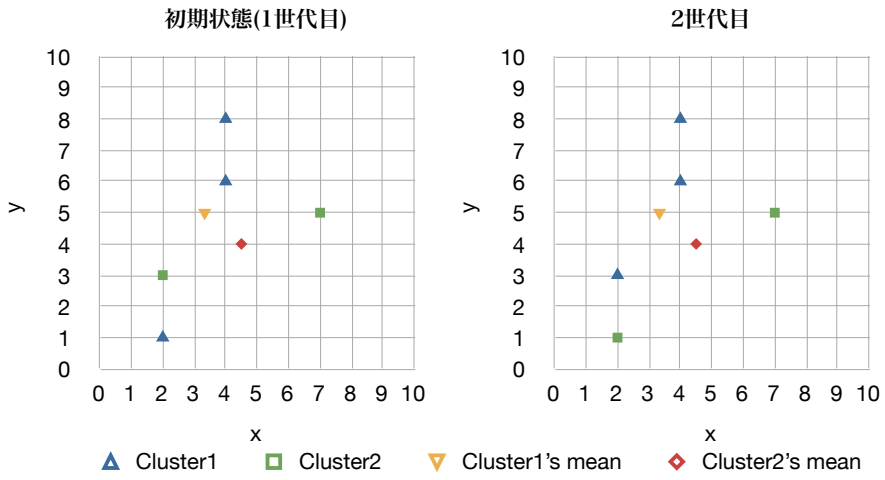

図 1 k-means クラスタリングの例

2.2 準同型暗号

準同型性を満たす暗号 (以下，準同型暗号) とは，平文や秘密鍵なしで $Enc[m]$ の m を計算できる暗号方式である．情報を秘匿したまま，加算 (乗算) することができる性質を加法 (情報) に関して準同型性を持つ，という[6]．加法準同型性を満たす暗号として，RSA 暗号や変形 ElGamal 暗号，Paillier 暗号などがある．本研究では加法に関して準同型性を持つ Paillier 暗号を用いた．

2.3 Newscast[3)]

Kowalczyk らによって提案された，非同期に平均を計算するプロトコルである．P2P ネットワークにおいて各ノードが持っている情報を中央に集中することなく平均値 μ を計算する．あるノード i がある P2P ネットワーク内のノード j をランダムに選び，$\mu_i^{(t+1)} = \mu_j^{(t+1)} = \frac{\mu_i^t + \mu_j^t}{2}$ と二者間で平均を取る作業を複数回繰り返す．十分な長さの t_* サイクル後には全ノードのベクトルは $\frac{1}{n}\sum_i^n \mu$ となり，全体の平均値に近似する．t_* を収束地のサイクルとする．

$$\mu_1^{t_*} = \cdots = \mu_n^{t_*} \doteq \frac{\sum_i^n x_i}{n} \tag{1}$$

2.4 Private AAC[4)]

佐久間らは，Newscast を応用して，秘匿したまま効率良く平均を求める非同期平均計算プロトコルを提案している．

準同型性を満たした暗号では，2 の割り算が実行できないので，次のようにして値を更新しる．サイクル t_i のユーザ i とサイクル t_j のユーザ j が，一般性失うことなく $t_i \geq t_j$ とすると，

$$Enc(\mu^{(t+j)}) = Enc(\mu_i^{(t)}) \cdot Enc(\mu_j^{(t)})^{2^{t_i - t_j}} \tag{2}$$
$$= Enc(\mu^t + 2^{t_i - t_j} \cdot \mu_j^t) \tag{3}$$
$$= Enc(\mu_j^{(t+1)}) \tag{4}$$

により，各々の暗号文を秘匿したまま総和していく．最後に，$\mu_i^{(}*) = \frac{\mu_i^{(}t)}{2^T}$ とすることで，newscast を等価な平均を得る．

3. 提 案 方 式

3.1 概　　要

k-means によってクラスタリングする際には，各世代において重心を全ノードで共有する必要がある．本手法ではこの重心を全ノードで共有せず，局所的に所持している情報のみで自分の所属するクラスタを判別する．

3.2 提案プロトコル

ノード数を n，クラスタ数を k，ノードの持つデータ x の次元数を d とする．各ノードを $u_1, \ldots, u_n\}$，サーバを S とする．各ノードは d 次元のデータベクトル $\boldsymbol{x}= (x_1, \ldots, x_d)$，各クラスタの平均の集合 $M = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ を持つ．ここで，$\boldsymbol{\mu}_i$ は i 番目のクラスタの重心ベクトルである．また，サイクル数 t と世代 g の定義を図 2 に示す．

