G-19

事例に基づいた最大事後確率分類
Instance-based maximum a posteriori classification

山田泰大* 犬塚信博† 世木博久‡
Yamada Yasuhiro Inuzuka Nobuhiro Seki Hirohisa

1 はじめに

知識発見やデータマイニング，機械学習の多くの手法ではオブジェクト間の類似度がその結果を左右する最も重要な要因の一つである．例えば，k 個の類似した事例を参考にして，事例の属する分類を予測する方法に k-最近隣法がある．これまでにオブジェクト間の類似度を求める多くの手法が提案されている [1,2]．しかし，その類似度を用いて導出された結果が何を意味するのか明確ではない．また，類似度は対象問題に依存して定まるはずであるが，問題から独立して類似度を定めてしまうこともある．そこで，我々は類似性尺度に次の 2 つの性質を要請する．(1) 理論的に目標達成との間の関係が説明ができる事．(2) 対象としている問題 (分類対象) 毎に，異なった視点で異なった類似度を求められる事．本論文では事後確率を最大にする分類を導くための類似性尺度の帰納法を提案する．また，この類似度を利用し k-最近隣法による投票結果から，分類に対する事後確率を計算できる事を示し、これを用いた最大事後確率分類法を与える．

2 類似性尺度

2.1 準備

U は事例空間，$\mathcal{C}$ は可能な分類の集合，$\langle x, c_x \rangle$ は訓練事例 (事例 $x \in U$ とその分類 $c_x \in \mathcal{C}$ のペア)，D_{base} は訓練事例の集合，X, Y は事例空間 U からその分布に従って選択された事例を表す確率変数，C_X, C_Y は事例 X，Y が属するクラスを表す確率変数を表すものとする．また，本論文では訓練事例にはノイズを含まず，事例に対してその分類は決定的に決まり，さらに，事例は独立かつ同一の分布にしたがって存在すると仮定する．

2.2 類似性尺度の定義

分類問題のために類似性尺度を必要とする時，「同じ分類に属するであろう事例は互いに類似している」，「類似していれば，同じ分類に属する可能性が高い」と云う事ができる．これらの文を確率の言葉で表し，事例 x，y 間の類似度 $sim(x,y)$ を式 (1) で定義する．

$$sim(x,y) = P(C_X = C_Y | X = x, Y = y) \qquad (1)$$

即ち，二つの事例が与えられたという条件の元で，それらの事例が同じ分類に属する条件付き確率で類似度を定義する．

2.3 類似度と分類投票

式 (1) で定義した類似度を使って，重み付 k-最近隣法と類似した分類投票を行なう．そのために，事例 x の分類 c に対する投票 $vote(x,c)$ を以下のように定義する．

$$vote(x,c) = \sum_{y \in U} sim(x,y) \cdot P(Y = y) \cdot \delta(c, c_y) \qquad (2)$$

ここで，$\delta(a,b)$ は $a = b$ なら 1，そうでなければ 0 をとる．式 (2) による投票と k-最近隣法の投票は事例の存在確率を表す追加的な項 $P(Y = y)$ が乗ぜられている点で異なっている。

2.4 類似度と事後確率

定理 事例の分類 $c \in \mathcal{C}$ に対する事後確率は $vote(x,c)$ と分類に対する事前確率 $P(C = c)$ から式 (3) で与えられる．

$$P(C_X = c | X = x) = \frac{vote(x,c)}{P(C = c)} \qquad (3)$$

証明.

