3D7-1

概念の分解・再統合に基づいてデータ抽象を求める分枝限定アルゴリズムの提案

A Method for Finding Data Abstractions By Integrating Concept Components

谷口 智彦 Tomohiko TANIGUCHI　原口 誠 Makoto HARAGUCHI　大久保 好章 Yoshiaki OKUBO

北海道大学大学院情報科学研究科コンピュータサイエンス専攻
Division of Computer Science, Graduate School of Information Science and Technology, Hokkaido University

We discuss in this paper a method for finding data abstractions for classification problems. Our abstraction can be obtained from a concepts in a given taxonomy by dividing the concept and then integrating the components. By the operations, we can create new concepts which are not explicitly in the original taxonomy.

1. はじめに

所与の (大規模) データから有用な知識を抽出することを目的とした**データマイニング**の研究が精力的に行なわれているが，そこでの大きな問題点として，抽出されるルール数の膨大さが指摘されて久しい．その大きな要因のひとつとして，データの**記述粒度**がしばしば詳細過ぎることが挙げられる．詳細に記述されたデータに多くの情報が含まれることは事実であるが，そこから抽出されるルールが必要以上に詳細なものとなり，ルールの容易な理解・解釈を妨げることが少なくない．ルールを適切な抽象レベルで表現することで，それらの可読性が高まり，その結果，ユーザにとってより有用なものとなることが期待できる．

こうした問題意識のもと，著者らはこれまで，**決定木** [Quinlan 93] に代表される**分類ルール**の可読性を高めることを目的として，**データ抽象** (*Data Abstraction*) の研究を行なってきた [Haraguchi 02, Kudoh 03]．そこでは，目標クラスの分類ルール抽出のための適切な抽象化を定め，その抽象化のもとでデータベースを汎化した後，汎化データベースから，より簡潔に記述された分類ルールを抽出する．

そこでの抽象化とは，もとのデータベース中の属性値を，より抽象的なものへ置き換える操作を意味する．すなわち，これまで区別していた属性値が同一のものと見做されるため，抽象化は一般に何らかの情報損失を伴う．ここで失われる情報が，目標クラスの分類に不可欠なものであった場合は，もはや汎化後のデータベースから有用な分類ルールが得られることはない．逆に，こうした情報が抽象化後も保存される場合は，汎化データベースからは，有用かつ簡潔な分類ルールの抽出が期待できる．

このアイデアを基礎に，文献 [Haraguchi 02, Kudoh 03] では，目標クラスに関する**適切な抽象化**を，情報理論的尺度を用いて定義し，そのもとでの汎化データベースから目標クラスの分類ルールを抽出する *ITA*(*Information Theoretical Abstraction*) システムを設計・実装し，その有用性を確認した．そこでは特に，与えられた概念階層辞書をもとに，適切さの尺度のもとで最適な辞書登録された概念を，適切なデータ抽象として選択する．これにより，抽象化前の分類精度を出来るだけ保存した簡潔な分類ルールの抽出が可能であることを確かめた．しかし，得られるデータ抽象は辞書登録されたものに限定されていることから，適当な辞書概念がない場合は，十分な分類精度を保つことが出来ない．

こうした問題を緩和すべく，文献 [Okubo 03] では，任意のデータ抽象が成す分割束の一部を辞書中の概念を用いて部分的に切り出し，その部分分割束を探索することで，辞書には陽に登録されていないデータ抽象を求める枠組に拡張した．これにより，分類精度の劣化を抑えながら，より柔軟にデータ抽象を求めることが可能となった．しかし，得られるデータ抽象は，陽に辞書登録はされていないものの，登録概念の統合操作により得られる範疇にあり，その意味で，これまでの制限が十分緩和されたとは言い難い．また，部分分割束を探索するコストも小さくないことから，これらの点を含めてさらなる手法の改良が望まれる．

本稿では，辞書登録された概念の**分解・再統合**によりデータ抽象を求める手法の概略について述べる．特にここでは，所与の目標クラスに関する分類ルールの可読性を高め，かつ，分類精度の向上に寄与するデータ抽象を求める．既存概念の分解・再統合操作により，辞書には陽に現れない**兄弟概念**の抽出が可能となる．

2. 準備

A_i を**属性**とし，その取り得る値の集合，すなわち，**ドメイン**を $dom(A_i)$ と表す．特に，$dom(A_i)$ の各要素を A_i の**属性値**と呼ぶ．

属性 $A_1, \ldots, A_m$ について，属性値の m 組 $t \in A_1 \times \cdots \times A_m$ を，**関係スキーマ** $R(A_1, \ldots, A_m)$ の**インスタンス**，あるいは，**事例**と呼ぶ．事例 t における属性 A_i の属性値を $t[A_i]$ で参照する．

