データマイニングと意思決定

徐　良為

近年，データマイニング技術の発展により，データマイニングが合理的な意思決定に重要な役割を担うようになってきた．データマイニングで得られたモデルの予測・判定結果が直接，または，間接的に日常活動における意思決定に大きな影響を与える．一方，データマイニングが直面する問題のほとんどが，複雑かつ多様で，不確実性に満ちたものである．データマイニングで得られる結果も決して唯一のものではないので，複数結果から，現実世界を最も忠実に反映したモデルを選択する意思決定も必要である．本稿は，汎用データマイニング環境に，意思決定機能を導入する実践から，データマイニング過程の意思決定，および，データマイニング結果が意思決定における役割について述べる．

キーワード：意思決定，データマイニング，不確実性，モデリング，最適化，近似解法

1. はじめに

データマイニングとは，人間がコンピュータのデータ蓄積能力および計算能力を借りて，膨大かつ複雑なデータに潜む要素間の本来の規則性，対応関連を見つけ出す作業である．近年，データマイニング技術は，急速に発展し，さまざまな分野で普及してきている．その背景には，複雑性，多様性，不確実性に満ちた現代社会を理解し，意思決定を行うために，問題の本質を割り出し，原因と結果の関連を明確にすることの重要性が再認識されている．

人間社会では，絶えず意識的，あるいは，無意識的な選択が行われている．朝，出かけるときに傘が必要かどうか，休日の過ごし方，スーパーでの買物，Web 閲覧ページ，TV チャンネル番組，経営戦略，製造過程の選択などである．これらの選択は，直感を頼りに判断する部分もあるが，科学的に，合理的な選択，つまり，意思決定が必要な場合もある．意思決定 (τ) は一般的に次のように表すことができる．

$$\tau(y) = \arg\max_{z \in Z} \quad f(y, z)$$

ここで，f は効用関数，y は外部環境，z は制御，選択などの「行動」をそれぞれ表す．τ は与えられた外部環境 (y) に対して，選択可能な領域 (Z) から，効用が最大となるような行動 z を算出する (arg max は関数 f が最大となるような引数 (z) を求める記号である)．例えば，スケジューリング問題を考える．f は構成員の満足度合，y は必要な人員の数，z はスケジュール割り当てとした場合の意思決定問題となる．しかし，現実世界では，y が必ずしも直接観測可能なものとは限らない．在庫管理の場合は，商品の適正な在庫量を求めるために，商品の将来における「需要量（消費量）」を知る必要がある．一般的には，将来の時点での商品の需要は，さまざまな外部要因から影響を受けるので，直接観測可能なものではない．ここで，意思決定関数 (τ) を次のように拡張する必要がある．

$$\tau(x) = \arg\max_{z \in Z} \ f(y := g(x, z), z) \tag{1}$$

ここで，x は直接観測可能な外部環境（説明変数と呼ぶ）を表し，g は y（目的変数）と x, z 間の依存関係を表す．より一般的には，現実世界では，われわれが事前に収集可能な観測量に限度があり，y の値を確定するに必要な要因 x の一部しか知ることができないので，手元の観測量 x に対して，y が取りうる値に不確実性が存在する．不確実性を扱うためによく用いる方法は，τ を次の「期待効用」を最大となるような z を求めることになる．

$$\tau(x) = \arg\max_{z \in Z} \sum_{y} f(y, z) P(y|x, z) \tag{2}$$

ここで，$P(y|x, z)$ は条件確率を表す．

データマイニング分野では，関数 g と条件確率 $P(y|x, z)$ をモデルと呼ぶ．合理的な選択の成否は，モデルの精度に大きく依存する．データマイニングの主な目的の一つは，収集されたデータからモデルを作成すること，いわゆる「モデリング」である．

以下では，データマイニング手法，特にモデリングを紹介し，高精度なモデルを構築する方法を述べ，デー

じょ りょうい
（株）数理システム データマイニング部
〒 160–0022 新宿区新宿 2–4–3

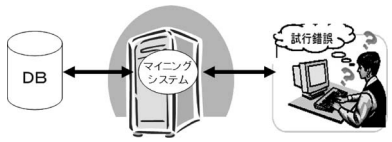

図 1　データマイニング作業図

タマイニングの適用シナリオ例を挙げる．

2. データマイニングモデリング

データマイニングの目的は，主にデータから，要因間の依存関係を表すモデルを推定したり，属性の値の近いもの同士をグルーピングしたり，同時発生事象を算出することである．データマイニングプロセスは主に次のステップを含む．

