

POSデータを用いたエンド陳列実施の判別

前田, 瑛 / Maeda, Yo

(出版者 / Publisher)

法政大学大学院理工学・工学研究科

(雑誌名 / Journal or Publication Title)

法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編

(巻 / Volume)

58

(開始ページ / Start Page)

1

(終了ページ / End Page)

4

(発行年 / Year)

2017-03-31

(URL)

<https://doi.org/10.15002/00014366>

POSデータを用いたエンド陳列実施の判別

A STUDY ON DISCRIMINANT METHODS OF DISPLAY-SETTING FOR RETAIL PRODUCTS ON SHELVES IN SUPERMARKET BASED ON POS DATA

前田 瑛

Yo MAEDA

指導教員 木村 光宏

法政大学大学院理工学研究科システム理工学専攻修士課程

It is of great importance for sales promotion management to optimize how the retail products should be displayed in a store. Among the various techniques of the sales promotion, we especially focus on the technique of displaying the products at the end of shelves in the store. It is widely believed that the efficient usage of the end of shelves can yield more chances of selling the products in the supermarket. In order to ensure this fact, we need to obtain the related data sets. However, it is difficult to collect/record the data sets with low cost which explain the relation between the performance of sales and whether the products were displayed at the end of shelves in the actual situation. In this research, we discuss the methodology to estimate whether the several specified products were actually displayed at the end of the shelves from other information obtained from the point-of-sales (POS) data sets. We apply several discriminant analysis models to the real data sets and discuss their performances.

Key Words : Discriminant analysis, POS data, SVM, logistic regression

1. はじめに

(1) 研究背景

小売業などの企業は、消費者からの購買反応を引き出すためのセールスプロモーション（以下 SP）と呼ばれる施策を実施している。SP は金額 SP、広告 SP、店頭 SP の 3 つに分けられ、特に店頭 SP は短期の売上に直結し、日本においてはその傾向が強い [1]。したがって、店頭 SP の効果を測定する市場反応モデルの研究が多くなされてきた [2]。本研究で取り扱うエンドも店頭 SP の一つである。エンドとは陳列棚の端のスペースであり、来店客の多くがエンドの前を通過するためエンドに置かれた商品は手に取られやすい（図 1）。従って、エンドに陳列される商品は値引率を抑えても売上が上がる傾向が見られ、エンドを考慮することにより精度の高い市場反応モデルの構築が期待できる。しかし実際には、店頭 SP の研究成果にはエンドを考察の対象とした研究は少ない。その主な理由として考えられるのは、いつエンドに陳列されているかが POS データには記録されず、エンドの陳列実施の有無を記録したデータがほとんどないためである。また、日本においてスーパーマーケットは 2 万店以上展開しており、調査員による観測も困難である。そのため、エンドの陳列実施の有無を他から得られる情報から推測することは、より精度の高い市場反応モデルの研究の一助となることから有用であると考えられる。

(2) 研究目的

本研究はエンドの陳列データを補うために、POS データの動向からエンドの陳列有無を判別することを目的とする。類

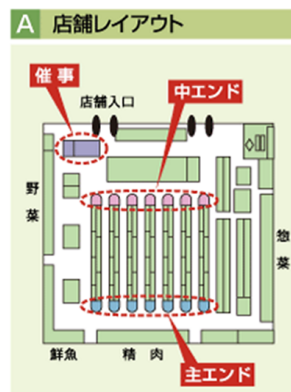


図 1 店舗レイアウト（[2] より転載）。

似研究として、佐藤・樋口 [3] は店頭 SP の一つである山積み陳列実施を考慮したモデルを提案しており、この中で山積み陳列実施の有無の判別を行っている。本研究では、傾向の異なる商品の POS データに対して分析を行う。判別には 4 つのモデルを用い比較・考察する。

