

不均一なデータに対する階層的分類

葛西正裕[†]
九州大学大学院経済学府[†]

古川哲也[‡]
九州大学大学院経済学研究院[‡]

1 はじめに

情報技術の発展とネットワーク環境の整備によって流通するデータ量は著しく増加した。同時にデータは数値、テキスト、画像、音声といった多様なものとなっている。蓄えられたデータを効率的に利用するためには、それらを適切に構成する必要がある。データのセマンティクスに基づく階層的な分類はデータの構成法として有用である^[1]。データ分類に関する研究は多くなされているが、データの意味は1つであり意味のレベルが一樣である均一なデータを想定する場合が多い。データは通常、不均一であり、そのようなデータを対象とする分類の議論は知られていない。本稿では、複数の意味を持つ不均一なデータに対する分類階層を提案する。

2 分類階層

対象となるデータをオブジェクト、オブジェクトの分類によって作られるオブジェクト集合をクラスとし、クラス C の要素を $m(C)$ で表す。クラス集合 C に対して、 $m(C) = \bigcup_{C \in C} m(C)$ とする。 $C = \{C_1, C_2, \dots, C_n\}$ が C の分類によるクラス集合であるとき、 C は C_i の親クラスであり、 C_i は C の子クラスである ($1 \leq i \leq n$) という。

オブジェクトがクラスの要素になるかどうかはオブジェクトとクラスの意味によって決定される。 $S(o)$ と $S(C)$ で o と C の意味を表す。2つの意味 S_i と S_j に対して、 $S_i \prec S_j$ で S_j は S_i の上位概念であることを表す。クラス C とオブジェクト o に対して、 $S(o) \preceq S(C)$ ならば、 o は C の要素である。よって $S(C_i) \prec S(C_j)$ ならば $m(C_i) \subseteq m(C_j)$ であり、分類階層において C_i は C_j の子孫になる。

分類階層に関する多くの研究は均一なオブジェクトを対象としており、一般に以下の性質が仮定される。

- 充足性：親クラスのオブジェクトは必ずいずれかの子クラスに属する。
- 排他性：親クラスのオブジェクトは最大1つの子クラスに属する。

$S(C)$ が、 $S(C)$ を被覆する集合 $S = \{S_1, S_2, \dots, S_n\}$ ($S_i \prec S(C)$) に分割され、 C の子クラス集合 $C = \{C_1, C_2, \dots, C_n\}$ が $S(C_i) = S_i$ ($1 \leq i \leq n$) であるとする。クラス C における任意のオブジェクト o に対して、 $S(o) \preceq S(C_i)$ となるような子クラス C_i が必ず存在し、 o はそのクラスに属するので充足性を満たす。

[定義1] クラス C の基本クラスは、 C の子孫クラスにおける終端クラスであり、その集合を $B(C)$ で表す。 □

[命題1] 分類階層が充足性を満たせば、クラス C に対して、 $m(C) = m(B(C))$ である。 □

分類階層における非終端クラスは、そのクラスのオブジェクトを記憶する必要がなく、オブジェクトの挿入や削除といった更新が効率的に行える。

3 意味のレベルが異なるオブジェクト

$S(o_i) \prec S(o_j)$ となる o_i や o_j のような様々な意味のレベルを持つ不均一なオブジェクトに対する分類階層は、子クラスの意味がその親クラスの意味を被覆していても充足性を満たさない。また、不均一なオブジェクトに対する分類階層では“クラスのオブジェクト”の意味には2つある。

- 分類オブジェクト：クラス C に分類されるオブジェクト
- 固有オブジェクト：クラス C の意味と同じ意味を持つオブジェクト

クラス C の分類オブジェクトを $m_C(C)$ で表し、 C の固有オブジェクトを $m_P(C)$ で表す。 $m_C(C)$ は、 $m(C)$ とされていたものである。不均一なオブジェクトにより分類階層は充足性を満たさなくなるが、クラスのオブジェクトは他のクラスの和集合で求めることができる。

