レシートデータによる同時購入商品の選定

2005MM018 猪子真奈　2005MM072 鈴木あゆ美
指導教員：田中 豊

1　はじめに

近年, ホームセンターには様々な商品が揃えてある. 何か目的を持って買い物に行ったが, 予定外の物を購入してしまったという, いわゆる「ついで買い」の経験は誰にでもあるのではないだろうか. レジ付近での商品の陳列はこれが狙いともいえる. また, 棚に陳列されている商品とは別にフックにぶら下げられている形で陳列されている商品を目にしたこともあると思う. あるホームセンターでは, このフックにかけられている商品陳列の仕方を「クリップフック陳列」と呼んでいる. 現在は, バイヤー個人の感覚や経験で同時購入されやすいと思われる商品を, 店側が工夫してクリップフック陳列している.

POS(Point Of Sales:販売時点情報管理) データが獲得しやすくなった昨今, 重要な課題の 1 つは蓄積されたデータを有効に活用することである.

本研究ではあるホームセンターのデータを頂き, どのような商品が同時購入されやすい傾向にあるのかを分析する. そしてその分析結果から, どのような商品をクリップフック商品として陳列したら効果的であるかを目的として解析を行う.

2　データについて

データは, ある一つの店舗のレシートデータを 1 か月単位で用いる. 商品の種類は 30 万種類以上存在するが, それらの商品は小分類というグループに分けられて管理されている. ホームセンターの全商品は 336 種類の小分類に分けられている. この小分類単位でのデータを用いる.

3　データ加工

同時購入の情報を抽出するために 1 点しか購入されていないレシートデータを除き, 2 点以上の商品が購入されているレシートデータを使用した. 行には小分類名, 列にはレシート番号が入っており購入されていれば 1, そうでなければ 0 が入っているトランザクションデータと, 小分類ごとの購入の組み合わせの度数を数えたクロス集計表を使用した. トランザクションデータについては後に詳しく示す.

4　分析方法

クロス集計表から多次元尺度法とクラスター分析を, トランザクションデータからアソシエーション分析を行った.

3 つの手法により分析をし, 同時購入されやすい商品の傾向を調べクリップフックの候補を探す.

5　多次元尺度法

5.1　多次元尺度法

多次元尺度法は, 個体間の類似度データを,2 次元あるいは 3 次元の空間に配置する方法である. 多次元尺度法は計量多次元尺度法（計量 MDS）と非計量多次元尺度法（非計量 MDS）に大別される. 計量 MDS では非類似度データあるいは距離データを低次元に配置し, 非計量 MDS では, 観測値に基づく非類似度データそのものを利用するのではなく, 適切な非線形変換（単調変換）を提案し, それによって変換した値を, 距離とみなして低次元空間に配置する手法である.[1],[2].

5.2　計量多次元尺度法による分析

12 月と 2 月のレシートデータから得られた, 小分類単位のクロス集計表を計量 MDS を用いて分析した. クロス集計表は, 各小分類同士の同時購入回数を集計したもので, 値が大きいほど類似していることを表す類似度データである.

計量 MDS は非類似度データあるいは距離データを用いるため, クロス集計表の数値を非類似度に変換する必要がある. 変換する方法として, クロス集計表の最大値以上の定数からそれぞれの数値を引く方法と, 逆数をとる方法（このとき値に 0 が含まれていないように考慮する必要がある. 本研究では全ての値に 1 を足している.）の 2 つの方法を用いた. 本研究では前者を方法 1, 後者を方法 2 とする. なお, 方法 1 の最大値以上の定数は 12 月 5650, 2 月 3210 としている. この 2 つの方法を用いて非類似度データに変換し, 分析を進めた.

5.2.1　方法 1 で変換した非類似度データでの分析

方法 1 で変換した 12 月, 2 月のデータから得られた計量 MDS の結果は元のデータとの適合度 GOF（主成分分析の累積寄与率に相当）が 12 月が 0.10724, 2 月が 0.07886 であり後述のストレス値がそれぞれ 0.89 と 0.88 であった.

5.2.2　方法 2 で変換した非類似度データでの分析

方法 2 で変換した 12 月, 2 月のデータから得られた計量 MDS の結果は, 元のデータとの適合度 GOF が 12 月が 0.08307, 2 月が 0.08857 であり後述のストレス値がそれぞれ 0.41 と 0.37 であった.

5.3　計量多次元尺度法による分析の考察

第 5.2.1 目と第 5.2.2 目の結果から, 計量 MDS による分析では元のデータとの当てはまりがよくないことがわかった. また, ストレス値に関してみると, 特に方法 1 の当てはまりの悪さが目立った.

