平成 14 年度　　修士論文概要

確率論を用いた類似性尺度による事例分類に関する研究

世木研究室　山田泰大
学籍番号　No. 13417109

1　序論

事例の分類を予測する方法に類似度に基づく方法がある．それらの代表的な方法に，k-nearest neighbor(k-NN) 法がある．類似度に基づく方法は，目標関数が複雑な問題領域において大きな利点を発揮する．また，現在の問題を解決するために過去の類似した事例を利用するため，対象問題上での事例間の類似性尺度が重要となる．ここで,「**類似度**」と云う概念は，データマイニングや情報検索などで工学的に利用されるほか，人間の認知的活動をモデル化するという点で類似度概念それ自体が興味深い研究対象である．本論文では，事例間の類似度を測る枠組である「**類似性尺度**」について考え，それを事例の分類に利用することを考える．

2　類似度に基づいた分類法

類似度に基づく分類を行う代表的な手法である k-NN 法では，近傍の事例を大雑把に同一視することによって，事例の事後確率分布を相対頻度で推定していた．この推定では，適切な尺度で「類似した近傍の事例間では，分類の事後確率が大きく変化しない」という仮定が成り立たねばならない．しかし，実際には類似性尺度として場当たり的なものが使われてきた．また，事例が数値表現ではない一般の場合にはユークリッド距離のような自然な距離尺度を利用することができない．そこで，これらの問題を統一的に扱える類似性尺度の枠組を定義する必要がある．

また，類似性尺度は状況 (文脈) に応じて変化するものであるが，従来の類似性尺度の多くはこの点で不十分である．そこで本研究では，ある分類を行うという文脈において，分類目標に変化する類似性尺度の定義を提案する．さらにその類似性尺度を用い k-NN 法と類似した分類法を与える．

3　類似性尺度

分類問題で類似性尺度を必要とする時，それに求められる性質は「同じ分類に属すであろう 2 つの事例は互いに類似している」,「互いに類似している事例は同じ分類に属する可能性が高い」と云うことができる．これを確率の言葉で表現し，事例 x, y 間の類似度をその事例を与えた下でそれらが同じ分類に属する条件つき確率で定義する．

定義 (類似性尺度) 事例 $x, y \in U$ 間の類似度 $sim(x, y)$ を次式で定める．

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

この定義では事例表現に依存せず，数値や記号表現などを統一的な理論的枠組で取り扱うことができる．また，事例間の類似度に対して,「同じ分類に属する事後確率」という確率的な意味が，明確に与えられていることも，大きな特徴である．そして,「どのような分類を行うか」という文脈ごとに，異なる類似度が得られるという性質も，この類似性尺度の定義は備えている．

次に式 (1) で定義された類似性尺度を利用した分類投票を定義する．これは、標本が無限の場合の、重み付き k-NN 法の分類投票と等しい．

定義 (分類投票) 事例 $x \in U$ の分類 $c \in \mathcal{C}$ に対する分類投票を次式で定義する．

$$vote(x, c) = \sum_{y \in U} sim(x, y) P(Y = y, C_Y = c) \quad (2)$$

ここで，U は事例空間，$\mathcal{C}$ は全ての分類の集合を表す．

次の定理はある条件を付けることで，類似度から事後確を率導出できることを示す．この定理はこの類似性尺度が妥当なものであることを示している．

定理 (類似度による事後確率導出定理) 事例に対する分類は決定的であり，独立かつ同一の分布 (iid) に従って生起するとき，事例 $x \in U$ の分類 $c \in \mathcal{C}$ に対する事後確率は投票 $vote(x, c)$ と分類の事前確率 $P(C = c)$ により，式 (3) で与えられる．

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

これは，通常の重み付き k-NN 法による投票の最大化による分類予測とは異なり，投票結果を分類の事前確率で修正すべきことを示しており，この点で通常の重み付き k-NN 法と，提案手法は大きく異なる．この結果から，事例 x に対する分類は式 (5) で予測できる．

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

$$= \operatorname*{argmax}_{c \in \mathcal{C}} \frac{vote(x, c)}{P(C = c)} \quad (5)$$

4　提案法の実現

前述の分類法の実現はデータからの学習と事例の分類からなる．まず，学習では訓練事例集合 S_{base} を入力とし，次の (a)〜(c) を行う．(a) 訓練事例集合 S_{base} を単に記憶する．(b) 分類の事前確率分布 $P(C)$ を 訓練事例集合 S_{base} から推定する．(c) 類似性尺度 $sim(\cdot, \cdot)$ を学習する．

