ISS-P-031

決定木アルゴリズムを用いたクラスタリング

原田瑞季† 三浦 孝夫†

† 法政大学理工学部創生科学科

1. はじめに

本研究では，情報量を用いた泡沫クラスタを回避するクラスタリング手法を提案する。

2. 提案手法

本章では，決定木を用いた新たなクラスタリングを提案する。この特徴はエントロピーに基づいて，明確な特徴を抽出し，十分なサイズを有するトップダウン型クラスタを作成する。特徴がない，または小規模な場合は泡沫クラスタとみなして捨て去る。まず，属性ごとにエントロピーを算出し，エントロピーが最小の属性を分割基準として選択する。これは，エントロピーが小さいものがデータを特徴づけていると考えられるためである。

そして，情報利得が最大となる属性で分割を行い，集合の要素数，もしくはエントロピーが閾値以下で分割終了し，更に分割後の要素数が閾値以下の集合は切り捨てる。閾値を設定する理由は，サイズやエントロピーが小さいものは泡末クラスタを作成している可能性が高いためである。

3. 実験

3-1.実験手順

閾値は，エントロピーの平均が小さいものがまとまりのある集合が多いとして，エントロピーの平均が最小の閾値を選択する。エントロピー=0.2 もしくは 0.3，要素数=5%の時がエントロピーの平均が最小であることから，今実験ではエントロピー=0.3，要素数=5%を閾値として採用する。

評価方法は，実験によって得られるエントロピクラスタ数とクラスタリングの手法である k-means のエルボー法によって得られる最適クラスタ数と一致するかを比較する。

また，k-means で得られるクラスタは，表面上の見た目の近さのみでクラスタを形成するため意味を正しく捉えることは容易ではない。そのため，エントロピクラスタを反映させ，ホテル情報からクラスタの特徴が実際に成り立つか信頼性を検証する。

3-2. 実験データ

本研究では，宿泊施設の口コミサイトである“楽天トラベル”を利用する。NII 情報学研究データレポジトリから取得した 2015 年 9 月 28 日，29 日のデータ 2000 件のうち，データの欠陥を除き，1 つのホテルに対して 2 件以上のレビューを持つ全 509 件の中から「その他」という項目を除いた 480 件を使用する。

属性は，「目的」，「同伴者」，「評価 1(立地)」，「評価 2(部屋)」，「評価 3(食事)」，「評価 4(風呂)」，「評価 5(サービス)」，「評価 6(施設・アメニティ)」，「評価 7(総合評価)」，「広さ」，「アクセス」，「価格」の 12 個を用い，要素が数値であるものは非数値に変換し，数値が広域的であれば，場合分けをして非数値に変換する。

3-3.実験結果

エントロピクラスタは，11個存在する。集合の詳細を表 1 に示す。集合の要素数に着目すると，要素数が閾値以下の集合を切り捨てたことから一定数の要素を持つ集合のみが存在し，泡沫クラスタを回避できていることが確認できる。また，エントロピーが 0 に近い値のクラスタを形成していることが確認できる。

表 1 集合の詳細

集合	20m^2未満の要素数	20m^2以上の要素数	集合内の要素数	エントロピー
1	32	1	33	0.1959
2	33	0	33	0
3	27	0	27	0
4	20	6	26	0.7794
5	32	2	34	0.3228
6	15	11	26	0.9828
7	25	2	27	0.3809
8	29	0	29	0
9	15	20	35	0.9852
10	10	17	27	0.9510
11	19	6	25	0.7950

次に，k-means のエルボー法により得られた最適クラスタ数4つとなり，エントロピクラスタ数とは不一致である。

K=11 の k-means のクラスタとエントロピクラスタで共通要素を持つ比率より，共通要素を含む比率が 10%以上あり，クラスタの要素数の過半数を超える共通要素が存在するエントロピクラスタは際立った特徴を持ち，クラスタをエントロピクラスタで表現できることから，クラスタにホテル情報から特徴づけを行ったものを表 2 示す。

表 2 エントロピクラスタで特徴づけされたクラスタ

クラスタ	エントロピクラスタ	特徴と比率		
C2	E2+E3	定員1人 98.3%(59/60)	朝食あり 78.3%(47/60)	洋室 100%(60/60)
C6	E9	宴会場あり 74.3%(26/35)	スウィートルームあり 68.6%(24/35)	温泉あり 65.7%(23/35)
C8	E5	定員1人 100%(34/34)	朝食なし 79.4 %(27/34)	洋室 70.6%(24/34)
C9	E4+E5	定員1人 96.6%(86/89)	朝食なし 75.3%(67/89)	洋室 80.9%(72/89)
C10	E1+E2	定員1人 100%(66/66)	朝食あり 75.8%(50/66)	洋室 94.0%(62/66)

すべての特徴は信頼性を 6.5 割以上持つ。

4.結論

情報量を用いたクラスタリングによって，エントロピーが 0 に近い値を持つクラスタを作成することができ，閾値の設定によって泡沫クラスタを回避することができた。しかし，最適クラスタ数は 4 個となりクラスタ数は一致しなかった。

また， k-means クラスタのうち，エントロピクラスタを反映させることで特徴的なものを抽出でき，すべての特徴において 6.5 割以上の信頼性を持つことから，意味構成を理解することが可能になった。

(参考文献)高村大也：言語処理のための機械学習入門，コロナ社，2010