5.4　非計量多次元尺度法による分析

非計量 MDS は観測値から得られた非類似度データを非線形変換（単調変換）した値を, 低次元に配置する手法である. 第 5.2 項と同様に, 集計結果から得られるクロス集計表を方法 1 または 2 の方法で非類似度データに変換し, 分析を行った.

非計量 MDS では個体間の距離 d_{ij} を単調変換した値 $\theta(d_{ij})$ と k 次元に配置された空間の距離 $\hat{d}_{ij}$ との差が小さくなるような座標値と, 単調変換の関数 θ を求める. 詳しくは次に示すストレス（stress）値（S）と呼ばれる統計

量を最小にするように座標値を決める.（計量 MDS の場合は, $\theta(d_{ij}) = d_{ij}$ である.）

$$S = \frac{\sum_i \sum_j [\theta(d_{ij}) - \hat{d}_{ij}]^2}{\sum_i \sum_j \hat{d}_{ij}} \quad (1)$$

ストレス値が小さい程, 得られた座標値がもとのデータをよく再現していることを示す. ストレス値による当てはまりの評価基準を表 1 に示す.

表 1　ストレスによる当てはまりの評価基準

S	当てはまり度
0.2	よくない
0.1	悪くはない
0.05	よい
0.025	非常によい
0.00	完璧

また, このストレス値を最小にする計算の過程で, ストレスの初期値（initial value）は計量 MDS の結果のストレス値が用いられている. この initial value がどれだけ減少するかで, 計量 MDS からの当てはまりの改善具合を判断することができる.

5.5　非計量多次元尺度法の分析結果

方法 1 を用いた計量 MDS の結果を初期値とした非計量 MDS の分析では, ストレス値が初期値 0.89 から改善できず, 収束したものと判定された. 方法 2 を用いた計量 MDS の結果を初期値とした非計量 MDS の分析では, ストレス値の改善がみられた. 次元数を 2〜5 として分析したときの 2 月のストレス値を表 2 に示す.

表 2　次元によるストレス値の変化

次元数	ストレス値
2	0.35
3	0.27
4	0.23
5	0.20

5.6　非計量多次元尺度法による分析の考察

第 5.5 項での結果より, 方法 1 で変換した非類似度データでの非計量 MDS の分析は計量 MDS から改善されていない. デーとの当てはまりが非常に悪く, ストレスの局所最小値になっている可能性がある. 方法 2 の結果より次元数が増えるにつれて, ストレス値が小さくなっていくことがわかる. 理論的には非計量 MDS は入力された非類似度の単調変換に対して不変であるので, 方法 2 の結果を採用し, 次元としては 4 次元での結果を考察した. その理由として, 次元数を増やせば増やすほどストレス値は小さくなるため, ストレス値の下がり幅を考慮したことによる. 第 5.2.1 目の結果で得られた座標値のプロット図は, {紙, 住居洗剤, 洗濯洗剤} と {キャットフード, スナック, ドックフード,　犬・猫用品} の 2 グループが目立っていた. この目立ったグループに着目して方法 2 の結果を考察した.4 次元での結果のプロット図で各小分類の位置を確認したところ, 小分類同士が近い位置に配置されていた. つまり 2 組は同時購入されやすい組の小分類とみてよいと考えられる. また, 方法 2 の計量 MDS の結果との違いは次元数やそれぞれのプロット図をみても明らかであるが, ストレス値に関して言えば, 非計量 MDS での分析のほうが, よりデータの再現ができていることがわかった.

6　クラスター分析

6.1　階層的クラスター分析による分析

階層的クラスター分析は結果として樹形図（デンドログラム）が得られる方法で, とくにクラスター数は定めず, 対象の階層的構造を求め, 目的に応じて大まかに分類したり細かく分類したりすることが可能である. 本研究での階層的クラスター分析の目的は, とくに類似度の高い小分類の組み合わせを見つけることである. データは, 第 5 節で用いた非類似度データに変換した 12 月と 2 月のクロス集計表（変換方法は方法 2 を採用）を用いた. 階層的クラスター分析の手法は最長距離法を採用している.[1],[2].

6.2　階層的クラスター分析による分析結果

いくつかのクラスターに分けてみたところ, どちらの月も, クラスター数が少ない場合（クラスター数 10 とする）, ある 1 つのクラスターに 200 種類以上の小分類が集中し, 残りのクラスターには小分類が 1〜4 種類という結果であった. また, クラスター数が多い場合（クラスター数 80 とする）, クラスター数が少ない場合と比べ 1 つのクラスターに集中する小分類が 100 種類ほどに減ったものの, 残りのクラスターには小分類が 1 つだけというクラスターがほとんどという結果が得られた.

