

## ニューラルネットによる牛乳販売量の予測

鈴木 督久(Tokuhisa SUZUKI)  
(株)日経リサーチ  
データベース第二部 部長

はじめに

マーケティング分野におけるデータマイニングは、顧客（個客）データベースや POS データを対象とすることが多い。両者は一般に「大規模」データベースとして蓄積されており、伝統的なマーケティング・リサーチが扱ってきた標本調査データとは性質が異なるためである。本稿ではその POS データに対するデータマイニングの事例として、ニューラルネットを利用した牛乳の販売量予測の事例を紹介する。

データマイニングの特徴には「大規模性」という量的側面ばかりでなく、データが比較的粗悪であるという質的側面もある。敢えて「粗悪性」と表現したが、決して“データラメで使いモノにならない”という意味ではない。

データの質を統計学的観点からみると、無作為化・局所管理・反復の3原則に従って要因を統制しながら計画的に収集した「実験データ」が最も良質である。次に、統制はできないものの無作為抽出はしている「調査データ」である。そしてデータマイニングが対象とするデータはビジネス過程で蓄積された「実績データ」であることが多い。必ずしも母集団からの確率標本ではなく、要因の統制もなければ無作為性もない。能動的・計画的に収集したデータではなく、受動的に集まったデータである。その意味では質的に劣悪だが、大規模であり時には母集団そのもの、ビジネス過程そのものでさえある。従って複雑な現実をそのまま含んで集積されていることが多く、データの構造を探索する最初期の作業が重要である。単純化して整理すると図1のように「データの質」のレベルを3種類で示すことができる。

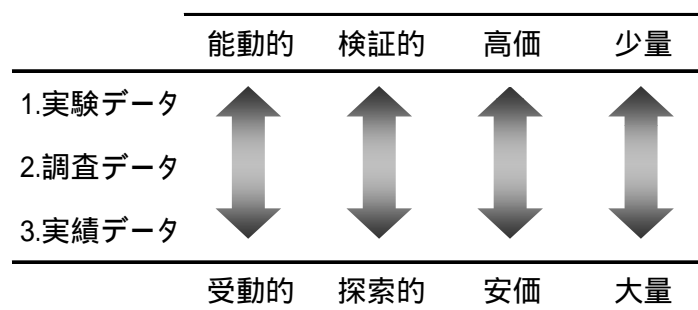


図1 データの質による3分類

POS データは粗悪とはいえないが、標本調査に比べれば受動的に集まるデータである。

商品分類コード管理のために膨大な人件費を投入するので安価でもないが、商品に関するデータ項目は「商品コード」「日付」「価格」「数量」の4つしかない。店舗コードもあるが通常のPOS情報サービスには含まれない。分析用には店舗別の「来店客数」が日単位で得られるが、パネルデータと違って来店客の属性項目はない。データが毎日発生するので時系列に沿って大規模であるが、特別な意図で調査した項目はないから、分析にあたっては「新しい変数」を加工・生成するアイデアが重要となる。

データマイニングは統計学者が「ゴミを入れればゴミしか出ない」と昔から強調してきた教訓に反して、結果的に集まった大規模で汚れたデータから知識を発見するという側面がある。確率標本の観測値から、母集団の未知の値を頻度論的表現で推定・検定するのが目的ではない。ビジネスに有用で相対的に低コストであれば「ゴミ」であろうが、そこから「宝」(知見)を搾り出す。推測より予測が、検証より探索的記述が目的であり、ゴミも積もれば知識となる(かも知れない)のである。「量が質を変える」ということはあり得る。

## 1. ニューラルネットによる予測モデル

ニューラルネットによる予測モデルは、伝統的な線形重回帰分析よりも高い予測成績を示すことが多いといわれるが、最近までプログラミングのできる専門家の協力が必要であった。1990年代後半からは、使いやすく処理速度も実用的に改善されたニューラルネットのソフトウェアが普及し、データマイニング用ソフトウェアには標準的に装備されるようになった。ビジネスマンが比較的容易に利用できるツールとなったのである。

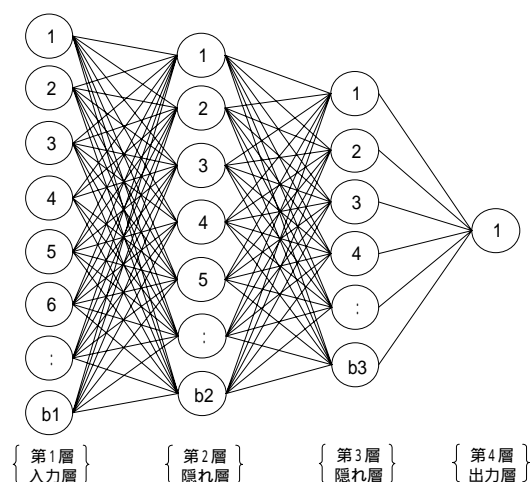


図2 階層型ネットワークモデルのトポロジー

図2が牛乳の販売量を予測する階層型ネットワークモデルのトポロジーである。4層で構成し、左から第1層を入力層、最終の第4層を出力層という。入力と出力の途中は中間層あるいは隠れ層という。各層は1つ以上のユニットで構成される。入力信号は入力層のユニットに入力され、ユニットのつながりを通して次の層へと伝達され逆戻りしない。ま

