

居酒屋の注文内容の分析における種々のデータマイニング手法の有効性比較実験

谷 研究室 平 雅志

masashi taira

概要

居酒屋の各グループが注文したメニューから,Clementine を用いてメニュー同士の相性を調べ比較実験をする

1 はじめに

スーパーマーケットやコンビニの店舗やオンラインショッピングなどでは、顧客は通常ある商品と一緒に別の商品を買うことがよくある。そこで、POS データなどから顧客がある商品を買ったときに、どんな商品を同時に買うかという規則性を分析して、その規則性にもとづいて関連する商品の売場を近づけたり、関連商品を推薦したりして売上を伸ばしている。居酒屋でも POS データをとっているが実際はそのデータをうまく活用できていない。メニューの相性を調べるため、本研究では居酒屋の伝票からデータを取りさまざまな手法からデータ分析を行う。

2 データマイニングについて

データマイニングとは、大量のデータを分析することでその中に隠された規則性や法則性を見つけ出すこと。『マイニング (mining)』とは、鉱山から鉱石を見つけ出すことであり、大量のデータを鉱山に、その中にある役に立つ知識を鉱石に例えて、データマイニングとしている。コンピュータの普及によって顧客の購買履歴データや POS データなどがそのままデジタルデータとしてデータベースに蓄積されるようになった。こうしたデータの特徴としてはデータの件数が膨大で、そのまま眺めていたのでは、そこから有効な知見を得ることは難しい。そうした限界にとらわれず、データの中から新たな知見を得て、ビジネスに有効活用するための手段として、データマイニングが注目されるようになっていく。データマイニングの手法としては、相関分析や多変量解析などの統計解析、決定木 (ディビジョンツリー)、ニューラルネットワーク、遺伝アルゴリズムなどの手法が使われる。

3 マーケット・バスケット分析

データマイニングの利用法の 1 つで、POS データや購買履歴データなど分析して、“一緒に買われる商品”の組

み合わせを発見する探索的データ分析のこと。一般にアソシエーション分析の手法が利用される。

3.1 相関係数

相関係数は 2 変数間の直線的な関係 (線形関係) の度を量的に表す。通常 r という記号で書かれ、その変域は、 $-1 \leq r \leq 1$ である。1 に近いときは 2 つのデータ列には正の相関があるといい、-1 に近ければ負の相関があるという。0 に近いときはもとのデータ列の相関は弱い。

< 定義 >

$$r = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - x)(Y_i - y)}{\sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - x)^2} \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - y)^2}}$$

3.2 アソシエーション分析

< 定義 > アソシエーションルール

$$A \Rightarrow B (A \text{ ならば } B \text{ である})$$

次にアソシエーションルールに付随する、サポート (支持度) とコンフィデンス (確信度) という概念を定義する。サポートはアイテム集合 A と B をともに含む割合である。コンフィデンスは A において B を含む割合である。

< 定義 > アソシエーションルールのサポートとコンフィデンス

$$\text{サポート} (A \Rightarrow B) \equiv P(A \cup B)$$

$$\text{コンフィデンス} (A \Rightarrow B) \equiv P(B|A)$$

例えばアソシエーションルールは以下のように表せる。

(例) ビール $\Rightarrow$ 枝豆 [サポート = 3%, コンフィデンス = 50%]

アソシエーション分析は SPSS 社のデータマイニングツール clementine を使用する。

3.3 飽和系列パターンของアルゴリズム

相関分析において膨大な数の互いに類似した頻出パターンが出力される事がある. こういった問題を飽和系列パターンのアルゴリズムを用いて, パターンの発見を効率化できる. 栗原が作ったプログラムを使用する

4 実験・結果

本研究で使用するデータは埼玉にある居酒屋から提供していただいた伝票を Excel で集計した.

