1F5-02

分類ルール獲得のための抽象化部品探索

Creating Abstract Concepts in Classification Rule Mining

大久保 好章　　原口 誠
Yoshiaki OKUBO　Makoto HARAGUCHI

北海道大学 大学院工学研究科 電子情報工学専攻
Division of Electronics and Information Engineering, Hokkaido University

In this paper, we present a method of creating abstract concepts for classification rule mining. For a target attribute, we try to find abstract concepts that are useful for the classification in the sense that assuming such a concept can well discriminate a target class and support data as much as possible. Our task is formalized as an optimization problem in which its constraint and objective function are given by *entropy and probability of class distributions*, respectively. Concepts to be found can be stated in terms of *maximal weighted cliques*, where each node is a class distribution and its weight is defined the probability. Nodes are connected based on a *closeness* relationship among nodes, where the closeness is defined *without* any explicit distance parameter. From the graph, as useful abstract concepts, top N maximal weighted cliques are efficiently extracted with two pruning techniques: *branch-and-bound* and *entropy-based* pruning. Our preliminary experimental results show that useful concepts can be created in our framework.

1. はじめに

本稿では，分類ルール [Quinlan 93] 獲得における抽象概念の生成手法について考察する．

データ抽象は，獲得される分類ルールの可読性向上のために有効であることが報告されている [Kudoh 02]．そこでのデータ抽象とは，分類ルールの条件属性値のクラスタ群 (属性ドメインの分割としての抽象概念群) を求める操作であり，分類精度を大きく劣化させることなく，かつ，できるだけ大きなクラスタ群を得る枠組であった．文献 [Kudoh 02] では，得られる抽象概念は，所与の概念辞書に強く依存し，ユーザが望む概念が辞書に適切に書かれていない場合は，それらを得ることが不可能であった．こうした点をより柔軟にすべく，[Okubo 03] では，辞書に制約された範囲で，抽象概念の探索を行う手法が提案されたが，計算量が膨大であり，十分な抽象概念を得るには，さらなる改善が必要であることがわかった．また，従来手法では，分類に寄与しない抽象概念も含めて概念群の評価がなされている，より有効な抽象概念が存在する場合もある．

本稿では，これら先行研究における問題点の解消を試みる．具体的には，概念辞書を用いずに，有用な抽象化部品 (抽象概念) を効率良く生成する手法を提案する．これまで，抽象概念は属性ドメインの分割として求められていたため，分類に寄与しない抽象概念も結果的に生成されていたことになる．本稿では，属性ドメインから分類に寄与するクラスタ (抽象化部品) のみを抽出する．こうしたクラスタ抽出問題は，重み付き最適クリークの抽出問題として定式化されている [原口 02] が，分類に寄与する抽象概念は，最適なものひとつに限らないことから，本稿では上位 N のクラスタを抽出する．

2. 準備

本稿では**離散属性**を扱うものとし，属性 A の取り得る値の集合，すなわち A **のドメイン**を $dom(A)$ で表す．

連絡先: 大久保 好章
北海道大学 大学院工学研究科 電子情報工学専攻，〒 060-8628 札幌市北区北 13 条西 8 丁目, TEL：011-706-7161,
E-mail: yoshiaki@db-ei.eng.hokudai.ac.jp

R を**関係名**，$A_1, \ldots A_m$ を属性とした時，$R(A_1, \ldots A_m)$ を**関係スキーマ**と呼ぶ．**関係** R を $R \subseteq dom(A_1) \times \cdots \times dom(A_m)$ と定義する．タプル $t = (a_1, \ldots, a_m) \in R$ について，その i 番目の要素 a_i を t の A_i-**値**と呼び，$t[A_i]$ で参照する．

関係スキーマを $R(A_1, \ldots A_m)$ とする関係 R について，A_i 以外のある属性 C を**目標属性**とし，R 中の各タプルは C-**値**を持つものと仮定する．ここで，$dom(C)$ の要素は**クラス**と呼ばれる．R のタプル t について，$t[C] = c$ である時，t **のクラスは** c **である**という．

本稿では，目標属性 C に関する分類ルールを扱う．特に，属性 A_i を条件属性とした時，$(A_i = a) \rightarrow (C = c)$ なる形式の分類ルールについて考える．なお，以下では簡単のため条件属性がひとつの場合について議論するが，複数の場合にも同様な議論を展開できる．

タプル $t \in R$ の生起確率 $\Pr(t)$ を $\Pr(t) = 1/|R|$ と定める．すなわち，R の関係スキーマ中の各属性 X は以下の確率変数と捉えられる．

$$\begin{aligned} \Pr(X = x) &= \Pr(\{\ t \mid t \in R \ \wedge\ t[X] = x\ \}) \\ &= |\ \{t \mid t \in R \ \wedge\ t[X] = x\}\ |\ /\ |\ R\ |. \end{aligned}$$