6.3　階層的クラスター分析による分析の考察

第 6.2 項より, 他と同時購入される頻度のあまり高くない多くの小分類がそれぞれ単体でクラスターを形成し, 同時購入される頻度の高い小分類が 1 つのクラスターを形成することがわかった. つまり, クラスター数が多くても離れない同じクラスター内にある小分類は, 同時購入されやすいということを意味するので, デンドログラムの低いところに位置する小分類が意味のある組み合わせとなる可能性があり, デンドログラムを距離の小さいところで切断すれば, 意味のある組み合わせを見つけられることがわかった. これを踏まえて, 同時購入されやすい小分類の組み合わせを見つけるため, いくつかの小分類を抽出して調べた.

6.3.1　小分類の抽出と多次元尺度法との比較

意味のある小分類の組み合わせを見つけるため, いくつかの小分類を抽出し, それらがデンドログラムのどの位置にあるか調べた. 多次元尺度法との比較も同時に行った. 小分類は第 5.6 項で取り上げた, キャットフード, ドックフード, スナック, 犬・猫用品, 住居洗剤, 紙, 洗濯洗剤を 2 月のクラスター分析の結果から抽出した.

その結果, 抽出した小分類はデンドログラムの非常に低い位置に見つけられた. そして, 第 5.6 項で組となった小分類が同様に同じクラスターを形成することがわかった. 多次元尺度法との比較により, ピックアップした小分類 7 種類は, {住居洗剤, 紙, 洗濯洗剤} の組と {キャットフード, ドックフード, スナック, 犬・猫用品} の組として, 同

時購入されやすい小分類であることがより確かになった.

7 アソシエーション分析による分析結果

7.1 導入

POS データの入力方式の一つとして表 3 のようなデータをトランザクション (取引) データと呼ぶ. アソシエーション分析は (相関ルール学習法 association rule learner ともよばれる), 小売店や店舗などで集めている表 3 のようなトランザクションデータを活用するために, 商品間の関連性について分析することを目的として 1990 年代初めに IBM 研究所によって開発された手法である. アソシエーション分析は表 3 のようなトランザクションデータから頻出する商品の組み合わせの規則を漏れなく抽出し, その中から価値ある規則を探し出すことを主な目的としている. [1],[3].

表 3 より, トランザクションデータの各行は 1 枚のレシートに対応し, 商品を購入していれば 1 が入っている. 例えばレシート ID=”23749” のデータでは洗濯洗剤と洗面・浴用品を購入しているので, それらに 1 が入りその他には 0 が入っている.

表 3 トランザクションデータ

ID	洗濯洗剤	洗濯用品	洗面・浴用品	装粧品
23742	0	0	0	0
23746	1	0	0	0
23748	0	0	0	0
23749	1	0	1	0

7.2 相関ルール

トランザクションデータに頻出する商品間の何らかの組み合わせの規則を相関ルールと呼ぶ. 「商品 X を買うと商品 Y も買う」のような相関ルールを X⇒Y の形式で表す. 相関ルールの「⇒」の左辺を条件部, 右辺を結論部と呼ぶ.[1],[3].

7.3 相関ルールの評価指標

相関ルールを検出する際, 用いられてる指標としては support 値 (支持度), confidence 値 (確信度), lift 値 (リフト) がある. データの中の, 商品集合 X を含むトランザクション件数を $\sigma(X)$, 全トランザクション件数を M を用いて表す. 例えば $X = \{$ 紙 $\}$ を含むトランザクションが 3 つあれば $\sigma(X) = 3$ である.

support 値は, 相関ルールが全データの中でどの程度出現するかを表す割合であり, 下記の式で表される.

$$\mathrm{supp}(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{M} \tag{2}$$

σ (X ∪ Y) は商品集合 X と Y の両方を含むトランザクションの件数を表し, 全トランザクションの中で X と Y の両方を購入する相対度数 (確率) を表す. support 値が高い相関ルールは, レシートデータにおける出現頻度が高く, 売り上げに大きく影響する重要な相関ルールであるといえる.

confidence 値は X(条件部) が起こった時に Y(結論部) が起こる割合であり, 次式で表される.