た同一層内ではユニット間で信号の伝達はされない。

ユニット間の信号は加重和として変換されて次の層のユニットに対する入力信号となる。重みはユニット間の結び付きの強さであり、図2ではユニットを結ぶ線分で表示している。一方、第2層以降のユニット内部では非線形のシグモイド関数（ロジスティック関数のような連続関数）によって信号が変換される。重みは適当な初期値からスタートして最終出力信号と教師信号との誤差を最小化（最適化）する目的に向かって反復計算しながら重みを更新して収束を目指す。

重回帰分析の用語と対比させると、入力層は独立変数でありユニットの個数 $b_1$ が変量数である。入力信号とは入力データのことであり各変数の観測値に相当する。最終（第4層）の出力層は従属変数であり、出力信号は予測値である。教師信号とは従属変数の値である。出力ユニットは1個なので非線形重回帰分析モデル、または判別分析モデルに相当する。ユニットを結ぶ重みは偏回帰係数に相当し、重みの推定のことをニューラルネットでは学習という。隠れ層は多変量解析の概念に対比させれば合成変量に相当する。入力ユニットよりも隠れユニットが少なければ情報縮約をすることに相当する。

## 2. 分析者の関与

隠れ層という用語はブラックボックスを連想させ、ニューラルネットはデータに合わせて自動的に非線形関数を選ぶという機械的印象を与える。実際、ニューラルネットによる予測では伝統的な重回帰分析で常識となっていた変数選択を考えなくてもよい。その代わりにモデルの安定性はクロスバリデーションを使う。またパラメータに関して線形である必要はないので、分布の対称化や線形化のための再表現（変数変換）も特に実施しない。

しかし実際にニューラルネットを応用する場面では、やはり分析者である人間が関与し、分析者の裁量・判断が重要となる点がいくつかある。表1はそれを主要な6項目としてまとめたものだが、概説を加えてから具体例を示すことにしよう。

表1 ニューラルネットにおける分析者の関与

1. 従属変数の選択	なにが分析の目的なのかを決めることに相当
2. 独立変数の選択	予測に影響しそうな変数の洞察・考案する
3. 隠れ層の個数	通常は1個だが2個以上の必要があるか検討
4. 隠れユニットの個数	さまざまな個数が選べるの試行錯誤になる
5. クロスバリデーション	訓練・妥当・検証の3種類のデータの割り当て
6. 予備解析	生データからの加工や外れ値の検討など雑多

第一に独立変数（入力信号）の選び方である。POSデータの変数は少ないが、そこから新しい予測変数を作ることはできる。何が牛乳の販売量に影響しているかという洞察力や

業務知識がここで役に立つ。変数選択で冗長な変数を減らす作業は不要ではあるが、役に立つ変数を見逃すことは分析者の怠慢である。

第二に従属変数であるが、これは分析目的に応じてほぼ自動的に決まる。ここでは牛乳の販売量であるが、具体的には販売本数を使う。販売金額も候補であるが、売上管理よりも在庫管理が目的なので本数が予測目標となるのである。

以上は伝統的な重回帰分析でもほぼ同じ事情を持つものであるが、ニューラルネットの場合は、第三の判断として隠れ層の数を選択する判断がある。一般には隠れ層が1つの三層モデルが頻用される。隠れ層が1つでもユニット数を増やせば非線形関数を近似できるという数学的性質のためである。しかしここでは図2のように隠れ層を2つにした。これは予測の安定性を試行錯誤した結果である。

第四に隠れ層内のユニット数である。入力ユニットが多ければ、隠れユニットの個数も多くの選択があり得るので、候補として試すモデルは相当な数となる。ニューラルネットのソフトウェアが実用的に高速化した現在でも、データマイニングに取り掛かると何週間も分析に没頭する結果になる。小さなモデルでは、ブートストラップ法を適用してEIC(エフロンの情報量規準)が最小となる個数を選択する方法を辻谷(2001)が示している。

