テクニカルノート

事例間の非類似度を用いたデータベースからのクラス間関係の獲得

森中 雄[†] 大原剛三[††]
馬場口登[††] 北橋忠宏[††]

本稿では，データベース中に複数のクラスが与えられる場合に，それらの間のクラス間関係を獲得する手法を提案する．提案手法では，クラスごとの属性値の発生確率を反映した事例間の非類似度を定義する．これを用いて，他クラスの事例が，着目クラスに属するかどうかを判定し，この結果からそれぞれのクラス間関係を獲得する．

Acquiring Relations between Two Classes Based on Dissimilarity between Instances

Yuu Morinaka,[†] Kouzou Ohara,[††] Noboru Babaguchi[††] and Tadahiro Kitahashi[††]

The purpose of this study is to acquire the relation between two classes, from the database that includes plural classes. In our method, we calculate the probability of each attribute value occurrence for an individual class, and based on this probability, we define the dissimilarity between two instances belong to the same class. If a target class is given, we investigate whether the instance in another class belongs to the target class, using this dissimilarity. The relation between the target class and another class is discovered from the number of instances in another class to which the target class belongs.

1. はじめに

データベース（DataBase: DB）中の事例は，あらかじめ何らかの基準により複数のクラスに分類されている場合がある．また，ある属性に着目したとき，各属性値を持つクラスに分類することが可能である．これらのクラスは，クラス間に共有される事例が明示されない限り，お互いに相容れないと判断し，クラスの相互排反性を仮定するのが通常である．しかし，DB中の各項目は，入力者の個人差により，同じ意味を持つ項目が別のラベルを付けられる，あるいは相互排反ではないクラス名が記入される場合がある[1)]．このような場合，これらの関係を事前に予測することが重要な意味を持つと考えられる．また，事例をクラスに正しく分類する問題においても，クラス間の関係が事前に予測できれば，より分類精度の高い知識の獲得，あるいは分類効率の向上等も期待できる[2)]．

そこで，本稿では，離散値属性のみから構成される複数のクラスを含むDBを対象に，事例間の類似性をもとに，与えられたクラス間の関係を獲得する手法を提案する[3)]．提案手法では，同一のクラスに属する事例は，性質が類似しているが，その類似性には限界が存在するという仮定のもと，この限界を数量化するために，事例間の非類似度を定義する．具体的には，クラスを特徴付けるために，各々の属性値の発生確率を基準に属性値の評価値をクラスごとに定義し，非類似度をクラスごとの評価値を用いた基準で算出する．この非類似度を用いて，他クラスの事例について，クラス内のいずれかの事例との非類似性がある程度以下であるならその事例は，そのクラスに共有されると判断する．クラス間関係は，着目する2クラスについて，一方に属する事例を他方が共有するか否かで獲得される．実験では，複数のクラスを含む2つのDBを用いて提案手法の信頼性を検証する．

2. 対象データベースとクラス間関係

本稿では，離散値属性のみからなるデータベースを対象とし，DB中の1レコードは，

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† 大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University

表 1 動物に関する DB の一例
Table 1 A part of a database about an animal.

事例名	クラス名	milk	cover	gills	egg	leg	water	fly
horse	mammal	1	hair	0	0	4	0	0
dog	mammal	1	hair	0	0	4	0	0
bat	mammal	1	none	0	0	2	0	1
platypus	mammal	1	feather	0	1	2	1	0
dolphin	mammal	1	none	0	0	0	1	0
cow	mammal	1	hair	0	0	4	0	0
eagle	bird	0	feather	0	1	2	0	1
penguin	bird	0	feather	0	1	2	1	0
Yuu	human	1	hair	0	0	4	0	0

$$I, C, v_1, \ldots, v_n$$

という構成をとる．ここで I は事例名，C はクラス名，$v_1, \ldots, v_n$ は属性 $P_1, \ldots, P_n$ の属性値である．また，与えられる DB には，2 つ以上の異なるクラスが存在するものとする．本稿では，与えられる複数のクラスに対して，それらの間の 2 クラス間関係を，以下に定義するうちのいずれかに同定するものとする．

定義 1（クラス間関係） 異なるクラス A, B は以下のいずれかの関係にあるものとする．

包含 B(A) 中の全事例が A(B) に共有される．

排反 B(A) 中のいずれの事例も A(B) に共有されない．

一部共有 B(A) 中の事例の一部が A(B) に共有される． □

例として “milk[1,0]”，“cover[hair,none,feather]”，“gills[1,0]”，“egg[1,0]”，“leg[4,2,0]”，“water[1,0]”，“fly[1,0]”（[ ] 内はとりうる属性値）という属性を持ち，複数のクラスを含む動物に関する DB の一例を**表 1** に示す．表 1 の場合，存在する 3 つのクラス『mammal』，『bird』，『human』のそれぞれの 2 クラス間の関係を獲得することが目標となる．

