データマイニングにおけるアルゴリズムの応用例

徳山　豪
東北大学大学院　情報科学研究科

IT(Information Technology) に不可欠な作業として, データベースにたまったデータからわかりやすい情報や知識を引き出す技術を引き出すこと, すなわちデータマイニングがある. データベースの巨大化, 多様化に従って, データマイニングはとても人手でできるものではなくなりつつあり, 自動化, もしくは半自動化が要求されている. 計算理論の観点からも, 学習理論や論理関数理論, さらにアルゴリズム理論の良い応用場所になっている [6].

顧客データベースで, 一人の顧客のレコードは一つのベクトルになるが, その各要素は, ある商品をその顧客が買ったかどうか (買えば 0 , 買わなければ 1) というような 2 値要素, あるいは, その顧客の総買い上げ高や年齢などの数値要素, 住所などの範疇要素などに大別される. ここで,「新しい顧客にどのような商品のダイレクトメールを送るかを, その顧客の既知データから判定しよう」等と言うのが最もシンプルなデータマイニングである. データマイニングをどのように捉えて, アルゴリズム理論を応用できる形に定式化できるだろうか.

データマイニングの代表的な道具として, Association Rule (結合ルール, または相関ルール) というものがある. 簡単な例は,「子供が 3 歳なら子供和服を買う」(七五三だから) ファストフード店で「ピザを買えばコーラも買う」(このルールから, ピザとコーラのコンビネーションセットを売り出すのが良いことになる)「年収が 350 万円以上, 600 万円以下の男性はカードローンの利用が多い」等と言ったものである.

実際のデータマイニングでは, これらのルールを単独ではなく多数発生させて, 組み合わせることによりシステムを構築する. これらは統計学や学習理論などと共通した考え方である. 例えば, ある属性 C の値 (たとえば, ある顧客のカードローン利用頻度) を, 他の属性値 (年齢, 年収など) から推定するシステムを作るには, 決定木や回帰木といった構造がポピュラーであり, 情報科学を学んだ読者にはお馴染みであろう.

これらの木構造を持ったシステムは, 木の各節点で判定をし, Yes ならば右, No ならば左という決定を行って, いく. そのために, 節点に結合ルールを置き, その仮定部を満たせば右, 満たさなければ左という判定を行うのである. 例えば, 図 1 の木では,「性別」がまず最初に判定され, 男ならば「年収」, 女ならば「年齢」が次に判定されるという仕組みになっている. (念のために, この決定木は実際のデータから構築したものではないことをお断りしておく).

データベースの要素値がすべて 2 値要素であった場合は, これは非常に素直なシステムである. 仮定部が 1 なら右, 0 なら左となる. 目的属性 C に対し, データでのその他の属性の数が k 個で, $\{A_1, A_2, \ldots, A_k\}$ ならば,「属性 A_i が 1 なら目的属性 C も 1 」という

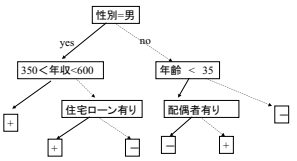

図 1: 決定木

形の結合ルールは k 個しかない. 属性の数は, 属性の絞込みやクラスタリングで, 数十程度に抑えられるのが常なので, これはそれほど大きい数ではない. 「属性 A_i と属性 A_j が共に 1 なら, 目的属性 C も 1」という形の結合ルールや, さらに複雑な結合ルールも考えられ, 研究されている. しかし, 決定木の形に直すと, 上記の結合ルールは高さ 2 の決定木で,「属性 A_i が 1 なら, 目的属性 C も 1」「属性 A_j が 1 なら, 目的属性 C も 1」を用いて書き表せるので, 上記の k 個の結合ルールと決定木を用いて表現できる.

これらの結合ルールを決定木に組み立てて, 木の高さを限定して, 推定率を目的以上にする木を構築する問題は理論的に難しい (NP 完全である) ことが判る. 従って, アルゴリズムとしては, いろいろな実用的な発見手法が考えられており, 例えば, 情報エントロピーを用いた貪欲法による構築 [7] は, 現実的にかなり性能は良い.

さて, 2 値データでなく, 数値データの場合は, さらに新たな問題が生じてくる.「年収が 350 万円以上, 600 万円以下の男性はカードローンの利用が多い」というルールで「年収が 350 万円以上, 600 万円以下」とさらりと書いたが, この 350 万, 600 万という数字の組み合わせを考えると, 結合ルールの候補数は非常に大きくなる. 年収の区切りを, 例えば,50 万円毎に均等に区切って 0 から 2000 万 (以上) までを 40 個の組に分けていれば,「年収」という属性一つで 800 以上の結合ルールが生じてくる. したがって, 何らかの基準で 350,600 などの区切りを前もって求めてから決定木の構築に入らないといけない. これを高速に行うのは一つのアルゴリズム問題である [3].