組数………300 円

客数………855 人

売上合計…2,376,854 円

4.1 相関係数の分析結果

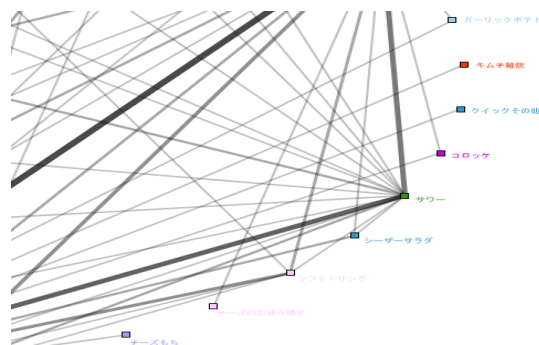
(クレメンタインを使用)

商品名		ビール	サワー	カクテル
ビール	Pearson の相関係数 有意確率 (両側) N	1 300 300	-.005 .929 300	.096 .098 300
サワー	Pearson の相関係数 有意確率 (両側) N	-.005 .929 300	1 300 300	.020 .726 300
カクテル	Pearson の相関係数 有意確率 (両側) N	.096 .098 300	.020 .726 300	1 300 300
枝豆	Pearson の相関係数 有意確率 (両側) N	.260** .000 300	.040 .487 300	.257** .000 300
たこわさ	Pearson の相関係数 有意確率 (両側) N	.122** .035 300	.058 .314 300	.216** .000 300
キムチ	Pearson の相関係数 有意確率 (両側) N	.261** .000 300	-.071 .222 300	-.073 .206 300

正の相関	負の相関
ビール&キムチ	カクテル&キムチ
ビール&串焼き	焼酎&サワー
ビール&唐揚げ	
ビール&大根サラダ	
カクテル&デザート	
カクテル&シーザーサラダ	
ししゃも&焼酎	
なすの浅漬け&びんちょう	

4.2 アソシエーション分結果

(クレメンタインを使用)



結果	前提条件	サポート %	確信度 %
デザート	ナスの浅漬け and なんこつあげ	4.33	100
デザート	カクテル and ナスの浅漬け and なんこつあげ	4	100
デザート	ビール and ナスの浅漬け and なんこつあげ	3.33	100
デザート	カクテル and 豆腐サラダ and 豚しそ	3	100
デザート	サワー and ナスの浅漬け and なんこつあげ	3	100
デザート	ビール and カクテル and ナスの浅漬け and なんこつあげ	3	100
デザート	ビール and サワー and ナスの浅漬け and なんこつあげ	3	100
サワー	ビール and ナスの浅漬け and ガーリックポテト	3.67	100
デザート	サワー and たこわさ and ナスの浅漬け	4	91.67
ナスの浅漬け	果実酒 and お茶漬 and デザート	3	88.89
たこわさ	カクテル and 焼酎 and 枝豆	3	88.89

4.3 飽和系列パターンのアルゴリズムの分析結果

```
[s5402013@bump gr]$ gcc -o howa-test howa-test.c
[s5402013@bump gr]$ ./howa-test home-1.txt
ノンアルコール
冷製
冷製 冷製
冷製 炒め物
冷製 チーズ類
冷製 ごはん物
冷製 カクテル
炒め物
炒め物 ごはん物
サラダ
串焼き
串焼き 冷製
チーズ類
揚げ物
ごはん物
デザート
焼酎
ビール
ビール 冷製
ビール 炒め物
ビール サラダ
ビール 串焼き
ビール チーズ類
ビール 揚げ物
ビール ごはん物
ビール サワー
ビール カクテル
サワー
カクテル
カクテル 冷製
カクテル カクテル
```

5 考察

相関係数による分析	
2 変数間では有効	データの件数が増えると有効な組み合わせを見つけるのが困難
アソシエーション分析	
結果が明解 探索的搜索法	指数時間である あまり購入されないアイテムは出力されない
飽和系列パターンによる分析	
高速である	あまり購入されないアイテムは出力されない 適切な頻出値を見つけるのが困難

6 参考文献

参考文献

- [1] SPSS クレメンタインによるデータマイニング. 牛田 一雄 高井勉 小暮大輔 著
- [2] 次世代データベースとデータマイニング. 石川博 著
- [3] clementine 8.0 user guide. SPSS 社