Step1　各ノード u_i は自分の所持しているデータ $\boldsymbol{x}_i$ を Paillier 暗号を用いて暗号化し $Enc(x_1), \ldots, Enc(x_d)$ を取得し，自分の所属する j 番目のクラスタの重心の初期値とし，それ以外のクラスタの重心は $Enc(0)$ を要素とするベクトルとする．

Step2　privateAAC を行い，各ノードは暗号化された各クラスタの平均値 $Enc(\boldsymbol{\mu}_1), \cdots,$

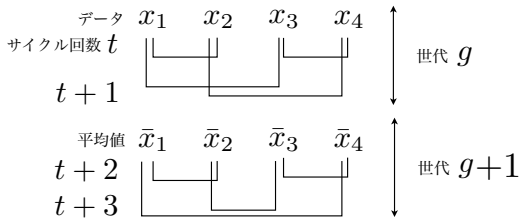

図 2 サイクル t と世代 g の定義

$Enc(\boldsymbol{\mu}_k)$ を取得する．[⋆1]

Step3 各ノード u_i は自分が所持しているデータ $\boldsymbol{x}_i$ と各クラスタの重心 $Enc(\boldsymbol{\mu}_1), \cdots, Enc(\boldsymbol{\mu}_k)$ との類似度 Sim を計算する．ノード u_i のクラスタ j への，類似度 $Sim_{i,j}$ の計算は次の式で定める．

$$Sim_{i,j} = \prod_{l}^{d} Enc(\mu_{j,l})^{x_{i,j}} = Enc(\sum \mu_{j,l} x_{j,l}) = Enc(\boldsymbol{\mu}_j \cdot \boldsymbol{x}_i) \quad (5)$$

Step4 ノードは k 個の類似度の中からランダムに $Sim_{i,A}, Sim_{i,B}$ の二つを選びとし，$(Sim_{i,A}/Sim_{i,B})^n$ を計算してサーバ S に送信．

Step5 サーバは送信された値を復号し，正の値か負値かを判別し，正の場合は 1, 負の場合は 0 の値をユーザに返す．

Step6 ノードはサーバから受け取った符号を見ることによって $Sim(A)$ と $Sim(B)$ のどちらが大きいかを判別することができる．ユーザは再び大きいと判定されたものを $Sim(A)$ とし，新しい $Sim(B)$ を選び，再びサーバに送信する．この Step4-6 を このとき，ノードは A と B のクラスタ番号を記憶しておく．トーナメント形式で，最も大きい一つが残るまで $k-1$ 回繰り返すことで，最も類似度の高いクラスタを探索する．

Step7 ノード u_i の所属を最も類似度の高いクラスタに変更する．

Step8 Step1-4 を T 回繰り返し，最終的に所属しているクラスタの番号を出力する．T は k-means を繰り返し実行する回数である．

⋆1 ただし，今回は privateAAC の最後の段階で行う 2^{T+1} の除算は行わない．

4. 評　　価

4.1 Newscast の性能

newscast プロトコルで算出した平均 μ^* は，真の平均 μ の近似値である．この二つの値の誤差 $\Delta\mu = |\mu - \mu^*|$ があるのかを調査した．newscast プロトコルは約 40 サイクル行うと収束するといわれている[3]．実際にノード数によってどのくらいで真の平均へ収束するのかを調査したグラフを図 3 に示す．

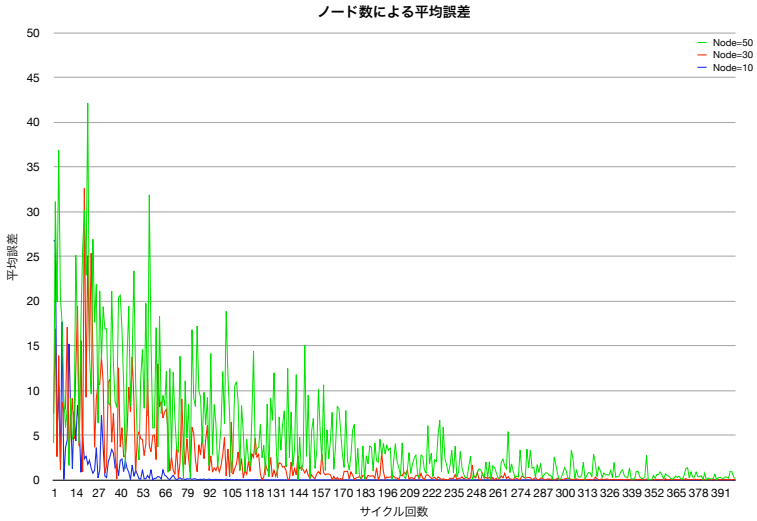

図 3 サイクル数による真の平均に対する誤差

newscast による平均値の分布をを図 5 に示す．この図は $n = 100$ で，データ x の値に 100 以下の整数を割り当てた時のものである．図に示す通り, サイクル数が高い方が平均値に収束していることがわかる．