3. 事例間の非類似度

2 つのクラス A，B 間の関係を獲得するためには，B 中の事例が A に属するかどうか，および，その逆を予測する必要がある．本稿では議論の都合上，まず，2 事例間の非類似度を両者の間で互いに属性値が異なるものの個数を基準に定義する．さらに，同一クラスに属する事例は性質が類似するという観点から，非類似度を用いて，着目クラスの全事例からの非類似度が，閾値以上である事例は，そのクラスに属さないと判断する．この閾値として，**限界非類似度**を定義する．

非類似度としては，事例間で属性値の異なる属性の個数を反映させるため，事例の持つ属性値系列をベクトルと見なした場合の n 次元空間における事例間のユークリッド距離が用いられることがある．ただし，どの属性値が異なるかは重要である．本稿では，クラスにおける属性値の重要性を考慮するため，属性値の発生確率に着目する．表 1 におけるクラス『mammal』を例にあげると，属性 “cover” においては，属性値 hair は 1/2 の確率で発生し，none は 1/3 の確率で発生する．このことより属性値 hair や none は，クラス『mammal』において発生における曖昧さが大きいといえる．一方，属性値 feather は確率 1/6 で発生し，『mammal』においては，発生する確率が低く曖昧さが小さい．発生確率の曖昧さの大きい属性値は，そのクラスにおいてさほど重要とはならないが，曖昧さの小さい属性値は，その属性値を持つ，あるいは持たないことは非常に重要である．したがって本手法では，属性値の当該クラスにおける必要性，もしくは重要性を評価し，その結果を非類似度に反映させる．以下に発生確率に基づいた属性値の評価値を定義する．

定義 2（属性値の評価値） クラス C に対し，属性 P における属性値 v の評価値 $E(v)$ を，以下に定義する．

$$\begin{cases} 0 < p(v) < 1 & E(v) = 1 + w\{1 - H(p)\} \\ p(v) = 0, 1 & E(v) = \infty \end{cases}$$

ここで，$p(v)$ は属性 P における属性値 v の発生確率であり，w は発生確率を非類似度に反映させる重みである．なお，$H(p)$ は発生確率によるエントロピーで，以下の式で表される．

$$H(p) = -p\log(p) - (1-p)\log(1-p) \qquad \square$$

定義 2 における発生確率によるエントロピー $H(p)$ は，曖昧さが最も大きいとき，最大値 1 をとる．したがって，評価値 E は，属性値の発生の曖昧さが少なくなるほど大きな値をとり，曖昧さが増すほど小さな値をとる．また，重み w を大きくとると，属性の曖昧さに関する評価が評価値に大きく反映される．例として，表 1 に示すクラス『mammal』の 6 事例を全事例とした場合の，クラス『mammal』における各属性値の発生確率と，$w = 1$ としたときの評価値を**表 2** に示す．

次にあるクラスにおける 2 事例 a，b 間の非類似度 $D(a, b)$ の算出方法を示す．

定義 3（非類似度） $P_k(a)$ が事例 a の k 番目の属性値を表すものとする．このとき 2 事例 a，b 間の非類似度 $D(a, b)$ は次式で与えられる．

$$D(a,b) = \sum_{\substack{k=1 \\ P_k(a) \neq P_k(b)}}^{n} \{E(P_k(a)) + E(P_k(b))\}$$

□

表 2 クラス『mammal』における属性値の発生確率と評価値
Table 2 The evaluation values of attribute values in the class『mammal』.

属性	milk		cover			gills		egg		leg			water		fly	
属性値	1	0	h	n	f	1	0	1	0	4	2	0	1	0	1	0
発生確率	1.0	0.0	0.5	0.33	0.17	0.0	1.0	0.17	0.83	0.5	0.33	0.17	0.33	0.67	0.17	0.83
評価値	∞	∞	1.0	1.08	1.35	∞	∞	1.35	1.35	1.0	1.08	1.35	1.08	1.08	1.35	1.35