$$\mathrm{conf}(X \Rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{3}$$

X を購入した中で X と Y の両方を購入する相対度数 (確率). confidence 値の高いルールは,X と Y の結びつきが強く信頼できるルールといえるが, そのXがほとんど出現しない場合は, あまりその商品を買う人がいない商品の組み合わせを意味するルールとなり, confidence 値だけ高くても意味がない. すなわち, ルールの重要性を判断するうえで, confidence 値が高いことは必要であるが, support 値もある一定度の高さが必要であるといえる.

lift 値は confidence 値を supp(Y) で割った値で定義されている. 確率論の記号を用いて, $\tilde{X}$ と $\tilde{Y}$ は商品集合 X,Y を購入するという事象, 従って $\tilde{X} \cap \tilde{Y}$ に X と Y を同時購入する事象を表すことにすると,lift 値は, トランザクション数が大きくなると確率 $\Pr(\tilde{X} \cap \tilde{Y})/\Pr(\tilde{X})\Pr(\tilde{Y})$ に近づく. X を購入する確率が Y と無関係 (独立) なら 1 に等しく,lift > 1 なら一方を購入すれば他方も購入しやすいという正の関連性. 逆に lift < 1 なら一方を購入すれば他方を購入しにくいという負の関連性があることを表している.

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{conf}(X \Rightarrow Y)}{\mathrm{supp(Y)}} \tag{4}$$

lift 値が小さいと Y の導入効果は小さいと判断される.

以上のことから, ルールの評価は support 値, confidence 値, lift 値を総合的に考慮する必要がある.[1],[3].

7.4 小分類の分析結果

一店舗の 12 月, 2 月, 3 月, 4 月のデータで解析を行った. データ数が膨大なため, まずは小分類単位で分析を行った. ある一店舗のデータに対して,support 値が 0.003 以上,confidence 値が 0.1 以上に該当するデータを検出するようにして, 解析を行った. 小分類単位で分析を行った後, 同時購入の頻度が高くクリップフックの候補となり得る小分類を単品にしてさらに詳細な分析を行う.

7.4.1 12 月小分類の相関ルール

support 値 0.003 は,12 月全体の件数 41292 件の 0.3% 約 123 件に相当する. 表 4 は検出されたルールの結果の一部抜粋で, support 値の上位 5 位である. 2 月,3 月,4 月のデータも同様に表にまとめた. 今回は紙面の都合上,3 月,4 月の表は割愛する.

表 4 12 月小分類の相関ルール (supp≥0.003,conf≥0.01) 一部抜粋

X		Y	supp 値	conf 値	conf 値	lift 値
			X⇒Y, Y⇒X	X⇒Y	Y⇒X	X⇒Y, Y⇒X
紙	⇒	洗濯洗剤	0.042	0.318	0.307	2.326
住居洗剤	⇒	洗濯洗剤	0.040	0.349	0.292	2.552
住居洗剤	⇒	紙	0.028	0.244	0.212	1.846
ボディケ	⇒	洗濯洗剤	0.025	0.398	0.183	2.907
ラップ	⇒	紙	0.023	0.405	0.173	3.063
…	⇒	…	…	…	…	…

7.4.2　2 月小分類の相関ルール

support 値 0.003 は,2 月の全体の件数 32336 件の 0.3% 約 97 件に相当する.

表 5　2 月小分類の相関ルール (supp≥0.003,conf≥0.01) 一部抜粋

X		Y	supp 値	conf 値	conf 値	lift 値
			X⇒Y, Y⇒X	X⇒Y	Y⇒X	X⇒Y, Y⇒X
紙	⇒	洗濯洗剤	0.0327	0.2763	0.2616	2.2113
ドックフ	⇒	スナック	0.0271	0.4572	0.3566	6.0239
スナック	⇒	犬・猫用	0.0266	0.3500	0.3310	4.3618
住居洗剤	⇒	洗濯洗剤	0.0239	0.3921	0.1910	3.1374
キャット	⇒	犬・猫用	0.0231	0.4202	0.2882	5.2364
…	⇒	…	…	…	…	…

7.5　考察

消耗品は購入数が多いため, 消耗品を含む相関ルールの support 値はどの月も高くなっている. また,support 値が低くても lift 値が高いものは全体でみると買われている数は少ないが商品同士の関連性が高いことを示す.
12 月は, 消耗品の中でも洗剤など掃除関係のものが上位になり,2 月は 12 月に比べるとペット関係のものが上位に入ってきた. 3 月は 2 月とあまり変わらないようであったが 4 月は用土や花苗といった園芸関係のものが上位になった.