第五にクロスバリデーションに関する判断として、トレーニング(訓練)データとバリデーション(妥当化)データの割合を決めることである。最近のデータマイニング用ソフトウェアには、割合を指定するとランダムにデータを分割する便利な機能が標準的に用意されている。さらにモデルの性能を確認するためのテスト(検証)データを用意する。このデータはトレーニングにもバリデーションにも使っていないデータである。ニューラルネットは過学習をしやすいので、クロスバリデーションは極めて重要な作業である。

第六に予備解析は依然として重要で不可欠である。特に極端な外れ値がある場合は、素性を確認したうえで、適切であれば除外するなどの判断をする。ここでは探索的データ解析の精神はそのまま有効である。また広義の予備解析として事前のデータ処理が必要なことが多い。データマイニングのデータは複雑で大規模であり、特定の分析手法に対応した形式で用意されているわけではないので、マイニング時間の多くはこの作業に費やされる。しかし、ニューラルネット用の分析データ形式に加工するまでの過程でデータに関する知識が蓄積されるので軽視できない作業である。

データマイニングには「ツールを使って一獲千金」のイメージがあるが、実は暗い坑道を這うようにして掘り進んで鉱脈を探しあぐねたり、地道な農耕作業の果てにじっくり収穫するような作業をするのである。

### 3. 従属変数と独立変数

予測目的の牛乳販売量として本数と金額が選択できるが、両者の相関は一般に極めて高く、どちらを使っても目標は達成できる。店舗によっては極端な値引き戦略を計画的に実施しているために、単純な線形関係が微妙に崩れる場合もあるが、ここでは在庫管理が目的なので本数を従属変数とする。

またモデルは店舗ごとに作成する。最終目標はある地域全体の牛乳出荷量の調整を想定しているが、販売の構造は店舗で異なるため、店舗単位の予測値を最終的に合計する方式

を考える。

独立変数としては表 2 に示す 75 変数を作成した。POS データの変数は少ないが、そこから新しい変数を作成することが重要な技術である。たとえば「日付」変数からは販売動向に影響しそうな曜日という変数を作成した。また「ある銘柄」だけの予測ではなく店舗全体の牛乳の販売予測を目的とするのだが、各店舗には主力銘柄があるので安定的に置かれている銘柄のうち上位 4 銘柄を識別する変数を作成した。販売動向は時系列の影響を受けると考えられるので、それをネットワークに学習させるためにラグ付き変数を 3 日前に遡って作成した。

この他、POS データではないが、店舗地域の天候データを外部から購入できるので、それも独立変数として使用した。「当日」の天候データが使われている項目は、前日に高い精度の予報値を得ることができるもので、実際の予測では天気予報値を代入する。

表 2 独立変数（入力ユニット）の一覧（75 変数）

変数の内容	個数
曜日 (2 値変数化)	7
3 日前までの販売個数 (店舗全体と上位 4 商品)	15
3 日前までの販売価格 (上位 4 商品)	12
3 日前までの来店客数	3
3 日前までの降水量	3
3 日前までの日照時間	3
当日および 3 日前までの不快指数	4
当日および 3 日前までの最高気温	4
当日および 3 日前までの最低気温	4
当日および 3 日前までの平均気温	4
当日および 3 日前までの湿度	4
当日および 3 日前までの天気 (2 値変数化)	12
合計	75

#### 4. 隠れ層とクロスバリデーション

隠れ層の数とユニットの数を決定する作業はクロスバリデーションと関連する。ここではモデル構成用に、ある店舗の 1998 年 4 月 1 日～1999 年 3 月 31 日までの 1 年分を使った。トレーニングデータとバリデーションデータの比は 6 : 4 から 7 : 3 の範囲で試行錯誤して最終的に 7 : 3 とした。抽出は単純無作為法による。モデルの安定性を確認するためのテストデータは推定には全く使用しない 1999 年 4 月 1 日～4 月 30 日までの 1 か月分とした。すなわち、この予測モデルは過去 1 年分の POS データで構成し、その後の 1 ヶ月間の予測業務に利用するのである。1 ヶ月後にはモデル構成データを 1 ヶ月ずらして再構成する。過去のデータは 2 年でも 3 年でも利用可能であるし、推定の立場からはデータが多い方が好ましい。しかし市場構造は常に変化しているのであまり遠い過去のデータを利用する意味もない。モデルを動的に月次更新することが現実的だと判断した。

表3はユニット数のさまざまな組み合わせによる予測結果の比較である。記号 $b_2$ と $b_3$ はそれぞれ第2層と第3層の隠れユニットの数である。ユニット数が少ないと予測力は弱い。ユニット数を増やしていくとバリデーションデータでさえほぼ完全に学習してしまう。そこでテストデータに対する安定性を確認することが重要になる。このモデルでは多くの階層型ネットワークモデルと違って、隠れ層を2つに増やしているが、バリデーションデータだけでなく、テストデータに対する安定性を考慮した結果である。