(1)　適用対象ドメインと目標の明確化

まず，データマイニングによる目標を明確にする．半導体製造のマイニングの場合は，シリコンウェハーの歩留まりに影響する要因を知りたいのか，製造ラインの故障診断，生産計画に必要な需要予測，製造工程の最適制御，流通業でのお勧め商品のレコメンド，顧客の行動を分析するなどを明確にする．

(2)　データ収集

さまざまなデータの格納先（DB システム，CAD, CAM，MRP，ERP システム）から，マイニングの目的に適したデータを選択・収集・統合する．有効なマイニングを行うためには，目的に適した精度の高いデータを選択・収集することが必要不可欠である．ここでのデータの良し悪しがマイニングの成功を左右する決定的なファクタである．データマイニング作業の大部分がデータ準備にあると言っても過言ではない．

(3)　データ加工

取得されたデータを分析に適した形式にするために，データの整理・整頓を行う．具体的には，データに含まれる欠損値やはずれ値を補填もしくは除外したり，分析に適さない内容を除外したり，データの単位を統一したりする．

(4)　マイニング

ここまで来てはじめてデータマイニングの核心部分に入る．ここは主に，マイニングアルゴリズムを選択し，データから有用なパターンを抽出したり，モデルを作成したり，仮説を立てたり，必要なデータを抽出したりする（図 1 参照）．

(5)　結果表示

意思決定者に分かりやすい形で，抽出された知識をレポートしたり，モデルを生産制御システムへ展開したりする．

通常，マイニング対象データ（収集，加工済み）は，次のようなテーブル形式で表す．

表 1　製造工程記録データ

工程A.温度	工程A.材料	工程A.加工時間	工程B.温度	工程B.材料	工程B.加工時間	合否
28.108	M	4.901	28.424	W	9.346	Yes
42.379	M	5.044	36.758	W	11.171	Yes
20.597	L	2.802	22.688	T	14.914	No
21.860	L	2.909	10.230	S	16.317	Yes
23.921	M	5.005	23.880	W	11.112	Yes
32.736	M	3.828	22.464	T	10.783	Yes
27.400	M	3.942	27.066	T	12.349	Yes
15.211	L	3.119	19.572	S	13.873	Yes

表 1 はある精密機器の製造データを表している．各列（変数と呼ぶ）は，製造条件および製品の最終検査結果（合否）を表している．各行は，製品が経由した製造過程を表している．マイニング作業は，この製造データから，製造条件（説明変数と呼ぶ）が製品の合否（目的変数と呼ぶ）に与える影響を調べたり，製造条件と合否間の関数（モデルと呼ぶ) を求めたりする．また，データの各列（変数）は，実数，整数，カテゴリに分類される．例えば，「合否」の取り得る値（Yes と No）のようなものがカテゴリで，「温度」は実数である．

マイニングのコア部分は，主に次のように分類される．

1. モデリング（回帰・分類）
2. クラスタリング
3. アソシエーション分析
4. 時系列分析
5. その他

本稿は主にモデリングについて述べる．モデルは入力（x）と出力 (y) の対応関係を規定するものである．入力から，出力をユニークに決定可能なものであれば，モデルを関数で表すことができる．入力に対して，出力をユニークに決められないもの，つまり，不確実性を含むような場合は，モデルを条件確率 $P(y|x)$ で表す場合が多い．モデルは通常，既存データから作成（学習）される（人間の経験によるものもある）．モデルの出力データタイプによって，2 種類に分けられ，出力データが数値の場合は回帰モデル，出力データがカテゴリの場合は分類モデルという．データからモデルを作成する過程は学習，または，フィッティングと呼ばれる．図 2 は，フィッティングについて簡単な例を表している．

図 2 に，入力 x と出力 y の対応関係を表す 11 点（●，○）のデータがある．そのうち 4 点（黒丸●）のデータをモデル作成するための学習データとする．フィッティングは，その 4 点のデータから，入力 x と出力 y

