5M-4　次元の段階的なしぼり込みによる射影クラスタリング

黒木 薫　古瀬 一隆*　大保 信夫*
筑波大学理工学研究科 筑波大学電子・情報工学系*

1 はじめに

クラスタリングは、顧客分類、クラス分類、傾向分析などの問題としてデータベースの分野では良く知られた問題である。しかし、既存の手法を高次元データ(属性数が多いデータ)に適用するとクラスタの精度が悪化するという問題が起こる。これは、既存の手法では、全次元の情報を用いて(例えばユークリッド距離などで)クラスタリングを行うため、高次元データにおける特徴のある次元が、特徴の無い次元によって隠されてしまうという理由による。

これらの問題の解決策として projected clustering[1] と呼ばれる手法が提案された。この手法は、クラスタ毎に重要な次元を見つけ出しクラスタリングを行う。本研究では、この projected clustering のアルゴリズム PROCLUS[1] を改良し、より精度の高いクラスタリング方式を提案する。

2 特徴のある次元とその数が異なるクラスタ

2.1 PROCLUS

高次元データでは、クラスタ毎に特徴のある次元及びその数が異なる。したがって、クラスタ毎に特徴のある次元を見付けだし、その次元における部分空間でクラスタリングを行う手法が提案された。このような手法のことを射影クラスタリングと呼ぶ。

射影クラスタリングのアルゴリズム PROCLUS は、k-medoid 法と山登り法に基づいている。k 個の medoid を選んだ後、各 medoid の周辺にあるデータをサンプリングして、medoid との平均距離などをもとに、特徴のある次元を見つけ出す。データをクラスタリングする際には、データと部分空間において最も近い medoid へ分配していく。発見されたクラスタに対して、クラスタ評価関数をもとに、山登り法を用いて medoid の交換を行い最適な medoid の組合せおよび、特徴のある次元を選び出す。

A projected clustering algorithm using stepwise dimension determination
Kaoru Kuroki, Kazutaka Furuse*, Nobuo Ohbo*
Facluty of Science and Engineering Tsukuba University, Institute of Information Sciences and Electronics Tsukuba University*

2.2 PROCLUS の問題点

PROCLUS で特徴のある次元を見つけ出す際のサンプリングでは、一番近い種との「全次元距離」を半径とした超球 (locality) 内のデータを用いるため、locality に入るべきデータ以外のデータが入ってしまい、特徴のある次元を判定する精度が悪くなると言った問題がある。

入力データのクラスタ (A、B、C、D、E) と PROCLUS における locality(L1、L2、L3、L4、L5) 内のデータとの関係の一例を示す。

(データサイズ 10000、次元数 20、クラスタ数 5、平均特徴次元数 6、クラスタの特徴次元の分散値を 8 のパラメータをもつデータ)

	A	B	C	D	E	OUT
L1	1127	73	47	131	125	60
L2	83	2441	1065	398	45	135
L3	96	266	1144	271	40	70
L4	248	37	221	1266	78	104
L5	309	46	33	211	957	81

表 1: 入力クラスタの locality への分配

この図から、各 locality に様々な入力クラスタのデータが含まれてしまっていることがわかる。これにより、次元の決定の際に精度が悪化してしまうという現象が起こる。

3 提案手法

3.1 提案手法

PROCLUS の問題点を考慮して、改良を加えた。まず、次元を段快適にしぼり込みながら、データを分配し直すことで、それぞれの medoid に特徴をもったデータが集まりやすくなるようにした。

また、medoid を最初に多めにとり、段階的に減らして行くという方法を用いた。medoid のしぼり込みでは、各クラスタを他のクラスタのに分配してみて、その評価値を基に最適なしぼり込みを行う。

3.2 提案手法のアルゴリズム

ここでは、PROCLUS を改良した提案手法を紹介する。提案する手法は、次元のしぼり込み、クラスタ数のしぼり込み、再構築の 3 段階からなる。提案手法のパラメタ及び、アルゴリズムを以下に示す。

＜ パラメタ ＞

- クラスタ数 (medoid の数) : k
- クラスタの次元の数の平均値 : l
- 初期 メドイド数 : $\alpha * k$
- 次元をしぼり込む数 : l_{new}(初期設定はデータ空間の次元)
- 次元を減らす数 : l_{del}
- 最終しぼり込み次元 : l'

＜ アルゴリズム ＞

1. 次元のしぼり込み
   (a) データから種を任意に $\alpha * k$ 個選ぶ。
   (b) データを部分空間で最も近い種に分配し、$\alpha * k$ 個のクラスタをつくる。
   (c) 各クラスタにおいて、全次元から特徴のある次元を l_{new} 個選ぶ。(特徴のある次元の判定は、PROCLUS と同様のものを使う)
   (d) $l_{new} \leftarrow l_{new} - l_{del}$
   (e) 次元の数が l' になるまで (b),(c),(d) を繰り返す。
   (f) $l' * k$ 個の次元を平均 l 個になるように各クラスタに分配する。
2. クラスタ数のしぼり込み
   (a) 一つのクラスタを、その他のクラスタに再分配する。(それぞれの種と、部分空間での距離が最も近いクラスタにデータを分配する)
   (b) できたクラスタを評価し、評価値が最も高くなるような、クラスタの再分配を行う。(クラスタの評価は PROCLUS と同様のものを使う)
   (c) (a),(b) をクラスタ数が k 個になるまで繰り返す。
3. 再構築
   前段階で生成されたクラスタに対し、次元の再決定およびクラスタの再構成を行う。

3.3 PROCLUS との比較実験

データサイズ 10000、次元数 20、クラスタ数 5、クラスタの平均次元数 6、クラスタの次元における分散値を 4,6,8,10,12 のパラメータをもつデータを 5 セット用い、実験結果はその 5 セットに対する評価値の平均値とした。(提案手法において $\alpha = 3$、$l' = 10$、$l_{del} = 1$ とした)

PROCLUS によって発見されたクラスタの評価値と、提案手法によって発見されたクラスタの評価値と、元データのクラスタに対する評価値を比較した。

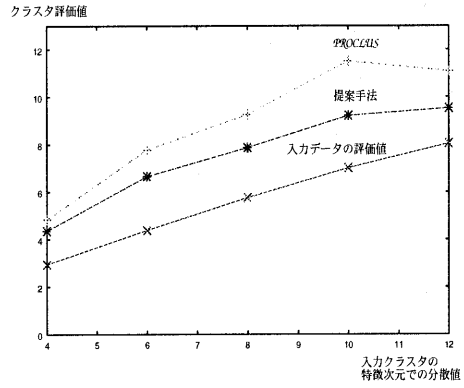

図 1: PROCLUS との比較実験

この図より、PROCLUS によるクラスタリングの結果よりも、提案手法の方が理想値に近付いていることがわかる。

4 考察

今回提案した手法によるクラスタリングは PROCLUS によるものよりも平均的に精度が上がることが分かった。この理由として、次元をしぼり込みながらデータの分配を行うことによる、特徴のある次元の判定の精度向上が推測される。また、クラスタの再分布によるクラスタ数のしぼり込みにおいて、特徴のある次元の判定がうまく行っている medoid だけが残るため、さらに精度があがったと推測される。

5 今後の課題

今後の課題として、以下のことを検討中である。

- 実データ (画像の特徴抽出データなど) を用いた実験。
- ORCLUS[2] との比較実験。

参考文献

[1] Charu C.Aggarwal, et al., Fast Algorithms for Projected Clustering. SIGMOD '99

[2] Charu C .Aggarwal,Philip S. Yu., Finding Generalized Projected Clusters in High Dimensional Space. SIGMOD 2000