事例の (多重) 集合を，$R(A_i, \ldots, A_m)$ を関係スキーマとする**関係データベース**と呼び，$\mathcal{D}_R$ と表記する．特に混乱がない場合は，単に $\mathcal{D}$ と書く．

関係データベース $\mathcal{D}$ において，属性 A の属性値が a_i であるタプルの数を $frec_{\mathcal{D}}(A = a_i)$ で表す．すなわち，

$$frec_{\mathcal{D}}(A = a_i) = |\ \{\ t \in \mathcal{D} \mid t[A] = a_i\ \}\ |$$

である．

連絡先: 原口 誠・大久保 好章
北海道大学大学院情報科学研究科コンピュータサイエンス専攻
〒 060-0814 札幌市北区北 14 条西 9 丁目
TEL：011-706-7106
E-mail：{ mh, yoshiaki }@ist.hokudai.ac.jp

以下では，属性 A を，$\Pr(A = a_i) = \frac{freq_{\mathcal{D}}(A=a_i)}{|\mathcal{D}|}$ なる確率分布を有する確率変数とみなす．属性 A の確率分布は $(\Pr(A = a_1), \ldots, \Pr(A = a_k))$ なる表記で明示することが出来る．

3. データ抽象に関する情報理論的尺度

関係データベース $\mathcal{D}$ における属性 A の**エントロピ** $H(A)$ は，

$$H(A) = -\sum_{a \in dom(A)} \Pr(A = a) \log_2 \Pr(A = a)$$

で与えられる．$H(A)$ は，$\mathcal{D}$ における属性 A の確率分布の様子を示すものである．$H(A)$ は，確率分布が一様である場合に最大値をとり，分布に偏りがある程，小さな値をとる．

関係スキーマ $R(A_i, \ldots, A_m)$ 中のある属性 C を**クラス** *(***目標***)* **属性**とする．ある属性 A の属性値 a_i に注目し，$A = a_i$ のもとでの C の**事後確率**の分布は，

$$(\Pr(C = c_1|A = a_i), \ldots, \Pr(C = c_m|A = a_i))$$

で与えられる．よって，そのエントロピ，すなわち，$A = a_i$ のもとでの C のエントロピ $H(C|A = a_i)$ は，

$$H(C|A = a_i) = -\sum_{c \in dom(C)} \Pr(C = c|A = a_i) \log_2 \Pr(C = c|A = a_i)$$

となる．この期待値は，

$$H(C|A) = \sum_{a_i \in dom(A)} \Pr(A = a_i) H(C|A = a_i)$$

となり，これを，A のもとでの C の**条件付きエントロピ**と呼ぶ．

$H(C) - H(C|A)$ は**相互情報量** (あるいは**情報利得**) と呼ばれ，しばしば，$I(C; A)$ と表記される．相互情報量 $I(C; A)$ は，A を条件属性とする目標属性に関する分類ルールの性能を示すものであり，値が大きいものほど分類精度が高いことを意味する．なおこの時，条件属性 A のエントロピー $H(A)$ は**分割情報量**と呼ばれる．

以下では，目標属性を C，条件属性を A と仮定して，C に関する分類ルールを抽出する際の属性 A のデータ抽象を考える．

4. 概念の分解・統合に基づくデータ抽象

条件属性 A のデータ抽象を，$dom(A)$ の分割 $\varphi = \{D_1, \ldots, D_k\}$ と同一視する．ここで，$D_i \subseteq dom(A)$ であり，D_i 中の属性値は，このデータ抽象のもとで，すべて同一の抽象属性値となることを意味する．

いま，属性 A のデータ抽象 $\varphi = \{D_1, \ldots, D_k\}$ および $\varphi' = \{D'_1, \ldots, D_\ell\}$ を考える．任意の $D_i \in \varphi$ について，$D_i \subseteq D'_j$ なる $D'_j \in \varphi'$ が存在する時，$\varphi \preceq \varphi'$ なる順序を与える．すなわち，φ' は，φ 中のセルを統合して得られるデータ抽象であることを意味する．

$dom(A)$ 中の各属性値を唯一の要素とするセルから成るデータ抽象を $\varphi_\perp$ とする．また，$dom(A)$ 中の属性値を，所与の概念階層辞書に従って，出来るだけまとめ上げたデータ抽象を $\varphi_\top$ とする．文献 [Okubo 03] に従って抽出可能なデータ抽象とは，$\varphi_\perp \preceq \varphi \preceq \varphi_\top$ なるデータ抽象 φ であり，かつ，φ の各セル中の属性値のもとで C の事後確率分布が類似したものである．つまり，$\varphi_\perp$ から，概念階層に沿ったセルの統合操作を行なって得られるものに限定される．それらは，必ずしも階層中に対応する概念を陽に持つとは限らないが，辞書の定める階層構造から外れるものではない．