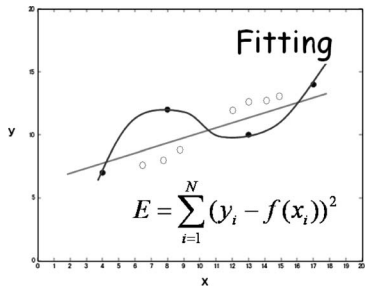

図 2　フィッティング

間の本来の関係を推定する．モデルの x と y の関係を一本の線で表した場合は，直線で表したり，四つの点をすべて通るような曲線で表したりするなど，無数の可能性がある．それらの線の良し悪しを評価するためによく使われるのが，図 2 に示した誤差（点から線までの y 軸における距離 E）式である．この評価基準の下では，直線より，すべての点を通るような曲線の方が圧倒的によい（誤差＝ 0）ことになる．この直線のように，比較的学習データにあまりフィットしない単純なモデルを UnderFitting という．学習データが四つの点に限って言えば，すべての点を通るような曲線が一番データにフィットしたように見えるが，図のように，学習データ以外に，さらに，学習時，導入しなかった 7 個のサンプルデータ（白丸 ◯）で検証を行ったところ，明らかに，曲線より直線の方の誤差が小さいことがわかる．このような学習データに対してはよくフィットするが，検証データ（あるいは，本当の母集団データ）にはフィットしないことを OverFitting（過剰学習）という．モデル本来の目的からすれば，学習データにだけフィットしてもよいモデルとは言えない．過剰学習を検証するための有力な手段の一つは，交差検証である．

交差検証とは，図 3 のように，データを学習データと検証データに分割し，学習データでモデルを作成し，このモデルを用いて，検証データに対する予測を行い，モデル精度を評価する方法である．データ分割は一度だけでなく，検証データをそれぞれ異なるデータブロックで，万遍なく選択して検証が行われる．

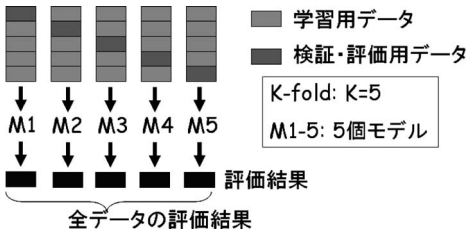

図 3　交差検証

3.　モデル作成ための意思決定

データマイニングのモデル作成（学習）方法は数多く存在する．現在も新しい技術が研究，開発され続けている．代表的な学習方法としては，線形回帰モデル，決定木，ニューラルネットワーク，k-NN，サポートベクトルマシン，Naïve Bayes などがあり，複数モデルを組み合わせる集団学習方法もある．モデルの推定（予測）精度は学習データに大きく依存する．モデルが一旦作成されたあと，モデルの推定精度だけではなく，データへの頑健性（学習データの内容によって大きくぶれないこと），説明能力（モデルそのものが，人間の経験などに照らし合わせても，十分な説得力があるもの）などを検証，評価する．期待した結果に達しない場合は，データの収集段階へ立ち返って，場合によっては別の説明変数を集め直す必要がある．モデルは図 4 のように試行錯誤の繰り返しで構築される．

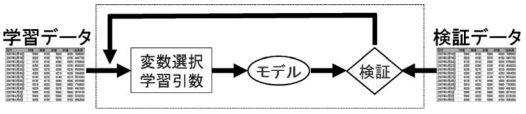

図 4　モデリングの試行錯誤

一般的には，与えられたデータに対して，どんなモデル作成方法で，作成時のパラメータをどのように与え，どのような説明変数を用いれば，精度のよいモデルを得られるかを事前に知ることができない．そのために，マイニング担当の技術者は，モデル作成と検証を繰り返す試行錯誤でよいモデルを選択する意思決定が必要である．モデル構築時の意思決定は，次のように定式化することができる．

$$\arg\max_{\theta,\eta} \; e(f, a, \theta, \eta)$$

ここで，e は交差検証のようなプログラムルーチン，f はモデル（説明変数から目的変数への関数，または，条件確率 P），a はデータ，θ はモデル構築パラメータ，η は説明変数候補（0-1 テーブル）をそれぞれ表す．意思決定関数の結果は，最良の評価値（e の結果）となるようなモデルの構築パラメータ（θ）および説明変数の組合せ（η）である．評価関数の e は交差検証のようなプログラムであり，一般的には数式で表すことができ