モデルの評価指標として重回帰分析と同様にさまざまな統計量を考えることができるが、ここでは牛乳の販売本数を予測したいという現実的な目的があるので、同じ単位すなわち本数で見積もって、それが実用的な誤差であるか否かを判断するのが端的で分かりやすい。そこで残差の絶対値の平均を使った。ここではモデル3が良さそうである。

表3 ユニットの個数とモデル評価

Model	予測値との相関係数		残差の絶対値の平均	
	Valid	Test	Valid	Test
(1) $b_2=20, b_3=10$	0.496	0.873	223	135
(2) $b_2=30, b_3=15$	0.999	0.622	6	209
(3) $b_2=40, b_3=20$	0.996	0.678	6	196
(4) $b_2=50, b_3=20$	0.999	0.534	4	203

## 5. 予備解析によるデータの吟味

順序としては最初を実施するのが予備解析である。多くの話題があるが、まずは図3や図4のように販売状況を折れ線グラフやヒストグラムで表現することから始まる。この店ではいくつかの極端な外れ値が観察できる。素性を調べた結果、特売日や大晦日などの人為的な特異日だったので、予測モデル構成にとっては除外して問題はないと判断した。

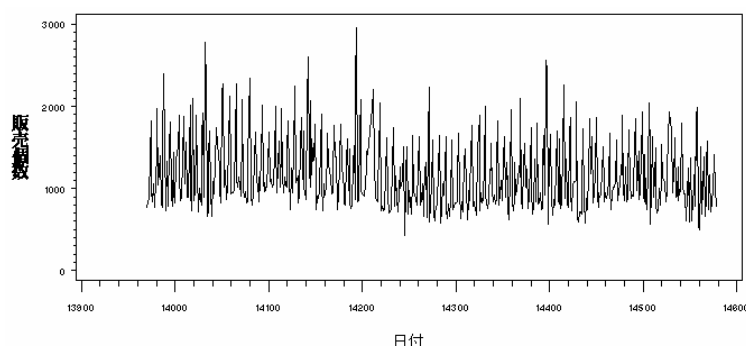


図3 ある店舗における牛乳の日別販売個数の推移

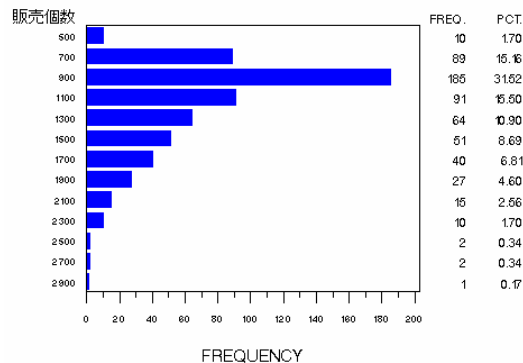


図4 ある店舗における牛乳の日別販売個数の分布

この他にも予備解析では多くのことを検討した。店舗ごとに販売実績のある商品を調べると、店舗によって扱う商品がかなり異なることや、短期間で消える商品も多いことが分かり、これらの扱い方を検討しなければならなかった。また1リットル牛乳だけで全体の9割前後を占めるので予測対象を限定することにした。値引きがどのように実行されているかも店舗ごとに観察した。休店日や外れ値を除外しても「前日」などのラグ付き変数に影響しないようにデータ処理するなど、瑣末な作業がデータマイニングを支えている。

おわりに

紹介した事例は日経リサーチがデータマイニング事業を開始するに当たって、研究的な目的で実施した事例のひとつである。ニューラルネットで使用したソフトウェアは SAS Institute Inc. の Enterprise Miner, Release 2.00 である。

日本経済新聞社・電子メディア局には同社の POS 情報サービス「NEEDS-SCAN」を使用する便宜をはかっていただいた。豊田秀樹氏（早稲田大学文学部教授）とは協力しながらデータマイニングを勉強する機会を与えていただき、特にニューラルネットの理論と実践に関して有益なご教示をいただいた。お礼申し上げます。

参考文献

- 1) 日経リサーチ(2000):『POS データに対するデータマイニング事例集』,日経リサーチ。
- 2) 鈴木督久(2000):『POS 情報を利用したデータマイニング事例』,データマイニング・シンポジウム論文集,日中統計シンポジウム。
- 3) 辻谷将明(2001):『ニューロ判別モデルとリサンプリング法』,日本科学技術連盟・多変量解析シンポジウム発表要旨。
- 5) 豊田秀樹(1996):『非線形多変量解析—ニューラルネットによるアプローチ—』,朝倉書店。
- 6) 豊田秀樹(2001):『金鉱を掘る統計学 データマイニング入門』,講談社。