[定義2] $m_C(C)$ におけるオブジェクト o は、 $S(o) \preceq S(C_i)$ となる子クラス C_i が存在しないならば、 C の不分類オブジェクトである。 C の不分類クラス C_U は、 C の不分類オブジェクトを要素とする仮想クラスである。 □

[定義3] クラス C の分類オブジェクトに対する基本クラスは、 C と C の子孫クラスにおける不分類クラスであり、その集合を $B_C(C)$ で表す。 □

[命題2] クラス C に対して、 $m_C(C) = m(B_C(C))$ である。 □

クラス C の意味と同じ意味のオブジェクトは、 C_U に含まれている。

[定義4] クラス C に対して、 $m_C(C)$ におけるオブジェクト o は、 $S(o) = S(C)$ ならば C の固有オブジェクトである。 C の固有クラス C_U^P は、 C の固有オブジェクトを要素とする仮想クラスである。 □

固有クラスを用いることで、 $m_P(C)$ は C_U^P より直接求めることができる。

Hierarchical Classification of Heterogeneous Data
[†] Masahiro Kuzunishi, Graduate school of Economics, Kyushu University
[‡] Tetsuya Furukawa, Faculty of Economics, Kyushu University

4 複数の意味を持つオブジェクト

意味 S_i と S_j ($i \neq j$) に対して、 $S \prec S_i$ かつ $S \prec S_j$ となるような意味 S が存在しないとき、 S_i と S_j は意味的に排他である。分類階層における各クラス C とその子クラス C に対して、任意の $S(C_i)$ と $S(C_j)$ ($C_i, C_j \in C, i \neq j$) が意味的に排他であれば、分類階層は排他性を満たす。同時にその分類階層では、2つ以上の親クラスを持つクラスは存在しないので、階層は木になる。

[命題 3] 分類階層が排他性を満たせば、クラス C の基本クラスにおける任意のクラス C_i, C_j ($i \neq j, C_i, C_j \in B_C(C)$) に対して $m(C_i) \cap m(C_j) = \phi$ である。□

不均一なオブジェクトには複数の意味を持つものがあり $S(C)$ は意味の集合となる。例えば、アメリカと日本の経済比較に関する報告書の意味は {アメリカ, 日本} である。1つの意味 S と複数の意味 $S = \{S_1, S_2, \dots, S_m\}$ に対して、すべての S_i ($1 \leq i \leq m$) が $S_i \preceq S$ となると、 $S \prec S$ である。

複数の意味のオブジェクトに対応する複数の意味のクラスを作れば分類階層は排他性を満たすが、そのような階層ではクラスの数が増えすぎて現実的ではない。本稿では、クラスの意味は1つとし、オブジェクトがクラスの要素となることの定義を拡張する。

[定義 5] オブジェクト o とクラス C に対して、 $S_i \preceq S(C)$ となる S_i が $S(o)$ に存在するとき、 o は C の要素である。□

分類階層は排他性を満たさず、オブジェクトが複数の子クラスに分類されることがある。オブジェクトが重複分類される分類階層は構造が簡単であり複雑とはならないが、次のような問題が生じる。オブジェクトがクラス C の子孫の兄弟クラスで重複分類された場合、 $B_C(C)$ において $m(C_i) \cap m(C_j) \neq \phi$ となるクラス C_i, C_j が存在する。また、クラスの要素となる条件が緩和されたので、クラス C には $S(o) \preceq S(C)$ とはならないオブジェクト o が含まれる。

5 重複分類を許した分類階層

分類階層におけるクラス C の深さを、根クラスから C までの経路の長さとし、 $d(C)$ で表す。 C のオブジェクトが1つの子クラスに分類されたとき、 C の子クラスではオブジェクトに重複はない。オブジェクトが2つ以上の子クラスに分類されたとき、子クラスにおけるオブジェクトから1つを選び、 C のオブジェクトとして代表させる。クラス C のオブジェクト o について、 $l(o, C)$ を o が選ばれなかった最後のクラスの深さとする。

- 根クラス C のオブジェクト o に対し $l(o, C) = 0$.
- クラス C のオブジェクト o と C の子クラス C_i に対して、 o が C における o を代表するとき $l(o, C_i) = l(o, C)$ 、そうでないとき $l(o, C_i) = d(C_i)$.