ず，もちろん，微分することもできない．また，説明変数選択は整数計画問題であり，一般的にグローバルな最適な解を求めるのがNP–困難と呼ばれ，計算機が現実的な時間内で解を求めることはできないので，近似的な解法を求めるしかない．

近似解を求めるには，局所探索，タブーサーチ，アニーリング探索，遺伝的アルゴリズム，ES アルゴリズム，PSO などの方法がある．ここは，われわれが実践したタブーサーチ（Tabu Search）[3, 8] について説明する．タブーサーチは巡回セールスマン問題のような組合せ最適化問題を解くためのメタヒューリスティクス（meta-heuristics）アルゴリズムである．タブーサーチは停止条件を満たすまで解の近傍探索を行う．タブーサーチの特徴として，調査済み解をすべてタブーリストに登録することにより，一度調査した解を繰返し調査することを防ぐと同時に，未調査の近傍解に対しては，既知解よりも評価値が悪いものであっても調査を行う．図 5 のように，さらなる高い山を目指すために，ローカル最適解（山）から一旦降りることもある．

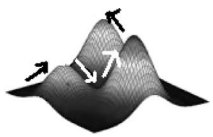

図 5　タブーサーチ

巡回セールスマン問題を解くときに用いるタブーサーチとは異なり，モデルチューニングでパラメータおよび説明変数を選択のために用いる際は，次のような点を考慮する必要がある．

1. 目的関数（e）は必要に応じて，ユーザが自由に定義可能なものでなければならない

2. モデルパラメータが連続値の場合は，事前にデータを手動でカテゴリ化するのではなく，自動的に最良な値を求めるのが望ましい

3. 目的関数は，複数回のモデル作成・評価が含まれるため，一回の評価の計算コストが高い．そのため探索空間内の近傍値選択に工夫が必要である．つまり，少ない探索回数で，よりよい解を求める必要がある

4. 探索が長時間にわたる可能性があるので，途中で計算が中断されても，あとで中断した時点から計算を再開できる必要がある

われわれの実践では，上記 1 に関しては，データマイニングシステムに含まれる既存のスクリプト定義で実現した．2 に関しては，近傍計算の場合は，PSO（Particle Swarm Optimization）[5] に近いアルゴリズムを導入した．3 に関しては，実験計画法での考え方に近い方法で実現し，4 は探索途中のすべての結果，状態を HDD に保存するように実現した．

上記 1 で述べたような目的関数はユーザが自由に定義することが可能であれば，第 1 節で述べたすべての意思決定のための式を解くことが可能となる．

タブーサーチでの実践は，通常の数理計画のベンチマーク問題，およびモデルチューニングの両面で実施したところ，満足できるような結果が得られた．

4.　意思決定の適用例

ここまでは，意思決定のための最適解を求める方法を示した．本節は，実例を通して，1 節で述べた意思決定の適用シナリオ [1, 4, 6, 7, 9] を示す．

【製造業における品質管理】データマイニングの製造業への適用は，製造過程，制御，メンテナンス，品質改善，欠陥検出，エンジンニアリングから，顧客管理（CRM），意思決定まで幅広く行われている．

製造条件（製造パラメータ）の最適化は，品質改善にとって極めて重要なファクタである．製造の初期段階において，特に製造データが少ない場合は，最適な製造パラメータを得るために，実験計画法がよく使われる．

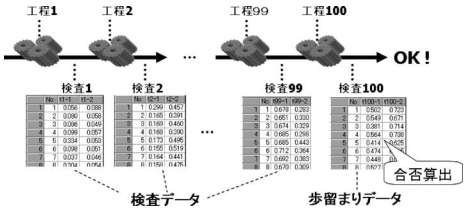

図 6　製造工程の検査データ収集

図 6 のように，各製造工程で検査モニターを設置し，そこから製造データを蓄積することにより，データマイニング手法が適用可能となる．データマイニング手法を用いることによって，従来では，単純化された制御パラメータの推定しか行えなかった状況を改善し，品質に直接影響する制御可能な要因と制御不可能な要因の相互作用を考慮した，より現実に近いモデルを利用して，最適な制御パラメータを探索することが可能と

なる．

1節の式 (1)，(2) を対応させると，f は品質指標（例：歩留まり値），y は製品の品質指標，z は制御可能な製造条件，x は観測可能だが制御不可能な外部要因であり，g と P はデータマイニングによる得られるモデルとなる．