たとえば，表 1 の事例 'dog' と 'bat' に関しては，属性値が異なる属性は，"cover"，"leg"，"fly" である．これらの各属性について，'dog' および 'bat' の持つ属性値の評価値 E の総和が非類似度となり，以下のように算出できる．

$$D(\mathrm{dog},\mathrm{bat}) = 1.0 + 1.08 + 1.0 + 1.08 + 1.35 + 1.35 = 6.86$$

未知事例 u が，クラス A に属する場合，u は，クラス A 中のいずれかの事例と類似しているはずである．本手法では，この類似性の限界を**限界非類似度** D_L と呼び，非類似度を用いて以下のように定義する．

定義 4（限界非類似度） クラス C に属する事例集合を $\{a_1, a_2, \cdots, a_n\}$ とするとき，C の限界非類似度 $D_L(C)$ は，次式によって与えられる．

$$D_L(C) = \max_{i=1,2,\ldots,n} [\min_{j=1,2,\ldots,n(i\neq j)} \{D(a_i, a_j)\}]$$

□

ここでは，あるクラスの各事例と，各事例から非類似度が最も小さい事例との非類似度を計算し，それらの最大値を限界非類似度 D_L としている．これは，未知の事例が与えられたときに，クラス A のすべての事例から限界非類似度 $D_L(A)$ 以上であれば，この事例が A に属さないと予測するための基準である．

4. クラス間関係の獲得

クラス A とクラス B の関係は，B 中の事例で A に属するものが存在するかどうか，および，A 中の事例で B に属するものが存在するかどうかから獲得する．B 中の事例 b が A に属するかどうかの判定は以下に示す方法で行うものとする．

step1 A の属性値の評価のもとで，b と最も非類似度の小さい A 中の事例 a_{near} と b の非類似度が，$D_L(A)$ 以下であるならば，step2 へ，そうでないならば step4 へ．

step2 B の属性値の評価のもとで，a_{near} と最も非類似度の小さい B 中の事例と a_{near} との非類似度が，$D_L(B)$ 以下であるならば，step3 へ，そうでないならば step4 へ．

step3 b は A に属すると判定し，終了．

step4 b は A に属さないと判定し，終了．

なお，手順 2 は，2 クラスが事例を共有する場合に，それぞれのクラスにおいて他方のクラスに属する事例が少なくとも 1 つは存在するという仮定に基づく．この仮定のもとでは，b が A に属するならば，b に最も近い a_{near} が B に属していなければならず，それを判定している．以上の方法を用いて，B に与えられる事例が A に属するかどうかを判定することができる．

この結果から，以下の基準に基づき，A と B のクラス間関係を獲得する．

1 $A(B)$ に与えられる全事例が $B(A)$ に属する
→ $B(A)$ は $A(B)$ を**包含**．

2 A に与えられる全事例が B に属さない，かつ B に与えられる全事例が A に属さない．
→ A と B は**排反**．

3 1，2 以外
→ A と B は**一部共有**．

5. 実験と考察

提案手法を C 言語で実装し，簡単な実験を行った．実験対象としてはともに複数のクラスを含む DB《animal》と《zoo》を用いた．《animal》は，ILP システム Progol[4] とともに配布されている動物に関する学習データに，クラスと事例を加えたものである．具体的には，既存の『bird（鳥類）』,『fish（魚類）』,『mammal（哺乳類）』,『reptile（爬虫類）』という相互排反性の成り立つ 4 クラスに加えて,『amphibian（両棲類）』,『fly_animal（飛ぶ動物）』,『human（人）』,『insect（昆虫）』,『water_animal（水と共生する動物）』という同事例を含まない 5 クラスの事例を加えた．事例の持つ属性情報については，客観的事実に基づき決定した．最終的な属性数は 11 個，事例数は 57 個とした．これを用いて，本手法が実世界において相互排反性の成り立たない複数クラスに対しても有効であるかどうかを検証する．また《zoo》は，UCI の機械学習用 DB の 1 つ（URL; http://www.ics.uci.edu/mlearn/MLRepository.html）であり，相互排反な 7 つのクラスを含む．全属性数は 16 個，事例数は 101 個である．以下，各々の DB に対する実験結果を示す．なお，実験では，属性値の評価値における重み w の値は 1 とした．

表 3 《animal》に対する実験結果
Table 3 Experimental result for《animal》.