2. データセット

(1) データセットの概要

本章では本研究で扱うデータについて説明する。本研究では POS データと呼ばれる多くの小売業で取得可能な販売データを用いる。POS データには「いつ」、「どこで」、「何が」、

「いくらで」、「何個」売れたかが記録されている。本研究で用いる POS データは 2013 年 9 月から 2014 年 9 月までに収集されたデータで、A～D の 4 店舗のスーパーマーケットから収集された 53 週間の週次データである。POS データにはポテトチップス（以下ポテチ）、柿の種、チョコパイの 3 商品について週、売上金額、販売数量、平均価格、来客数の 5 項目が記録されている。さらに、各週のエンド陳列の有無を表したエンド FLG（エンドフラグと呼ぶことにする）を用意する。エンド FLG は各店舗において調査員がエンド陳列を観測し、商品がエンドに陳列された週（陳列時）を 1、陳列されなかった週（非陳列時）を 0 として記録したものであり、特別な費用を投入して採取したものである。POS データ及びエンド FLG は KSP-SP 社より提供された。本研究では POS データにエンド FLG を加え、データセットとする。例として表 1 に A 店・ポテチのデータセットの一部を示す。本研究では各商品に対してデータセットが 4 つあり、計 12 のデータセットを扱う。なお、店舗によって販売されない期間が存在する商品があるため、データセットのサイズは最大 53 週 × 6 項目である。

表 1 A 店・ポテチのデータセット。

週	金額(円)	数量	価格(円)	来客数	FLG
201331	2688	32	84	12648	0
201332	2771	33	84	12064	0
201333	2519	30	84	11368	0
201334	840	10	84	12016	0
201335	1748	21	83.2	12589	0
201336	2604	31	84	11982	0
201337	2687	32	84	11822	0
201338	1327	16	82.9	11795	0
201339	2268	27	84	11793	0
201340	4946	66	74.9	11922	1