7.6　単品を含む分析結果

小分類同士の相関ルールをみてきたが, クリップフックされる商品を明確にするために小分類と単品の相関ルールを調べることにする. 小分類から単品にしたものは, ねじ・釘, 住居洗剤, 洗濯洗剤, 筆記用具, 事務用品である. 単品にするために選んだ小分類は 12 月,2 月,3 月,4 月の結果を踏まえ,X⇒Y と Y⇒X の confidence 値の差が 2 倍以上,support 値が 0.003 以上,lift 値が 2.0 以上という 3 つの条件を満たすものとした. X⇒Y と Y⇒X の confidence 値は X と Y のどちらの商品がついで買いされやすいかを示している. そのため confidence 値の差が 2 倍以上ある相関ルールのうち,confidence 値が大きい相関ルールの Y に該当する小分類はクリップフック商品の候補になり得ると考えられる. また,support 値が 0.003 以上というのは各月の全体客数の 0.3% に相当し,lift 値が 2.0 以上ということは関連性があるという事を意味する. これら 3 つの値を考慮し, 単品にする小分類を決定した. 3 月のレシートデータは約 11 万件で, 単品の商品名と小分類の種類も 3724 種類という大きなデータになったため, レシートデータを 1 万件ずつランダムに抽出し, トランザクションデータに加工したのち分析を行った.
表 6 と表 7 は, それぞれ 1 万件のレシートデータをトランザクションデータに加工しアソシエーション分析をしたものである. support 値 0.0003 以上,confidence0.01 以上で, それぞれ Y が単品商品になっている相関ルールのみを表にした. 表 6 は一部抜粋, 表 7 は support 値上位 5 位である.

表　6　　単品を含む 3 月の相関ルール 1 (supp≥0.0003,conf≥0.01) 一部抜粋

X		Y	supp 値	conf 値	lift 値
スナック	⇒	エリエールトイ	0.0021	0.0210	4.2945
犬・猫用品	⇒	エルモア　ティ	0.0021	0.0232	2.4354
スナック	⇒	エルモア　ティ	0.0021	0.0210	2.2053
…	⇒	…	…	…	…
加工材	⇒	ステン釘	0.0008	0.0429	41.6357
…	⇒	…	…	…	…

表　7　　単品を含む 3 月の相関ルール 2 (supp≥0.0003,conf≥0.01) 一部抜粋

X		Y	supp 値	conf 値	lift 値
ボディケア	⇒	エルモア　ティ	0.0044	0.1128	3.8841
台所洗剤	⇒	エルモア　ティ	0.0033	0.1068	3.6779
ボディケア	⇒	ハミングレギュ	0.0027	0.0677	6.9207
ラップ・ホイ	⇒	快適上手ＥＣＯ	0.0027	0.0957	5.7703
芳香・消臭	⇒	エルモア　ティ	0.0027	0.1304	4.4920
…	⇒	…	…	…	…

7.7　考察

表 6 より, 加工材とステン釘のように似た状況で使用されるような商品は lift 値が高く同時に購入されやすいことが確認できた. またティッシュなど普段から必要な消耗品は X の小分類の種類に関係なく, 同時購入されやすいことがわかった.

8　まとめ

多次元尺度法とクラスター分析から, {住居洗剤, 紙, 洗濯洗剤} , {キャットフード, ドックフード, スナック, 犬・猫用品} の組が同時購入されやすい小分類として見つけられたが, 意外性のある小分類同士の組を見つけることにはつながらなかった. しかし, 小分類同士の位置関係は確認でき, クリップフック商品として検討する目安になることが結果として得られた.
アソシエーション分析から,confidence 値や lift 値を考慮することで, ねじ・釘や紙などどの小分類がクリップフック商品に向いているかわかった. 小分類で各月みてみると, 多少ではあるが季節性が見られることができた.12 月の掃除関係の消耗品や 4 月の花苗などが季節性を表していると考えられる. 単品を含む分析では, トランザクションデータの作成にかなり時間がかかり, また 1ヶ月分のレシートデータをトランザクションデータにまとめることがRでは不可能であったため単品を含んだ相関ルールが 3 月分の一部しかできなかったが, 今後大規模データを扱えるような方法を検討し, より確実なクリップフック商品の選定につなげていきたいと考えている.

参考文献

[1] 金明哲：『R によるデータサイエンス』. 森北出版株式会社,2007.

[2] 田中豊, 脇本和昌：『多変量解析』. 現代数学社,1983.

[3] 元田浩, 津本周作, 山口高平, 沼尾正行：『データマイニングの基礎』. オーム社,2006.