5 匿名加工再識別コンテストの加工アルゴリズム— PWS Cup 2017 優勝チームより—

濱田浩気 | NTT セキュアプラットフォーム研究所／理化学研究所

本稿では，匿名加工再識別コンテスト PWS Cup 2017 で優勝した筆者の参加チーム「君の名は〜ユアネーム〜」が使用した加工アルゴリズムを紹介する．なお，ページ数の制約のため，ルールを実際よりも単純化している．

コンテストのルール

PWS Cup 2017 では，図 -1 のような顧客の購買履歴データを対象に，個人の特定が起こりにくいように，かつ，元のデータとの差異が大きくならないように加工する技術をチーム対抗で競った．コンテストは大きく**加工フェイズ**と**再識別フェイズ**により構成され，最後に各チームが加工したデータが評価されて勝者が決定する．

加工フェイズ

加工フェイズでは，各チームは全チームに共通の公開の購買履歴データ（図 -1 に例を示す．**元データ**と呼ぶ）を受け取り，元データを加工したデータ（図 -2 に例を示す．**加工データ**と呼ぶ）を作成する．加工データの作成は以下の 3 種類の加工とその組合せにより行われる：(1) **行の削除**，(2) **仮名化**（たとえば，Alice を X4 に置き換える．ただし同一の名前はすべて同じ仮名に変更しなくてはならない），(3) **セル値変更**（たとえば，「桃」を「梨」に変更する）．

名前	購入日	商品
Alice	9/7	梨
Bob	9/2	梨
Bob	9/10	ブドウ
Carol	9/2	桃
Carol	9/12	ブドウ
Dan	9/2	ブドウ
Dan	9/5	梨
Dan	9/12	ブドウ

図 -1　元データ

仮名	購入日	商品
X4	9/2	梨
X2	9/2	梨
X3	9/2	梨
X1	9/2	梨

図 -2　加工データ

再識別フェイズ

再識別フェイズでは，各チームは他の全チームの**公開加工データ**（加工データから削除された行を除去してランダムに行の順序を入れ替えたもの．図 -2 に対応した公開加工データの例を図 -3 に示す）を受け取り，名前と仮名の対応を推測する．

加工データの評価

各加工データは**安全性**と**有用性**の 2 つの観点から評価され，2 つの評価値の和である**総合点**が最小のチームが勝者となる．

安全性は，正しく再識別された顧客の割合に基づいて評価される．元データの顧客数を n，正しく再識別された顧客数を m とするとき，m/n を**再識別率**と呼ぶ．たとえば，図 -3 に対して Alice=X1, Bob=X2, Carol=X3, Dan=X4 と推定した場合の再識別率は 2/4=0.5 となる．安全性の評価値は，その加工データに対して行われたすべての再識別の再識別率の最大値（小さいほど良い）により定義される．

有用性は，元データと加工データの遠さに基づいて評価される．有用性の評価値は，削除された行の割合と値を変更されたセルの割合の和（小さいほ

仮名	購入日	商品
X1	9/2	梨
X2	9/2	梨
X3	9/2	梨
X4	9/2	梨

図 -3　公開加工データ

ど良い）により定義される．たとえば，図 -2 は削除された行数が 4，値を変更されたセル数が 3 であるので，有用性の評価値は 4/8+3/16=0.6875 となる．

加工アルゴリズム

以下では筆者らが PWS Cup 2017 で用いた加工アルゴリズムを紹介する．筆者らはまず，総合点を容易に見積もることができる基本アルゴリズムを作成した．その後，見積もられる総合点が小さくなるように基本アルゴリズムを改良した．

基本アルゴリズム：全員を同じにする

図 -2 のような加工を考える．加工後の 4 名の顧客は，加工データ上は順序と仮名以外に何も違いがない．したがって，仮名が名前と無関係にランダムに作られている場合，公開加工データ上でどれが Alice であったかは完全に分からない．すなわち，Alice が正しく再識別される確率は 1/4 となる．今，すべての人が見かけ上同じになっているため，同様に他の顧客も正しく再識別される確率は 1/4 であり，正しく再識別される人数の期待値は 1/4+1/4+1/4+1/4=1 である．これは正しく再識別される人数の期待値としては理論上最良である．安全性の評価値の見積もりとして再識別率の期待値を用いることにすると，総合点は 1/4+0.6875=0.9375 と見積もられる．

改良 1：バランスを取る

図 -2 の加工は理想的な安全性を実現したが，総合点は 0.9375 と大きい（悪い）値であった．総合点をより小さくするため，図 -2 とよく似た図 -4 のような加工を考える．図 -2 では 4 人全員が同じになるように加工したが，図 -4 では Alice と Bob，Carol と Dan がそれぞれ同じに見えるように加工をしている．この場合，仮に Alice が Y2 か Y3 のいずれかであることが突き止められたとしても，Y2 と Y3 は図 -2 のときと同様に仮名以外は完全に一致しているので，Alice が正しく再識別される確率は 1/2 となる．同様に Bob, Carol, Dan も正しく再識別される確率は 1/2 となる．よって，図 -4 は仮に絞り込みが成功したとしても，正しく再識別される人数の期待値は 1/2+1/2+1/2+1/2=2 となる．図 -2 の場合に比べると安全性の評価値は 2/4=0.5 と少し悪くなったが，削除された行数は 2，変更されたセル数は 2 で有用性の評価値は 2/8+2/16=0.375 であり，総合点は 0.5+0.375=0.875 と改善できた．

仮名	購入日	商品
Y2	9/2	梨
Y3	9/2	梨
Y4	9/2	ブドウ
Y4	9/12	ブドウ
Y1	9/2	ブドウ
Y1	9/12	ブドウ

図 -4　加工データ（改良 1）

仮名	購入日	商品
Z1	9/7	梨
Z3	9/2	梨
Z3	9/12	ブドウ
Z4	9/2	梨
Z4	9/12	ブドウ
Z2	9/2	梨
Z2	9/12	ブドウ

図 -5　加工データ（改良 2）

改良 2：より良いグループ分けの選択

さらに良い総合点を得ることはできないだろうか．図 -4 の加工を振り返ると，同じ安全性の評価値を達成するには，必ずしも（Alice, Bob）と（Carol, Dan）のグループ分けをする必要がないことが分かる．たとえば，(Alice, Carol)，(Bob, Dan) でも，(Alice)，(Bob, Carol, Dan) でも正しく再識別される人数の期待値は同じ 2 となる．実際，(Alice)，(Bob, Carol, Dan) とグループ分けをした図 -5 では削除された行数は 1，変更されたセル数は 3 で有用性の評価値は 1/8+3/16=0.3125 となり，総合点は 2/4+0.3125=0.8125 と見積もられる．このように，分け方が決まると総合点を見積もることができることを利用し，グループ分けの選択を組合せ最適化の問題と捉えて求解することにより，総合点を小さくすることができる．

以上のようなアルゴリズムを用いることによって筆者らのチームは総合点が小さい加工データを作成することができ，結果として好成績を得ることができた．

（2018 年 1 月 31 日受付）

■濱田浩気（正会員）　hamada.koki@lab.ntt.co.jp

2009 年京都大学大学院情報学研究科通信情報システム専攻修士課程修了．現在，日本電信電話（株）NTT セキュアプラットフォーム研究所研究主任，国立研究開発法人理化学研究所革新知能統合研究センター客員研究員．