1Q-3 大規模クラスタリング技術検証†

仲光 廣晃　　野口 喜洋

松下電器産業株式会社　マルチメディアシステム研究所‡

1. 背景

近年、インターネットやデジタル放送ビジネスに関連して、ユーザへの広告配信を前提としたビジネスモデルが多く採用されている。そのようなモデルでは、ユーザがより興味をもった、言い換えればそれにアクセスする確率が少しでも高い広告を配信することが、収益確保上の要件となる。そのため、各ユーザの広告に対する操作履歴を保存・解析して、各ユーザの嗜好に適合した広告を選択するという、いわゆるレコメンデーションの手法が使われる。しかし、これらのビジネス領域では、ユーザ数が100万人単位と大規模であること、そして、アクティブなユーザの集合が時間とともに入れ替わり、全体が把握しにくいことから、1人1人のユーザに対する既存レコメンデーション手法の適用が必ずしもうまくいかないという問題点がある。

そこで、操作履歴や嗜好の近いユーザ同士をクラスタリングし、そのクラスタを単位としてレコメンデーションを適用する手法が有効である。本稿では、100万人単位のユーザに対して適用でき、かつ高速にクラスタ生成が行える手法を開発したので報告する。

2. 大規模データ数に適したクラスタリング手法

一般的にこれまでに提案されているクラスタリング手法は、大きく3つに分類することができる。1つは、階層型クラスタリング、もう1つは、非階層型クラスタリング、最後の1つは、このいずれにも属さないものである。

まず、最初に階層型クラスタリングについて検証する。最短距離法や、最長距離法に代表される階層型クラスタリングは、データ間の距離(一般的に、ユークリッド距離)から、対象を分割、または統合してクラスタを形成する。この時、この距離の計算に多大なメモリ領域が必要とされる。

例えば、データ数(ユーザ数)が1万だった場合には、距離を1万総当りで計算する必要が生じる。この時、この1万総当りの距離データを保存するのに、データ形式をfloat型(4byte)で持った場合でも、

10000×10000×4(byte)=400000000(byte)=400M(byte)

となり、実質400M(byte)以上のメモリ領域が必要となる。実際用いられるデータ数は1万以上であると想定され、このことから階層型クラスタリングは、大規模データには、不向きであると判断される。

次に、非階層型クラスタリング手法について検証する。非階層型クラスタリング手法は、クラスターを代表とする点をクラスターの数だけ用意し、分類対象データは、その代表点の中から最も類似した点(一般的に、最もユークリッド距離の近い点)に分類される。ここで、非階層型クラスタリングは、適切なクラスタ数が既知にわかっていなければならない、という問題点がある。今回のようなデータ(ユーザ)がいくつのクラスタに分類できるか既知でない場合、非階層型クラスタリングは用いることができない。また、クラスタ数が既知でない場合でも用いることができるFCM*[1]アルゴリズムは、クラスターの統合に距離を用いるため、階層型アルゴリズムと同様、多大なメモリ領域を必要とし、適切であるとは言えない。

3. SOM[1]

階層型、非階層型クラスタリングのどちらも適切でなく、その他のクラスタリング手法を検証した結果、SOM[1]によるクラスタリングが、メモリ使用量、及び速度の面から見ても望ましいという結論にいたった。

図 1に示すようにSOMは、ニューロンと呼ばれる2次元格子上に配置された多数のニューロンで構成される。ニューロン1個には、クラスタリング対象の次元と同じベクトル値(参照ベクトル)が格納されている。SOMは、クラスタリングを行う際に、総当りの距離データを必要とせず、各ニューロンの参照ベクトルの値のみをメモリ上に保持すればよい。結果として、大幅な使用メモリ減少を実現できる。また、SOMの学習アルゴリズムなどは、参考文献を参照されたい。

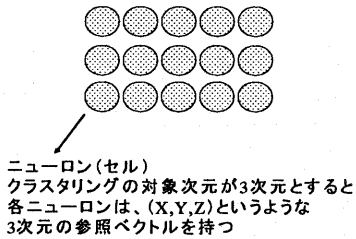

図 1　SOM構造

4. SOMからのクラスタ形成方法

SOMからのクラスタの形成方法は、大きく分けて2つある。1つは、各ニューロンを代表点とみなす方法。もう1つは、学習済みのSOM[2]を解析する方法である。

まず、各ニューロンをクラスターの代表点とする方法だが、この手法を用いるとニューロン数がおのずとクラスタ数となってしまう。今回のような、クラスタ数が既知でない場合のクラスタリングには、この手法を用いることができない。

よって今回は、学習済みのSOMを解析することによりクラスタ形成を行う。この形成するにあたりニューロンから代表点となりうるニューロン(便宜的に今後密ニューロンと呼ぶ)の検出[3]を用いる。この密なニューロ

---

†Studies on a Large Scale Clustering

Hiroaki Nakamitsu,Yoshihiro Noguchi

‡Matsushita Electric Industrial Co., Ltd.
Multimedia Systems Research Laborator.
E-mail: nakamitu@trl.mei.co.jp