(2) データセット分析

以下にデータセットの傾向をまとめる。

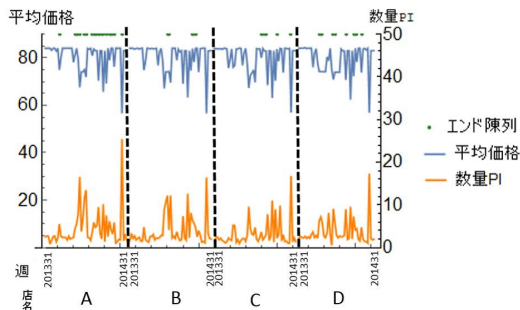


図 2 ポテチの時系列データ。

図 2～4 は各商品の平均価格、数量 PI (Purchase Index, 購買指数)、エンド陳列の有無の時系列データである。ここで、数量 PI は

$$\text{数量 PI} = \frac{\text{数量}}{\text{来客数}} \times 1000,$$

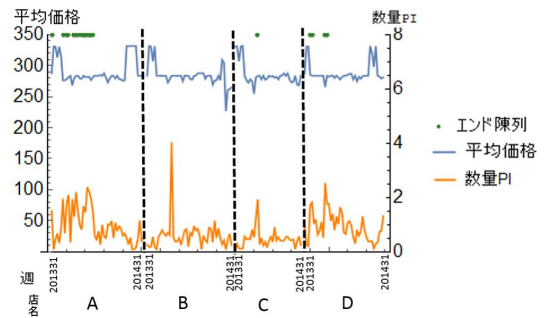


図 3 チョコパイの時系列データ。

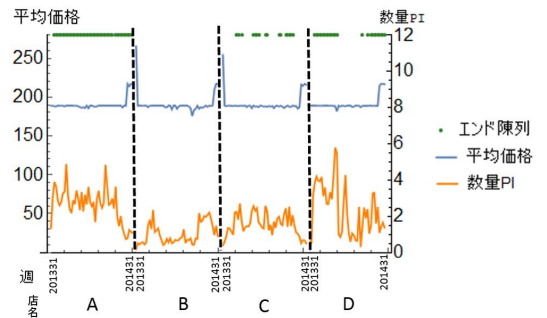


図 4 柿の種の時系列データ。

と表し、来客者 1000 人あたりの購買数を示す。これらの図からすべての商品に共通してエンドに陳列されると数量 PI が上がる傾向があること、商品によってエンド陳列の有無と価格の変動は異なること、ポテチにおいて価格が下がると数量 PI が上がることが言える。次に各商品の分布を見る。図 5～7 は各商品の数量 PI と価格の散布図である。チョコパイと柿の種については陳列時と非陳列時の分布に違いが出ていることが分かる。これらのことから本研究では、目的変数に「エンド FLG」、説明変数に「数量 PI」と「平均価格」を設定し、傾向の違う商品に対してモデルの正答率を比較する。

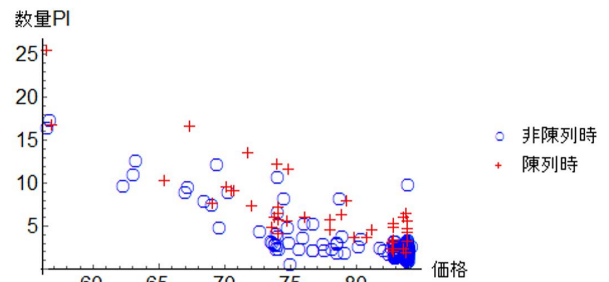


図 5 ポテチの散布図。

3. モデル

本研究の目的は、POS データの動向からエンドの陳列の有無を判別することである。したがって、POS データを 2 値分類モデルを用いて 0/1 の 2 つのクラスに判別する。本研究では以下の 4 つのモデルを用いる。

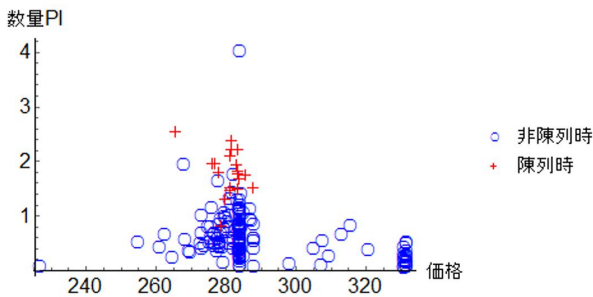


図 6 チョコパイの散布図.

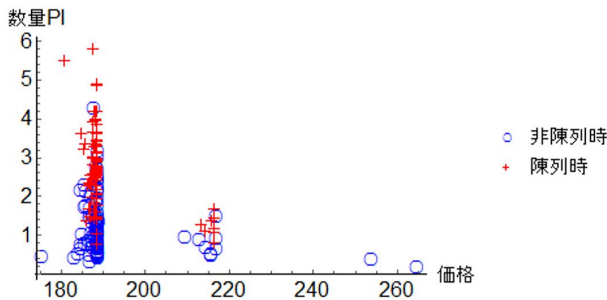


図 7 柿の種の散布図.

(1) サポートベクターマシン

サポートベクターマシン (Support Vector Machine, 以下 SVM) は、教師あり学習を用いるパターン認識モデルの一つである。2 クラス分類問題のために考案され、未知データに対して高い予測精度を持つ分類器を構築できる [4]。SVM では分類境界を形成する際に、各クラスのデータから可能な限り離れるように境界を定める。分類境界と学習データとの最短距離をマージンと呼び、マージンを最大化するような境界を選択する。

(2) ロジスティック回帰

ロジスティック回帰 (logistic regression) は統計的回帰モデルの一つである。医学や疫学において病因究明の方法論として開発され、リスク要因の解析に用いられることが多い。患者の生死や存在の有無などを表す 2 値変数の発生確率を予測する [5]。