newscast はランダムに相手ノードを選択するため，必ずしも自ノードと相手ノードのサ

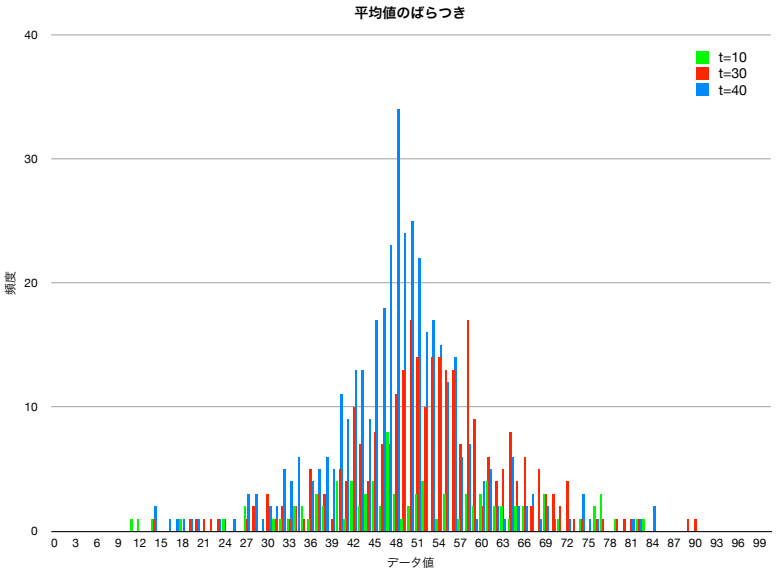

図 4 サイクル回数による平均値のばらつき

イクル回数が一致するとは言えない．ノード数 n で newscast を実行した時の二者間のサイクル数の誤差の分布を図 5 に示す．

4.2 コサイン類似度の精度

k-means クラスタリングを行う際に，各ノードと平均を計算する方法をユークリッド距離とコサイン類似度でそれぞれ計算した場合の精度を比較する．ユークリッド距離 $d(\boldsymbol{x},\boldsymbol{\mu})$ と，コサイン類似度 $\cos\theta$ はそれぞれ次の式で計算される．

$$d(\boldsymbol{x},\boldsymbol{\mu}) = \sqrt{\sum_{i=1}^{n}(x_i-\mu_i)^2} \tag{6}$$

$$c(\boldsymbol{x},\boldsymbol{\mu}) = \frac{\boldsymbol{x}\times\boldsymbol{\mu}}{||\boldsymbol{x}||\cdot||\boldsymbol{\mu}||} \tag{7}$$

精度 E_t を次のように定義する．

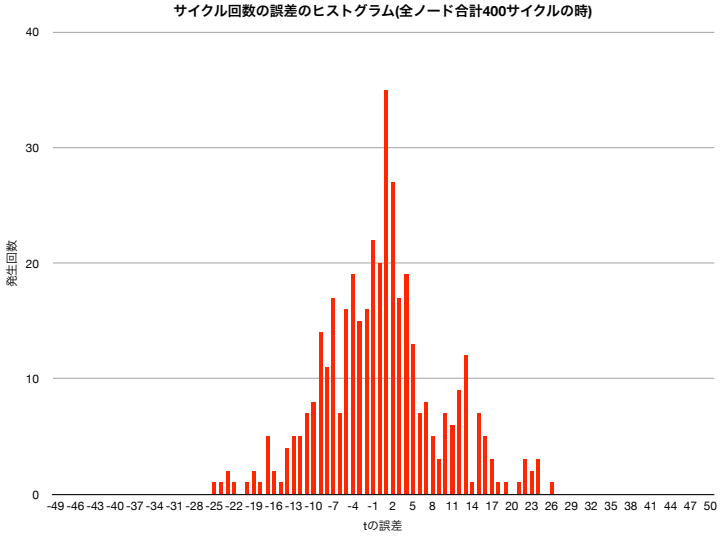

図 5 誤差の分布

$$E_t = \sum_{j}^{k} |G_{j*} - G_{jt}| \tag{8}$$

k-means クラスタリングにおいて，ユークリッド距離を用いて k-means クラスタリングを行った結果 G_{j*} を真値とし，コサイン類似度を用いて k-means クラスタリングを行った結果 $G_{j,g}$ を比較し，異なったクラスタへ識別されたノードの個数を誤差とする．ユークリッ