$0 \leq p_i \leq 1$ および $\Sigma_{i=1}^{n} p_i = 1$ を満たすベクトル $(p_1, \ldots, p_n)$ を**確率分布** と呼ぶ．

$dom(C) = \{\ c_1, \ldots, c_n\ \}$ なる目標属性 C と属性 A を考える．$a \in dom(A)$ について，以下で与えられる条件付き確率の分布を，a **のもとでの** C **の** (**事後**) **クラス分布**と呼ぶ．

$$\mathcal{D}_a^C = (\ \Pr(C = c_1 \mid A = a), \ldots, \Pr(C = c_n \mid A = a)\ ).$$

同様に，$G \subseteq dom(A)$，すなわち A 中のいくつかの属性値集合 (以下，クラスターと呼ぶ) についても，G のもとでの C のクラス分布を定義する．

$$\mathcal{D}_G^C = (\ \Pr(C = c_1 \mid A \in G), \ldots, \Pr(C = c_n \mid A \in G)\ ).$$

クラス分布は，C に関する分類問題を考える際，a あるいは G での条件付けによるクラス識別性を示しており，偏りのある分布であればあるほど，その条件付けが，あるクラスの識別に寄与することを意味する．

3. 重み付きクリーク探索による抽象化部品生成

C を $dom(C) = \{c_1, \ldots, c_n\}$ からなる目標属性，A を C に関する分類ルールの条件属性とする．本章では，C の分類に寄与するクラスター $G \subseteq dom(A)$ の生成を，**重み付きクリークの探索**により実現する．

3.1 クラス分布のエントロピー

先に述べた通り，あるクラスター $G \subseteq dom(A)$ のクラス識別性は，クラス分布 $\mathcal{D}_G^C$ により示唆され，分布が偏っているほど，G による条件付けが，クラスの識別に有効であることを意味する．すなわち，クラスの識別性能を測る尺度として**クラス分布のエントロピー**を考えることができる．

クラス分布 $\mathcal{D}_G^C$ のエントロピーを $H(\mathcal{D}_G^C)$ とし，以下で定める．

$$H(\mathcal{D}_G^C) = \sum_{a \in G} \frac{\Pr(a)}{\Pr(G)} H(\mathcal{D}_a^C).$$

ここで，

$$H(\mathcal{D}_a^C) = -\sum_{j=1}^{n} \Pr(C = c_j | A = a) \log_2 \Pr(C = c_j | A = a)$$

である．

クラス分布に偏りがあればある程，エントロピー値は小さくなり，そうした分布を与えるクラスタは，クラス識別性能が高いと言える．以下では，クラス識別性の高いクラスタが，分類問題における有効な抽象概念に成り得ると考え，こうしたクラスタの探索手法について考察する．

3.2 最適クラスタ

クラス識別性の低い分類ルールは役に立たないことから，ここで求めるべきクラスタには，クラス分布のエントロピーに関する制約を課す．すなわち，エントロピー値がある閾値より小さな分布を与えるクラスタのみを考える．

さらに，一般には，サポートの低い (適用範囲の狭い) 分類ルールも有用であるとは言い難い．よって，クラス識別性が高いことのみならず，できるだけ多くのデータに対して適用可能なことも要請する．

以上の観点から，本稿でのクラスタ生成を次の最適化問題として定式化する．

エントロピー制約:

$$H(\mathcal{D}_G^C) \leq \delta$$

目的関数 (最大化):

$$\Pr(G) = \sum_{a \in G} \Pr(A = a)$$

.

ここで，δ はクラス識別性をコントロールする入力パラメータである．この問題の解，すなわち，エントロピー制約を満たし，かつ，できるだけ多くのデータを支持するクラスタを**最適クラスタ**と呼ぶ．

十分な識別性があり，かつ，ある程度のデータを支持するクラスタは，例え上記の意味で最適でなくとも，分類問題を考える上では有用なクラスタであると言えよう．よって，本稿では，最適クラスターを含む**上位 N クラスタ**を求め，それらを有用な抽象化部品と考えることにする．

3.3 クラス分布の近さ

抽象化部品の生成は，ボトムアップ，すなわち，要素 1 のクラスタを順次統合し，より大きなクラスタを生成する戦略で行われる．その際，クラス分布のエントロピーに，クラスタの拡張に伴う単調性があれば，効率の良い探索が期待できるが，それは保証されない．効率良い探索を実現するためには拡張処理をうまく制御してやる必要がある．