式 (1) を式 (2) に代入すると以下が成り立つ．

$$vote(x,c) = \sum_{y \in U} sim(x,y) \cdot P(Y = y) \cdot \delta(c, c_y)$$

$$= \sum_{y \in U} P(C_X = C_Y | X = x, Y = y) \cdot P(Y = y) \cdot \delta(c, c_y)$$

可能な全ての分類に対して展開し，$P(C_X = C_Y | X = x, Y = y) = \sum_{i \in \mathcal{C}} P(C_X = i, C_Y = i | X = x, Y = y)$ を代入すると，次のようになる．

$$= \sum_{y \in U} \sum_{i \in \mathcal{C}} P(C_X = i, C_Y = i | X = x, Y = y) \cdot P(Y = y) \cdot \delta(c, c_y)$$

次に 事例は独立かつ同一の分布に従っているとの仮定から，

$$= \sum_{y \in U} \sum_{i \in \mathcal{C}} P(C_X = i, C_Y = i, Y = y | X = x) \cdot \delta(c, c_y)$$

事例に対する分類が決定的かつ，訓練事例にノイズがないと仮定すると，$i \neq c_y$ の時，$P(C_X = i, C_Y = i, Y = y | X = x) = 0$ である．このことから，次式が導かれる．

$$= \sum_{y \in U} P(C_X = c_y, C_Y = c_y, Y = y | X = x) \cdot \delta(c, c_y)$$

事例空間 U 中の事例を分類が c であるものと，そうでないものに分割し，次式を得る．

$$= \sum_{y \in U \wedge c_y = c} P(C_X = c_y, C_Y = c_y, Y = y | X = x) \cdot 1 + \sum_{y \in U \wedge c_y \neq c} P(C_X = c_y, C_Y = c_y, Y = y | X = x) \cdot 0$$

*名古屋工業大学大学院工学研究科電気情報工学専攻
†名古屋工業大学工学部電気情報工学科
‡名古屋工業大学工学部知能情報システム学科

$c \neq c_y$ の時 $P(C_X = c, C_Y = c, Y = y|X = x) = 0$ より，

$$= \sum_{y \in U \wedge c_y = c} P(C_X = c, C_Y = c, Y = y|X = x) + \sum_{y \in U \wedge c_y \neq c} P(C_X = c, C_Y = c, Y = y|X = x)$$

$$= \sum_{y \in U} P(C_X = c, C_Y = c, Y = y|X = x)$$

また，事例は独立かつ同一の分布に従っているので，

$$= \sum_{y \in U} P(C_X = c|X = x) \cdot P(C_Y = c, Y = y)$$

$$= P(C_X = c|X = x) \cdot \sum_{y \in U} P(C_Y = c, Y = y)$$

$$= P(C_X = c|X = x) \cdot P(C_Y = c)$$

$$= P(C_X = c|X = x) \cdot P(C = c)$$

以上から，$P(C_X = c|X = x) = vote(x, c)/P(C = c)$

2.5 最大事後確率分類

前節の結果から，事例 $x \in U$ の分類 $class(x)$ は最大事後確率推定法を用い，式 (4) で予測する．

$$class(x) = \operatorname*{argmax}_{c \in C} P(C_X = c|X = x) = \operatorname*{argmax}_{c \in C} \frac{vote(x, c)}{P(C = c)} \quad (4)$$

3 類似度学習と類似度計算の枠組

本節では，提案した類似度をデータから帰納する方法を示す．式 (1) の計算には，確率モデルの構築が必要である．

まず，事例とその分類ペア $\langle x, c_x \rangle$ の集合 D から，事例の分類に対する事後確率を学習する事後確率学習器 PPL(ex. Bayesian network) の存在を仮定する．

式 (1) の確率モデルをデータから学習する簡単な方法は二つの訓練事例 $\langle x, c_x \rangle$,$\langle y, c_y \rangle$ を組み合わせて 1 つの訓練事例に変換し，その目標値は同じ分類に属するか否かの真偽値とすることでこれを PPL に学習させることである．

即ち，我々は基本データベース $D_{\text{base}} = \{\langle x, c_x \rangle\}$ を変換し，新たに結合データベース D_{cmb} を作成する．

$$D_{\text{cmb}} = \{\langle cmb(x, y), \delta(c_x, c_y) \rangle | \langle x, c_x \rangle, \langle y, c_y \rangle \in D_{\text{base}}\} \quad (5)$$

ここで，$cmb(x, y)$ は D_{base} 中の 2 つの事例の記述を組み合わせた 1 つの事例の記述を表し，結合データベース D_{cmb} 中の事例の分類は $\delta(c_x, c_y)$ で与えられ，0 または 1 をとる．