[1] Self Oraganizing Maps

ンとは、隣接ニューロンとの参照ベクトルが非常に近似しているニューロンを表す。

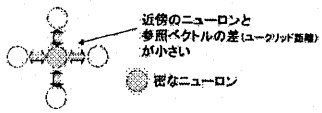

図 2 密なニューロンと近傍のニューロンの関係

隣接しているニューロンと参照ベクトルが近似しているということは、そのニューロンが何度も勝ちニューロンとなり、隣接するニューロンの参照ベクトルを更新し、互いに似た参照ベクトルを持ったと考えられ、学習データ密集個所の代表点といえる。

具体的には、すべてのニューロンで、4近傍のニューロンとの参照ベクトルの差を累積する。次に、求められた参照ベクトル差が、4近傍のどの参照ベクトル差よりも小さい場合、密なニューロンとし、グループの代表点とするものである。

5. 密なニューロン検出の改良案

データの大規模化に伴い、以下において蜜ニューロの検出方法に2つの改良案を提案する。

(i) 辺、角のニューロンにおける、参照ベクトル差修正

辺、及び角のニューロンは、上下左右いずれかのニューロンが存在しないため、求められた参照ベクトル差が、通常4近傍から得られるものよりも低くなる。さらに同様の理由で、密なニューロン検出時、比較対象のニューロンが通常の4近傍より少なくなるため、選出されやすいという現象が起きる。これは、データの次元が2次元でも起こるが、より高次元データの場合、各ニューロン間の参照ベクトル差は、次元数に比例するため、この傾向はより顕著になる。

この現象を防ぐため、辺、角のニューロンの持つ参照ベクトル差を以下のように修正する。

辺：存在しない1つのニューロンとの参照ベクトル差として、存在する3つのニューロンとの参照ベクトル差の平均を用いる。

角：存在しない2つのニューロンとの参照ベクトル差として、存在する2つのニューロンとの参照ベクトル差を用いる。

(ii) 8近傍との比較

密ニューロの検出に、4近傍のニューロンとの参照ベクトル差を比較するが、この方法だと、密ニューロンが斜めに生じるという現象が起こる。これは、SOMが学習の際に、勝ちニューロンから一定半径内のニューロンの参照ベクトルを変化させるので、斜めのニューロン同士も似通った参照ベクトルを持つようになり、結果として両者とも密ニューロンとして検出されることとなる。これは、検出する必要ないニューロンを密ニューロンとして選択してしまう。

この現象を防ぐため、従来密ニューロンとして4近傍と参照ベクトルを比較していたものを、8近傍と比較することで、斜め同士のニューロンが密ニューロンとして検出されるのを防ぐ。

6. 実験結果

今回、視覚的にわかるよう2次元上に任意数の中心点からガウスノイズをかけたデータを用いSOMを学習させ、そこからのクラスタ形成に、従来手法と今回の改良案を用いたものとの比較実験を行った。以下にその結果を示す。(この時、ニューロンの数は、100個、初期学習係数:0.1、初期学習半径:3で学習した。)

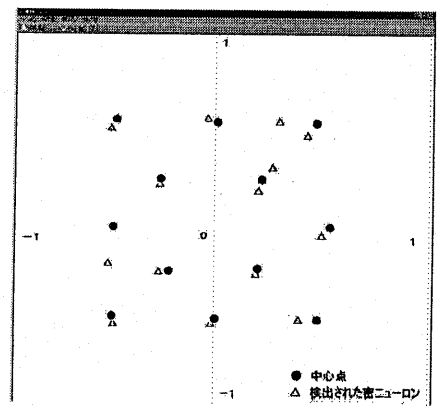

図 3 従来手法による実験結果

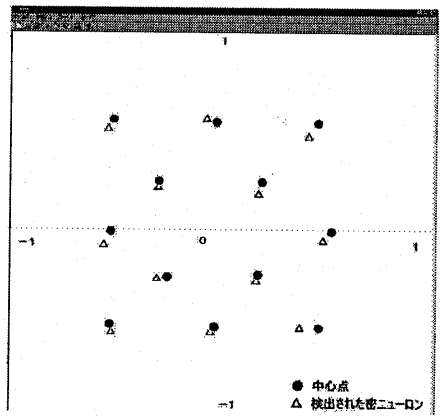

図 4 改良案による実験結果

7. おわりに

本稿の実験から、ＳＯＭによるクラスタリングが１００万人単位のユーザに対して適用でき、かつ、考案した密ニューロンによる代表点検出の手法が有効であることが判明した。今後は、高次元（１００次元以上）のデータに対する本手法の適用について検討する。

なお、本研究は、通信・放送機構(TAO)からの「情報家電インターネット」に関する委託研究に基づいて推進しているものである。

参考文献

[1]乾誠貴,亀井且有,井上和夫,「クラスター推定機能付き Fuzzy C-Means アルゴリズム」,電気学会論文誌,Vol.114-C,No.11,pp.1166-1171,1994

[2]T.Kohonen,"Self-OraganizingMaps",Springer Verlag, Berlin,1995.

[3]田中雅博,白谷文行,山本公明,「自己組織化マップ上のデータ密度ヒストグラムを用いたクラスタ分類法」,電子情報通信学会論文誌,D-II,Vol.J79-D II,No.7,PP.1280-1290,1996