類似した分布を与えるクラスタ同士を結合すると，その拡張クラスタは，もとの分布と近い分布を与えることから，**分布間の類似性**に基づいてクラスタの結合処理を制御することが考えられる．類似性を測るためには，**分布間の距離**を与えることが一般的であるが，適切な距離関数を与えること，さらに，そのもとで類似の度合を判断する適切なパラメータを与えることは容易でない．そこで本研究では，この様な明示的な距離を用いない**分布の近さ**を採用している [原口 02]．具体的には，エントロピー関数の凸性に基づく位置関係で定まる分布間の近さを考える[*1]．

エントロピー条件を満たすふたつのクラス分布 p_1 と p_2 を考える．

$$p_1 = \mathcal{D}_{G_1}^C = \{x_1, y_1, (1 - x_1 - y_1)\}, \quad H(p_1) \leq \delta$$
$$p_2 = \mathcal{D}_{G_2}^C = \{x_2, y_2, (1 - x_2 - y_2)\}. \quad H(p_2) \leq \delta$$

ここで，エントロピー値が δ となる分布の集合，すなわちエントロピーの等高線は以下の曲線で表わされる．

$$\begin{aligned} f(x, y) \quad = \quad & -x \log_2 x - y \log_2 y \\ & -(1 - x - y) \log_2 (1 - x - y) - \delta = 0 \end{aligned}$$

各 p_i からこの等高線に降ろした垂線の足の座標をそれぞれ (x_i^*, y_i^*) とすると，$f(x, y)$ に対する (x_i^*, y_i^*) における接線は，

$$g_i(x, y) = f_x(x_i^*, y_i^*)(x - x_i^*) + f_y(x_i^*, y_i^*)(y - y_i^*) = 0.$$

で与えられる．ここで，f_x (f_y) は，f の x (y) に関する偏微分である．以上の準備のもと，クラス分布 p_1 and p_2 の近さを以下の通り定義する．

定義 3..1 次のいずれかが成り立つ時，p_1 と p_2 は近いと言われる．

- $g_1(x_1, y_1) \times g_1(x_2, y_2) \geq 0$,
- $g_2(x_1, y_1) \times g_2(x_2, y_2) \geq 0$

簡単に述べると，p_1 あるいは p_2 から降ろした垂線の足における接線[*2]を境界として，両者が同じ側に位置する場合には近いと考える．図 1 はこうした分布の位置関係を図示したものである．左図における p_1 と p_2 は近いが，右図におけるそれらは近いとは判断されない．

この分布間の近さ概念に基づき以下の定理を得る．

定理 3..1 $G = \{a_1, \ldots, a_m\}$ を以下を満足するクラスタとする．$H(\mathcal{D}_{a_i}^C) \leq \delta$ $(1 \leq i \leq m)$ かつ，任意の i と j $(i \neq j)$ について，$\mathcal{D}_{a_i}^C$ と $\mathcal{D}_{a_j}^C$ は近い．

この時，$H(\mathcal{D}_G^C) \leq \delta$ である．

*1 以下では簡単のために，目標属性のクラス数が 3 の場合について考えるが，4 以上の場合についても同様に議論可能である．

*2 クラス数が 4 以上の場合は，接 (超) 平面となる．

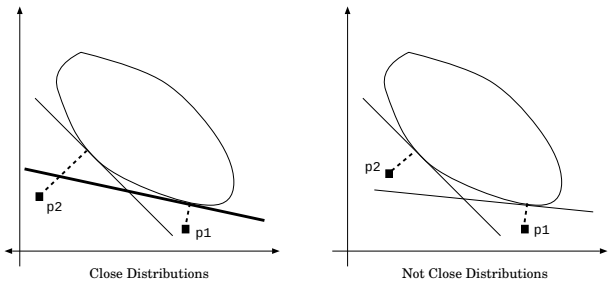

図 1: クラス分布間の近さ

つまり，エントロピー制約を満たし，かつ，互いに近い分布を与える属性値からなるクラスタは，必ずエントロピー制約を満たす分布を与える．

3.4 トップ N 重み付き極大クリークの探索

これまでの議論に基づき，上位 N の抽象化部品の生成を，**重み付グラフにおけるクリーク探索**として定式化する．

今，各条件属性値に対して，それらが与えるクラス分布をノード，その生起確率をノードの重みとするグラフを考える．ここで，エントロピー制約を満たし，かつ，近い関係にある分布間にはエッジを張ることにしよう．すると，このグラフにおけるクリーク (完全部分グラフ) は，定理 3..1 の条件を満たす分布から構成されることになる．

また，文献 [Kudoh 02] で報告されている様に，ある分布を与えるクラスタが例え十分なクラス識別能力を有していなくとも，その生起確率が相対的に低い場合は，それを例外的なものとして生起確率の大きな識別能力の高いクラスタに統合することで，クラス識別能力を保ったまま，クラスタの生起確率をあげることが可能となる．先に述べた通り，本稿で求めるべきクラスタの生起確率はできるだけ高いことが要請される．よって，エントロピー制約を満たさない分布については，無条件に任意の分布とエッジを張ることとし，エントロピ制約が満たされる限りは，上記クリークに取り込み，クラスタを拡張する．