【コールセンター】コールセンターにおいてはオペレーターのスケジューリング（人員配置）は直接経営効率に関わる問題である．過剰な人員配置は資源の無駄となり，過少配置は顧客対応を悪化させ，企業イメージに関わる重大な問題となる．最適な人員配置を求めるために，顧客からのコール需要が必要不可欠である．将来の時点において，顧客からのコールを直接観測することができないので，モデリングを用いる必要となる．

1節の式 (1)，(2) を対応させると，f は応答効率，y は時間帯別，用途別のコール需要，z は人員配置，x は今までの需要実績および外部要因，g と P はモデルとなる．

【省エネ制御】地球温暖化が社会生活に与える影響を緩和するためには，省エネルギの取り組みが必要不可欠である．省エネの取り組みにおいて着目される点としては，主に供給側のエネルギの生産効率，消費側の消費効率，および生産と消費のマッチングが挙げられる．

エネルギの供給側，例えば，風力発電の生産効率と言えば，意思決定式を対応させると，f, y は発電効率，z は制御可能な発電装置のヨー角度，ピッチ角など，x は観測可能だが，制御不可能な風向き，風速，気温，g と P は外部要因による発電モデルを表す．また，火力発電の場合は，f, y はボイラー効率，z は制御可能な空気投入量，石炭投入量，投入空気の温度など，x は観測可能だが，制御不可能な外部気温，石炭品質，g と P は発電モデルを表す．

エネルギの消費側に関しては，f, y はエアコン・冷蔵庫などの電力消費（このとき，max を min に変える），z はモータ印加電圧，圧縮機モータ回転数など，x は気温，人体の不快度数など，g と P はエネルギ消費モデルを表す．

5. おわりに

データマイニングと意思決定の関係について述べた．意思決定のための最適解 を求める機能実装の実践を紹介した．意思決定がデータマイニングの目的であると同時に，データマイニングのプロセスにおけるモデルの選択にも，意思決定が必要である．意思決定の観点から見れば，外部条件がすべて把握可能な場合はデータマイニングを行う必要がない場合があり，逆にデータマイニングの結果そのものが直接意思決定に使われることもある．しかし，多くの場合はより合理的な意思決定を行うために，データマイニングが必要不可欠である．本文で紹介した適用シナリオは，データマイニング応用例の全体のほんの一部分に過ぎない．現在も新しい適用事例が絶えず増え続けている．

データマイニングは統計・人工知能・データベースシステム・パターン認識など多くの分野の知見を必要とする実践的な総合技術である．データを保存，管理するために，データベースシステムが欠かせないのと同じように，意思決定にはデータマイニング技術が欠かせない．

参考文献

[1] A. Kusiak, H. Zheng and Z. Song: Power Optimization of Wind Turbines with Data Mining and Evolutionary Computation, *Renewable Energy*, Vol. 35 (2010), 1324–1332.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006

[3] F. W. Glover and M. Laguna, *Tabu Search*, Kluwer Academic Publishers, 1997.

[4] J. A. Harding, M. Shahbaz, Srinivas and A. Kusiak: Data Mining in Manufacturing: A Review, *Journal of Manufacturing Science and Engineering*, Vol. 128 (2006), 969–976.

[5] R. Poli: Analysis of the Publications on the Applications of Particle Swarm Optimisation, *Journal of Artificial Evolution and Applications*, Vol. 2008 (2008), 1–10.

[6] S. He, Z. He, G. A. Wang and L. Li: Quality Improvement Using Data Mining in Manufacturing Processes, in *Data Mining and Knowledge Discovery in Real Life Applications*, Edited by J. Ponce and A. Karahoca, 357–372, 2009.

[7] Z. Song and A. Kusaik: Constraint-Based Control of Boiler Efficiency: A Data-Mining Approach, *IEEE Transactions on Industrial Informatics*, Vol. 3 (2007), 73–83.

[8] 数理システム: Visual Mining Studio V7.1 マニュアル, 技術資料, 2011.

[9] 徐良為, 膨大なデータから規則性を抽出するデータマイニング技術—データマイニングによる品質管理と省エネ制御, 電気学会誌, Vol. 131 (2011), 617–620.