class1	class2	関係	class1	class2	関係	class1	class2	関係
bird	fly	共有	bird	water	共有	fish	fly	共有
fly	insect	共有	fly	mam	共有	fly	water	共有
insect	water	共有	mam	water	共有	rep	water	共有
water	amphi	包含	water	fish	包含	mam	human	包含
amphi	bird	排反	amphi	fish	排反	amphi	fly	排反
amphi	human	排反	amphi	insect	排反	amphi	mam	排反
amphi	rep	排反	bird	fish	排反	bird	human	排反
bird	insect	排反	bird	mam	排反	bird	rep	排反
fish	human	排反	fish	insect	排反	fish	mam	排反
fish	rep	排反	fly	human	排反	fly	rep	排反
human	insect	排反	human	rep	排反	human	water	排反
insect	mam	排反	insect	rep	排反	mam	rep	排反

amphi: amphibian, rep: reptile, mam: mammal
fly: fly_animal, water: water_animal

5.1 《animal》に関する実験

《animal》の 9 クラスに対して本手法を用いてクラス間関係を獲得した結果を**表 3** に示す．なお，包含関係では，class1 が class2 を包含するものとする．

表 3 に示す 2 クラス間関係は自然界の 2 クラス間関係と同一であり，誤ったクラス間関係が獲得されることはなかった．ここで，一部共有関係となった 2 クラス,『bird』と『fly_animal』に着目し，それぞれに与えられた事例と，共有された事例を以下に示す．

- 『bird』={swan,eagle,parakeet（インコ),ostrich（ダチョウ),parrot,penguin,chicken,swallow}
- 『fly_animal』={goose,kite,flyingsquirrel（ムササビ),cicada(セミ),sparrow,flyingfish(トビウオ)}
- 共有された事例={swan,eagle,parakeet,parrot, swallow,goose,kite,sparrow}

上の結果からも，鳥類で飛ぶ事例がクラス『bird』と『fly_animal』の共有部分となっている．このことから，提案手法における，他クラスの事例が着目クラスに属するか否かの判定が有効であるといえる．

5.2 《zoo》に関する実験

次に，DB《zoo》に対して同様の実験を行った.《zoo》では，各クラス名が $1, 2, \ldots, 7$ という形で与えられる．本実験においては，各事例の名前と性質から，クラス『1』から『7』は，それぞれ,『哺乳類』,『鳥類』,『爬虫類』,『魚類』,『両棲類』,『昆虫類』,『その他』という自然界に実在するクラスであると判断した．

実験結果としては，すべてのクラスの組合せで，それぞれ排反の関係が獲得され，これは自然界の 2 クラス間関係と一致した．この結果からも本手法が，実際の DB に適用可能であることが示された．また，得られた 2 クラス間関係から多重継承および，階層構造等を獲得することも可能である．

本手法では，排反な 2 クラスに属する事例に関して，ある程度以上の非類似性を仮定している．しかし，このような仮定は必ずしも成り立たない．たとえば，UCI の DB の 1 つである《投票》は，民主党と共和党に属する議員が，属性として与えられる法案に賛成したか否かを記したものである．これに本手法を用いた場合，民主党と共和党の議員の中に，属性値の系列が類似したものが存在するため，正しく『排反』の関係を獲得できない．本手法は，属性値の全系列を評価した非類似度を用いるため，一部の属性値によってクラスが特徴付けられる，あるいはクラスと事例の持つ属性値との間の相関が小さい場合には，関係の同定が困難となる．

6. むすび

データベース中に存在するクラス間の関係を獲得する手法を提案した．提案手法では，事例の発生確率に基づいて定義される非類似度を用いて，他クラスの事例が着目クラスに属するかどうかを判定した．実験では，複数クラスを含む DB を対象とし，クラス間関係の獲得に本手法が有効であることを示した．

さらに，異なる複数の DB を対象に，それらの持つ属性の多くが一致する場合，本手法を応用し，それらの DB 中に存在するクラス間の関係を獲得し，その結果を利用することが可能である．

今後の課題として，離散化された連続値属性の各属性値間の類似性の考慮等があげられる．なお，本研究の一部は，日本学術振興会科学研究費補助金奨励研究（11780269）の補助による．

参考文献

1) Schlimmer, J.C.: Learning Determinations and Checking Databases, *Proc. AAAI 91 Workshop on Knowledge Discovery in Databases*, Anaheim, CA, pp.64–76 (1991).
2) McCallum, A., Rosenfeld, R., Mitchel, T. and Ng, A.Y.: Improving Text Classification by Shrinkage in a Hierarchy of Classes, *Proc. 15th ICML*, pp.359–367 (1998).
3) 森中，高，大原，馬場口，北橋：類似性予測による分類学習用データベースからのクラス間関係の獲得，信学総大 (1999).
4) Muggleton, S.: Inverse Entailment and Progol, *New Generation Computing*, Vol.13, pp.245–286 (1995).

(平成 12 年 2 月 21 日受付)
(平成 12 年 9 月 7 日採録)