[定義 6] クラス C とクラス集合 C について、 C に対する C の代表オブジェクトは、 $rep(C, C) = \{o \mid o \in m(C), l(o, C_i) \leq d(C), C_i \in C\}$ である。□

[命題 4] クラス C に対して、 $m_C(C) = rep(C, B_C(C))$ である。□

[命題 5] クラス C の基本クラス C_i と C_j ($i \neq j, C_i, C_j \in B_C(C)$) に対して、 $rep(C, \{C_i\}) \cap rep(C, \{C_j\}) = \phi$ である。□

[定理 1] クラス C に対して、 $m_C(C)$ は $rep(C, C_i)$ ($C_i \in B_C(C)$) の直和に等しい。□

クラス C のオブジェクト o は、 $S(o) \preceq S(C)$ となると C の通常オブジェクトであり、そうでないとき C の追加オブジェクトであるという。 $m_N(C)$ 、 $m_S(C)$ をそれぞれ C の通常オブジェクトの集合、追加オブジェクトの集合とする。 C の追加オブジェクト o は、 $S_k \preceq S(C)$ かつ $S_l \not\preceq S(C)$ となる S_k, S_l ($k \neq l$) が $S(o)$ に存在するようなオブジェクトである。2種類の分類オブジェクトは区別できる必要がある。

クラス C のオブジェクト o について、 $l(o, C)$ をオブジェクトが重複している最初のクラスの深さとする。

- 根クラス C のオブジェクト o に対し $f(o, C) = null$.
- クラス C のオブジェクト o と C の子クラス C_i に対して、 o が C_i 以外の兄弟クラスに分類されないとき $f(o, C_i) = f(o, C)$ 、そうでないとき

$$f(o, C_i) = \begin{cases} f(o, C) & (f(o, C) \neq null) \\ d(C_i) & (f(o, C) = null) \end{cases}$$

[定義 7] クラス C とクラス集合 C について、 C に対する C の通常オブジェクトは、 $nat(C, C) = \{o \mid o \in m(C), f(o, C_i) > d(C) \text{ or } f(o, C_i) = null, C_i \in C\}$ であり、 C に対する C の追加オブジェクトは、 $sup(C, C) = \{o \mid o \in m(C), f(o, C_i) \leq d(C), C_i \in C\}$ である。□

[命題 6] クラス C に対して、 $m_N(C) = nat(C, B_C(C))$ であり、 $m_S(C) = sup(C, B_C(C))$ である。□

[命題 7] クラス C と $rep(C, B_C(C))$ におけるオブジェクト o に対して、 $o \in nat(C, B_C(C)) \cup sup(C, B_C(C))$ である。□

[定理 2] クラス C に対して、 $m_N(C)$ は $nat(C, rep(C, C_i))$ ($C_i \in B_C(C)$) の直和に等しく、 $m_S(C)$ は $sup(C, rep(C, C_i))$ ($C_i \in B_C(C)$) の直和に等しい。□

6 むすびに

本稿は、不均一なオブジェクトに対する分類階層を提案した。クラスの分類オブジェクトは、そのクラスの通常オブジェクトと追加オブジェクトに区分され、各々の基本クラスの代表オブジェクトの直和によって求まる。提案した分類階層により、不均一なデータを効率的に分類することが可能になる。

謝辞 本研究の一部は文部科学省科学研究費補助金基盤研究 (C) (2) (課題番号 15500072) の支援を受けている。

参考文献

- [1] Ke Wang, Senqiang Zhou, and Shiang Chen Liew, "Building Hierarchical Classifiers Using Class Proximity," *Proc. 25th Int'l Conf. on Very Large Data Bases (VLDB'99)*, pp. 363-374, 1999.