(3) ユークリッド距離による判別

ユークリッド距離 (Euclidean distance) とは幾何学的な距離を指す。日常生活上の距離のことで、平方根 (2 次元)、立方根 (3 次元) をとる [6]。クラスを構成するデータ点と未知の点とのユークリッド距離を求め、各データ点との距離の平均が小さいクラスに分類する。

(4) マハラノビス距離による判別

マハラノビス距離 (Mahalanobis distance) は変数間の相関を考慮し計算される [6]。そのため、相関係数により等距離線が異なり、相関係数が 0 のときユークリッド距離と一致する。クラスを構成するデータ点と未知の点とのマハラノビス距離を求め、各データ点との距離の平均が小さいクラスに分類する。

4. 判別

本研究の目的は、POS データの動向からエンドの陳列有無を判別することである。したがって、「エンド FLG」と前記のモデルによる判別結果を比較し、正答率により評価する。判別の予測は 4-分割交差検証により行う。4-分割交差検証とは、データを 4 個に分割、そのうちの 1 個をテストデータ、残る 3 個を学習データとし、予測を行う。この処理を 4 個に分割されたデータそれぞれをテストデータとして 4 回予測を行う。本研究では商品ごとに 1 店舗のデータセットを 4 つに分割されたデータとみなし、3 商品のデータに対して交差検証を行う。つまり、4-分割交差検証を 3 回行う。表 2 は 4-分割交差検証を行った際の学習データのデータ数であり、不均衡データが存在する。クラスに属するサンプル数に偏りがあるデータを不均衡データと呼び、判別結果に偏りが生じる場合がある。本研究ではポテチとチョコパイにおいて陳列時のデータ数が少ない。不均衡データへの対処法としてデータ数を調整する SMOTE アルゴリズムがある。

表 2 学習データのデータ数.

商品		店舗			
		A	B	C	D
ポテチ	陳列時	21	37	36	32
	非陳列時	134	118	119	127
チョコパイ	陳列時	5	19	18	15
	非陳列時	120	110	121	117
柿の種	陳列時	41	87	71	62
	非陳列時	104	57	73	87

(1) 学習データ数の調整

SMOTE (Synthetic Minority Over-sampling Technique) アルゴリズムは学習データが少ない少数派クラスに対してデータを人工的に作成するアルゴリズムである [7]。各データ点と k 近傍点との間の点をランダムに選択し、少数派クラスに加える。これを指定したデータ数になるまで繰り返し行うことでデータ数を増やしていく。ここで指定したデータ数は少数派クラスのデータ数の定数倍になる。また、多数派クラスからはデータをランダムに削除する。データ数を調整した学習データ (以下、調整データ) のデータ数を表 3 に示す。

表 3 調整データのデータ数 ($k=5$) .

商品		店舗			
		A	B	C	D
ポテチ	陳列時	63	111	108	96
	非陳列時	84	148	144	128
チョコパイ	陳列時	55	57	54	45
	非陳列時	50	76	72	60
柿の種	陳列時	123	228	213	186
	非陳列時	164	171	284	248

(2) 多数決による判別

3. で用いた4種類の判別モデルから3つを選び、判別結果から多数派の結果を採用することとする。各モデルの判別結果の違いから正誤を補い合わせることを目的としている。

(3) 判別結果

判別した結果を表4~6にまとめる。なお、便宜上各モデルの頭文字をとったアルファベットで表記した。つまり、SVMをS、ロジスティックをL、ユークリッドをE、マハラノビスをMと表記し、多数決においては例えば、SVM、ロジスティック、ユークリッドを選択した場合はSLEと表記する。

表4 ポテチ調整データ正答率(%)。

	S	L	E	M
陳列時	69.05	59.52	64.29	73.81
非陳列時	86.14	90.96	81.33	83.13
全体	82.69	84.62	77.88	81.25
	LEM	SLM	SEM	SLE
陳列時	66.67	69.05	78.57	78.57
非陳列時	90.36	89.16	86.75	86.90
全体	85.58	85.10	85.10	85.24