ここでは，より柔軟にデータ抽象を作り出す手法を提案する．特に，辞書の定める階層構造を積極的に外れたデータ抽象を抽出することを試み，それをセル (概念) の**分解・統合操作**により実現する．

条件属性 A のドメイン $dom(A)$ に関して，所与の概念階層辞書から定義可能なデータ抽象 (分割) を $\varphi = \{D_1, \ldots, D_k\}$ とする．ここで，各セル D_i を，C の事後確率分布が類似した属性値同士に**分解**し，それを $D_i = \{D_{i_1}, \ldots, D_{i_n}\}$ とする．異なる D_i および D_j 中のセルを**再統合**した結果得られるセルは，辞書には陽に現れない概念に相当し，特に，辞書に規定された階層構造からは外れた新規なものとなる．こうしたセルの分解・再統合操作により，辞書に陽に現れない新規な概念の抽出が可能なだけでなく，分類精度が向上する新たな有用な概念が抽出できる可能性もあることを強調しておく．分解・再統合操作を繰り返すことで，もとの φ からより離れた概念が抽出されるが，辞書登録された概念から過度に離れたものは，ユーザにとって解釈が困難なものとなろう．その意味で，φ の適当な近傍の範囲で分解・再統合操作を行なうことが妥当であると考えられる．

以上より，ここでのデータ抽象問題を次の通り定式化する．

Given: 関係データベース $\mathcal{D}$，目標属性 C, 条件属性 A，分割情報量の上限閾値 ϵ，および，概念階層辞書から定義可能な A のデータ抽象の族 Φ.

Find: A のデータ抽象 φ，ただし，以下の条件を満たすものの中で，目標属性 C との相互情報量が最大であるもの．

1) 分割情報量制約を満たす，すなわち，$H(\varphi) \leq \epsilon$.
2) Φ 中のあるデータ抽象から分解・統合操作で得られ，かつ，その近傍に位置する．
3) Φ 中の任意のデータ抽象と比較不能．

分割情報量の上限閾値 ϵ は，分割が過度に詳細化されることを抑制するための制約であり，これにより探索の枝刈りが可能となる．また，相互情報量の最大化は，出来るだけ分類精度が高いものを求めることを意味する．ただし，最大なものだけでなく，上位 N (Top-N) を求めてもよい．

5. おわりに

本稿では，所与の目標クラスに関する分類ルールの可読性を高め，かつ，分類精度向上に寄与するデータ抽象を求める手法の概略について述べた．これまでの手法 [Okubo 03] では，辞書登録された概念から合成可能な概念のみがデータ抽象として認められたが，ここでは，既存の辞書概念の分解・再統合操作により，辞書には陽に現れない兄弟概念の抽出も可能である．目標クラスの分類において，潜在的に有用な様々な抽象概念が抽出できることから，概念体系の構築や修正等にも有用であると期待している．

現在，計算機実験に向けてシステムを実装中であり，具体的なデータにおける本手法の有効性については，稿を改めて報告したい．

参考文献

[Quinlan 93] J. R. Quinlan: C4.5 - Programs for Machine Learning, Morgan Kaufmann, 1993.

[Tomita 03] E. Tomita and T. Seki: An Efficient Branch-and-Bound Algorithm for Finding a Maximum Clique, Proceedings of the 4th International Conference on Discrete Mathematics and Theoretical Computer Science - DMTCS'03, Springer-LNCS 2731, pp. 278 - 289, 2003.

[Kudoh 03] Y. Kudoh, M. Haraguchi and Y. Okubo: Data Abstractions for Decision Tree Induction, Theoretical Computer Science, 292(2), pp. 387 - 416, Elsevier, 2003.

[Haraguchi 06a] M. Haraguchi and Y. Okubo: A Method for Pinpoint Clustering of Web Pages with Pseudo-Clique Search, Federation over the Web, International Workshop, Dagstuhl Castle, Germany, May 1 - 6, 2005, Revised Selected Papers, Springer-LNAI 3847, pp. 59 - 78, 2006.

[Haraguchi 02] M. Haraguchi and Y. Kudoh: Some Criterions for Selecting the Best Data Abstractions, Progress in Discovery Science, Springer-LNCS 2281, pp. 156 - 177, 2002.

[Okubo 03] Y. Okubo, Y. Kudoh and M. Haraguchi: Constructing Appropriate Data Abstractions for Mining Classification Knowledge, Web-Knowledge Management and Decision Support - The 14th International Conference on Applications of Prolog, Revised Papers, Springer-LNAI 2543, pp. 276 - 289, 2003.