事例 $x, y \in D_{\text{base}}$ 間の類似度は，事例 x，y の結合事例 $cmb(x, y)$ を PPL によって求められた確率モデルに与え，その分類が 1 となる事後確率を求める事によってなされる．

4 提案手法と事後確率学習

直接，事例の分類に対する事後確率を学習する学習器を DPPL，事例間の類似度から，事例の事後確率を導く方法を sim_{MAP} 法と呼ぶ事にする．DPPL と sim_{MAP} 法との違いは，学習対象が異なることである．DPPL は対象問題に対する事例とその分類からなる基本データベースから学習するのに対し，sim_{MAP} 法は，2 つの事例とそれらが同じ分類に属するか否かからなるデータベースで学習する．

我々は以下の考察から，sim_{MAP} 法の方が DPPL による事後確率学習よりも優れていると考える．同じ分類に属するのかどうかを判定する方が，どの分類に属するのかを判定するよりも，直観的に簡単である事が多い．例えば，論文データベースでは，ある論文がどの分野に属するのか判定する時，参考にしている論文と同じ分野に属する可能性が高い．また，事例間の比較を行なうため，直接分類を求める時と違い，扱える情報量が増える．例えば，論文データベースのように，参考にしている論文と同じ分野に属すると云った知識は，ルールベースシステムに埋め込む事ができない．なぜなら，全ての参考論文の存在 (無限にある) を知っていないといけないから．また，決定木学習のような学習法では，事例を解析しルールを帰納する時に，事例毎に解析しているので，事例同士を比較するような知識を利用する事ができない．

5 アルゴリズムの近似

式 (2) では，理論上 事例空間中の全ての事例を用いて，事後確率を求めている．実際には限られた有限個の訓練事例しか利用できない．そこで事例空間 U を訓練事例の集合で近似する．また，式 (2) では，通常の k-最近隣法と異なり，余分に事例の存在確率 $P(Y = y)$ が乗ぜられている．この存在確率を D_{base} から経験的に求めると以下のようになる．

$$\hat{P}(Y = y) = \frac{N_{D_{\text{base}}}(y)}{|D_{\text{base}}|}, \quad (6)$$

ここで，$N_{D_{\text{base}}}(y)$ はデータベース中で事例の記述が y と一致した事例の個数である．つまり，データベース中の各事例が k-最近隣法のように投票すれば，$P(Y = y)$ の項は定数となり，最大事後確率分類を選ぶのに影響しない．そこで，式 (2) から，この定数項を取り除く．結果として，sim_{MAP} 法の投票は k-最近隣法の投票と一致する．ただし，sim_{MAP} 法では分類予測の段階で，投票結果を分類に対する事前確率で正規化している点が異なっている．以上をまとめると，式 (2) の近似式は式 (7) のようになる．

$$\hat{vote}(x, c) = \sum_{\langle y, c_y \rangle \in D_{\text{base}}} sim(x, y) \cdot \delta(c, c_y). \quad (7)$$

6 結論と今後の課題

最大事後確率分類のための類似性尺度の帰納方法を提案した．そして，提案した類似度と分類の間の関係を理論的に示した．また，提案した類似度帰納法は分類対象毎に異なった類似性尺度を導出する事ができる．本論文では事例に対して分類が決定的でノイズを含まない問題領域に対して，式 (1) で定義した類似度を用いて重み付投票をした結果，事後確率が求められる事を示した．しかし，分類が非決定的な場合に対応する式は明確にはなっていない．その場合の式の導出は今後の課題であるが，k-最近隣法は事例のノイズに強い事などから，ある程度のノイズに対しても本手法は耐性があると考えられる．

参考文献

[1] D. Lin, An Information-Theoretic Definition of Similarity, Proceedings of the Fifteenth International Conference on Machine Learning, pp296-304, 1998.

[2] L. Martin and F. Moal, A Language-Based Similarity Measure, in Proceedings of the 12th European Conference on Machine Learning, pp.336-347, 2001.