表5 チョコパイ調整データ正答率(%)。

	S	L	E	M
陳列時	63.16	63.16	78.95	78.95
非陳列時	94.87	94.19	94.87	94.23
全体	91.43	90.80	93.14	92.57
	LEM	SLM	SEM	SLE
陳列時	73.68	68.42	78.95	78.95
非陳列時	94.23	93.59	94.23	93.59
全体	92.00	90.86	92.57	92.00

表6 柿の種調整データ正答率(%)。

	S	L	E	M
陳列時	78.16	72.41	74.71	73.56
非陳列時	80.37	87.76	82.24	83.18
全体	79.38	80.54	78.87	78.87
	LEM	SLM	SEM	SLE
陳列時	73.56	73.56	73.56	73.56
非陳列時	83.18	82.24	83.18	82.24
全体	78.87	78.35	78.87	78.35

5. 考察

4. の結果から高い正答率を確認できた。多数決においては低いときでも7割近い正答率を得ることが出来た。また、他のモデルと比較してもユークリッドとマハラノビスの正答率が安定して高く、特にマハラノビスは全てにおいて7割以上

の非常に安定した結果となった。これはマハラノビスの特徴である相関を考慮した距離計算により判別性能が高かったと考えられる。これは相関の強いポテチにおいてマハラノビスよりもユークリッドの正答率が劣っていることから見てとれる。SVMやロジスティック回帰についてはポテチやチョコパイにおいて陳列時正答率が6割程度で、他手法と比較して判別性能は高くない結果となった。多数決においても3つの内両手法を選択した場合に正答率が他パターンよりも劣っている。これらからマハラノビスや多数決によるエンドの陳列有無の判別は有用であると言える。

次に、商品ごとに比較を行う。チョコパイにおいて他商品と比較してユークリッドとマハラノビスの正答率が高い。これは陳列時において数量PIの上昇が顕著であることから判別が容易であったと考えられる。柿の種にも同様の傾向が見られ、どのモデルでも同様の正答率が得られた。ポテチにおいても陳列時の数量PIの上昇は見られるが、図5から分布が一部重なっており、判別に影響したと考えられる。

6. おわりに

本研究ではエンドの陳列データを補うために、POSデータの動向からエンドの陳列有無を判別した。傾向の違う3つの商品に対して、4つのモデルを適用した。結果としてどの商品に対してもほとんどのモデルで7割以上の正答率を得ることができた。市場反応モデルの構築に対して十分活用できるといえる。今後の課題としては、時系列を考慮した判別である。図2~4から全ての商品において価格の変動が同時期に起きていることが分かる。価格の変動原因について調査し、これを考慮したモデルを構築することで更なる精度向上が期待できる。

参考文献

- [1] 山中正彦, 「プロモーションにおける店頭の新定番売り場(エンド)の重要性-マイクロとブランド・マーケティングの統合へ」, マーケティングジャーナル, Vol. 26, No. 3 (通号 103), pp. 77-94(2007).
- [2] 「店頭を活性化する POS データと定点観測」, <http://www.yomiuri-is.co.jp/perigee/feature02.html> (2016/10/4 アクセス).
- [3] 佐藤忠彦・樋口知之, 「ビッグデータ時代のマーケティング-ベイジアンモデリングの活用」, 講談社 (2013).
- [4] C.M. ピシヨップ著, 元田浩・栗田多喜夫・樋口知之・松本裕治・村田昇監訳 「パターン認識と機械学習 下」, シュプリンガー・ジャパン (2008).
- [5] Annette J. Dobson 著, 田中豊・森川敏彦・山中竹春・富田誠訳 「一般化線形モデル入門(第2版)」, 共立出版 (2008).
- [6] 永田靖・棟近雅彦, 「多変量解析入門」, サイエンス社 (2001).
- [7] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research 16, pp. 321-35 (2002).