さらに, 2 値属性の場合は 2 つの属性値から目的を推定するような結合ルールは考えなくても決定木構築にはそれほど支障は無かったのだが, 数値属性の場合は本質的に決定木で表せないような結合ルールがある. 例えば,(現実的でちょっと見るのも嫌になる読者もいらっしゃるかもしれないが) 「住宅ローン残高が年収の 2 倍以上 4 倍以下ならばカードローンの利用が多い」というような結合ルールは,「住宅ローン残高」と「年収」を別々に考えた結合ルールを用いて決定木構造で表すことは難しい. しかも, このように柔軟にルールを考え出すと, 2 つの属性だけでも無数にルールの候補が現れてくる. したがって, 何らかの候補の絞込み手法が必要になってくる.

今の場合, 数値データを考えると,「住宅ローン残高, 年収, カードローン利用数」という 3 つの数値のベクトルである. いわば, 3 次元の点だと思っても良かろう. 一般に, k 個の数値属性を考えると, n 人のデータから, k 次元空間の n 点が生成されると考えていい. 従って, 結合ルールの生成は, 点集合を目的属性値に関して性質の異なる 2 つの

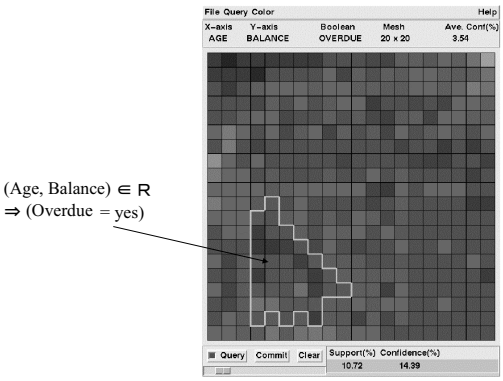

図 2: SONAR の出力

部分 S, S' に分離するという一種の幾何学データ処理 (クラスタリング) である. このとき, 条件としては, 目的属性以外の k-1 個の属性値に対応する k-1 次元部分空間への射影 $Pr(S)$, $Pr(S')$ が交わらないこと (分離性), S の形状がルールとして学習誤差に強い形状をしていること (普遍性) ユーザにとって, 理解のできるルールになっていること (説得性) の 3 点が必要になる.

特に, 上記の 3 次元の場合, 目的属性値の部分を色調で表す (例えばカードローン利用数と赤色の部分の濃さ) と, 3 次元の点集合でなく, 2 次元の色付きの点集合, すなわち, カラーの画像と思うことができる. 画像を色調に関して 2 つの (そのうち少なくとも一つは形状に意味のある) 部分に切り分ける作業は画像処理では「イメージ切り出し」と呼ばれ, いろいろなアルゴリズムが提案されてきている. データマイニングで大切な説得性については, 切り出されたルールが何らかの意味で最適化されており, その最適化基準がユーザにとって納得のいくものであることが望ましい. 例えば, 情報エントロピーを最小化する切り出しなどが判りやすく, 理論的な幾何学処理アルゴリズム (文献 [1]) を用いることにより, 高速に切り出しを行える. つまり, 数値属性を持つデータベースでの結合ルール生成は画像処理でのイメージ切り出し, もしくはその一般化であるクラスタリングとして捉えることができる.

このような考え方を用いて, 著者の関わった SONAR システム (System with Optimized Numeric Association Rules) の設計が行われた. [2, 4]. 例えば, 図 2 は, SONAR での結合ルールの出力例である. 詳細は解説記事 [5, 8] をご覧頂きたい. ただし, そこで与えられた定式化が絶対ではなく, また, 4 つ以上の数値属性の関わる結合ルールは (計算速度の問題点から) 与えられていない. データマイニングでのクラスタリング技術に関連しては数多くの論文が出ているが, 通常のクラスタリングでの要求との差異により, 更なるユニークな手法が求められている.

参考文献

[1] T. Asano, D. Chen, N. Katoh, and T. Tokuyama. Polynomial-time solutions to image segmentations. In *Proc. 7th ACM-SIAM Symposium on Discrete Algorithms*, pp. 104–113, 1996.

[2] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, pp. 13–23, June 1996.

[3] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Mining optimized association rules for numeric attributes. In *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, June 1996.

[4] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, Constructing Efficient Decision Trees by Using Optimized Association Rules. In *Proceedings of the 22nd VLDB Conference*, September 1996.

[5] 福田剛志 森本康彦 森下真一 徳山豪「データマイニングの最新動向– 巨大データからの知識発見技術 –」(解説記事) 情報処理 (1996)

[6] Kazuhisa Makino, *Studies on Positive and Horn Boolean Functions with Applications to Data Analysis*, 京都大学博士論文 (数理工学), March 1997.

[7] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[8] 徳山　豪「データベースからの知識発見に使われる計算幾何学」数理科学 1999 年 7 月号, 11-18 .