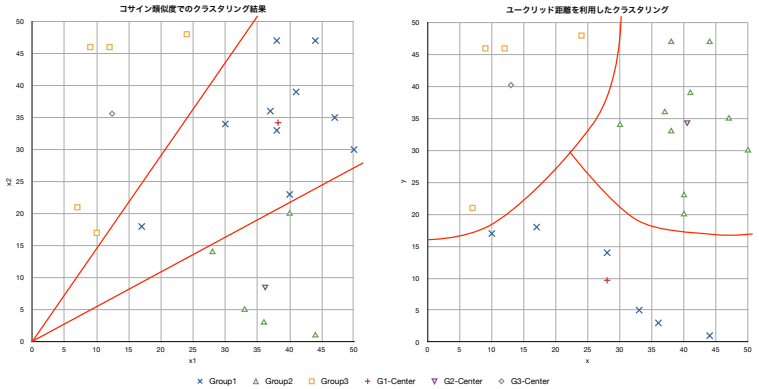

図 6 左：ユークリッド距離を用いた場合，右：コサイン類似度を用いた場合

5. Web 履歴に基づくユーザクラスタリング

相羽らによる検索履歴のプライバシーを秘匿したユーザクラスタリング[7] で用いたデータを利用し，k-means クラスタリングに適用した．このデータは被験者 6 名の検索履歴を一人 15 件，合計 90 件取得し，Yahoo!!Japan の分類をもとに検索履歴を分類分けしたものを利用する．分類分けは図 1 の通りである．

$d = 15, n = 6, k = 3$ クラスタリングを行う前と実行後の状態を表 3, 表??に示す．

表 1 各番号に対応するジャンル

番号	ジャンル名
1	エンターテイメント
2	メディアとニュース
3	趣味とスポーツ
4	ビジネスと経済
5	各種資料と情報源
6	生活と文化
7	芸術と人文
8	コンピュータとインターネット
9	健康と医学
10	教育
11	政治
12	自然か科学と技術

表 2 k-means を実行する前のデータとクラスタ

User	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	C1	C2	C3
A	20	0	38	3	4	13	2	7	2	2	1	2	0	1	5	0	1	0
B	8	2	50	0	7	4	0	5	10	1	0	2	3	0	8	1	0	0
C	30	6	21	13	7	1	0	20	0	1	0	0	1	0	0	1	0	0
D	0	1	2	16	23	4	38	7	8	0	0	0	0	1	0	0	1	0
E	9	19	26	7	13	1	0	4	0	0	14	3	0	4	0	0	0	1
F	8	7	18	0	20	1	6	10	11	0	4	3	2	0	10	0	0	1

表 3 k-means を実行後のデータとクラスタ

User	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	C1	C2	C3
A	20	0	38	3	4	13	2	7	2	2	1	2	0	1	5	1	0	0
B	8	2	50	0	7	4	0	5	10	1	0	2	3	0	8	1	0	0
C	30	6	21	13	7	1	0	20	0	1	0	0	1	0	0	1	0	0
D	0	1	2	16	23	4	38	7	8	0	0	0	0	1	0	0	1	0
E	9	19	26	7	13	1	0	4	0	0	14	3	0	4	0	0	0	1
F	8	7	18	0	20	1	6	10	11	0	4	3	2	0	10	0	0	1

Vol.2010-CSEC-50 No.49
2010/7/2

参 考 文 献

1) 総務省, “クラウドコンピューティング時代のデータセンター活性化策に関する検討会”, 2010 年 5 月 28.
2) G. Jagannathan, K. Pillaipakkamnatt, R. N. Wright, D. Umano,“Communication-Efficient Privacy-Preserving Clustering”, Transaction on Data Privacy, pp. 1-25, vol. 3, 2010.
3) W. Kowalczyk and N. Vlassis, “Newscast EM”, Advances in Neural Information Processing Systems 17, MIT Press, 2005.
4) 佐久間 淳, 小林 重信, “P2P ネットワークにおけるプライバシを保護した非同期平均計算プロトコル”, pp. 1-6, SCIS2007 3D4-1.
5) 佐久間 淳, 小林 重信, “P2P ネットワークにおけるプライバシを保護した k-means クラスタリング”, pp. 1-6, SCIS2007 3D4-2.
6) 木澤 寛厚, “プライバシ協調フィルタリングにおける利用者の評価行列の次元削減”, pp. 509-514, コンピュータセキュリティシンポジウム 2008 論文集.
7) 相羽 研次, “検索履歴のプライバシーを秘匿した ユーザクラスタリング”, 東海大学情報理工学部情報メディア学科 2009 年度卒業論文, 2009.

謝　　辞

本研究にあたり，実データを提供していただいた相羽研次氏に感謝を申し上げます．また，実験に協力してくださった方々に感謝の意を述べると共に，謝辞とさせていただきます．