以上まとめると，ここでの抽象化部品生成は，重み付きグラフ $\mathcal{G}$ におけるクリーク探索問題として定式化できる．

入力： 生成する抽象化部品の数 N，クラス識別性パラメータ δ，重み付き (無向) グラフ $\mathcal{G}$，ここで，

> ノード：条件属性値の与えるクラス分布．
> ノードの重み：対応する属性値の生起確率．
> エッジ：エントロピー制約を満たす近いノード間，および，エントロピー制約を満たさないノードと任意のノード間に存在．

出力： エントロピー条件を満たす (極大) クリークの中で生起確率が上位 N のもの．

以上のクリークは，重み付き最大クリーク抽出アルゴリズム $MWCC$ [若井 98] を改良して求めることが可能である．$MWCC$ は**分枝限定-深さ優先**探索に基づくアルゴリズムである．そこでの分枝限定は，クリークの重みに関して行われ，本研究でのクリーク探索においても同様に利用できる．ここでは，それに加えて，エントロピー制約に基づく探索の枝刈りが可能となっている．具体的には，クリークを拡張する際，分布のエントロピー値の昇順にノードを付加することで，制約を満たさなくなった時点で探索を打ち切ることが可能となる．

4. 実験

本章では，計算機実験の結果について簡単に述べる．

システムを C 言語で実装し，AT&T PC 互換機 (PentiumIII-1.2MHz, 512MB memory) 上で実験を行った．実験には，*The UCI ML Repository* [Marphy 94] の *Adult Database* を用いた (タプル数 5908, 属性数 15)．目標属性を *Workclass*(クラス数 8)，条件属性を *Education*(属性数) および *Sex*(属性数 2) とした場合の結果を以下に示す．なお，所与のパラメータは $N = 5$ および $\delta = 1.0$ とした．

Education に関して，以下の 3 つのクラスターが得られた．

> [`Preschool`, `1st-4th`, `5th-6th`, `9th`, `11th`, `12th`]
>
> [`Preschool`, `1st-4th`, `5th-6th`, `9th`, `10th`, `12th`]
>
> [`Preschool`, `1st-4th`, `5th-6th`, `9th`, `12th`, `Prof-school`]

いずれも，主に初等教育に関する属性値から成っており，`Self-emp-inc` を識別するには，これらをひとつにまとめた抽象概念 (抽象化部品) が有効であろうことを示唆している．属性 *Education* には高等教育に関する属性値も実際に含まれているが，これらはクラスタに取り込まれなかった．このことから，本手法により意味のある抽象化部品が生成されたと考えられるだろう．

5. おわりに

本稿では，分類ルール獲得のための抽象化部品生成手法について議論した．具体的には，あるクラスの識別能力が高いクラスタを意味のある抽象概念と考え，こうしたクラスタをクリーク探索により求める手法を考察した．

本稿での手法は，抽象概念を生成する際に，概念辞書を一切用いない点が大きな特徴である．一般に，辞書は多くの人にとって役立つ概念を記述する有用なものであり，データマイニングや情報検索においても重要な役割を果たしている．半面，特定の用途や目的を持つユーザにとっては不十分と感じることもしばしばあり，この様な場合は，辞書の修正が不可欠である．本稿での手法が，こうした辞書の不備を修正するための新概念 (抽象化部品) 生成の基礎技術と成り得ることを期待し，この観点からさらに考察を進めている．

参考文献

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

[原口 02] 原口 誠: “最適クリーク探索に基づくデータからの概念学習”, 第 50 回人工知能基礎論研究会資料，SIG-FAI-A202-11，pp. 63-66, 2002.

[Kudoh 02] Y. Kudoh: “A Study on Appropriate Abstraction for Data Mining”, Doctoral Dissertation, Division of Electronics and Information Engineering, Hokkaido University, 2002.

[若井 98] 若井 康，富田 悦次，若月 光夫: “最大重みクリーク抽出アルゴリズムの効率化”, 人工知能学会全国大会 (第 12 回) 論文集，pp. 250 - 252, 1998.

[Okubo 03] Y. Okubo, Y. Kudoh and M. Haraguchi: "Constructing Appropriate Data Abstractions for Mining Classification Knowledge", Web-Knowledge Management and Decision Support - 14th Int'l Conf. on Applications of Prolog, Revised Papers, Springer LNAI 2543, pp , 2003.

[Marphy 94] P. M. Marphy and D. W. Aha: "UCI Repository of machine learning databases", http://www.ics.uci.edu/mlearn/MLRepository.html, Univ. of California, Dept. of Information and Computer Science, 1994.