ステップ (c) では訓練事例集合 S_{base} を結合事例の訓練集合 S_{cmb} に変換し，その上で従来の確率モデル学習器を利用することにより，式 (1) の値を推定する．ここで，結合事例の訓練事例集合 S_{cmb} は次式で定義される．

$$S_{\mathrm{cmb}} = \{\langle cmb(x,y), \delta(c_x,c_y)\rangle | \langle x,c_x\rangle, \langle y,c_y\rangle \in S_{\mathrm{base}}\}$$

訓練結合事例 $cmb(x,y)$ は，2 つの事例 $x,y(\in S_{\mathrm{base}})$ が同じ分類に属すか否か $\delta(c_x,c_y)$ を分類とする新たな事例を表す．例えば事例が属性値表現される場合，事例の結合事例の生成法として，各属性が同じ属性値か否かからなる一致汎化結合法や，各属性の属性値のペアを属性とする直積結合法などが考えられる．

一方，事例の分類は，式 (2) と式 (5) で行う．ただし，実際は利用できるデータは訓練データに限られているので，訓練事例集合 S_{base} で事例の分布を近似し，投票の式 (2) は式 (6) で近似した．ここで，その際に式 (5) で無関係な定数倍は取り除いた．

$$\widehat{vote}(x,c) = \sum_{\langle y,c_y\rangle \in S_{\mathrm{base}}} sim(x,y)\cdot\delta(c,c_y). \quad (6)$$

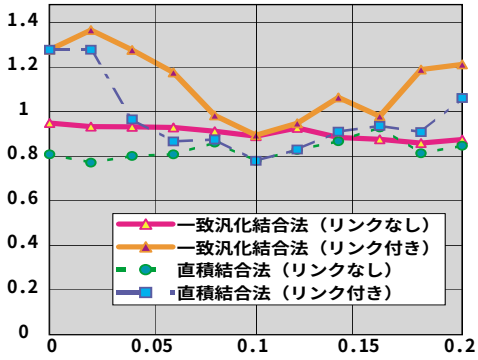
このように分類投票の手続きは通常の重み付き k-NN 法と一致する．しかし，最終的な分類が式 (3) に従って分類の事前確率で修正された結果を最大化する分類を出力する点で k-NN 法の分類予測とは大きく異なる．

5　提案手法と事後確率学習

ベイジアンネットワークのような通常の事後確率学習器も本手法も同様に定理 (式 (3)) によって，最終的には同じ事後確率を推定する．しかし，類似度 (式 (1)) を学習することは，直接事後確率を学習することは互いに性質の異なる学習対象であり，必然的に学習のし易さは異なる．本研究では，「論文 A の属す分野は参考文献 B と類似した分野に属すだろう」と云うように，事例の類似性が判断し易い場合や，事例の比較によって単一の事例解析では扱えない事例間の関係情報を扱える場合に提案手法の方が従来の直接的な事後確率学習よりも優れていると考える．一方，本手法は確率モデル学習器を内部で利用するため，計算量的な問題がある．

6　実験

UCI 機械学習用 DB および人工データで実験を行なった．本実験では類似度学習に事後確率学習器として，ナイーブベイズ (NB) 法を用いた．そして，NB 法と提案手法の精度比較を行なった．UCI の DB 上で本手法は，DB vote で 96.0%，balance. で 88.3%，flare2 で 72.1%となり，NB 法を通常通り適用した場合と同程度の結果を得，方法の動作を確認した．一方，人工データでは NB 法の仮定に従った事例生成を確率的に行なった．そのため，NB 法がモデルに従った最良の分類器といえる．また，同時に提案手法の有利な場合として，Web のリンクを模し，事例間に同じ分類であれば確率 $p = 0.1$ で，異なる分類の場合には確率 q(実験毎に q を変化させた) でリンク情報を付加した．提案法では，このリンク率 p,q に差があるほどよい分類予測ができると期待できる．NB 法の精度に対する本手法 (2 種類の結合事例生成法に対してリンクあり・なしの場合) の相対精度を図に示す．



図：提案手法の NB 法に対する相対精度．横軸は q の値を，縦軸は相対精度を表す．

7　結論と今後の課題

確率を用いた類似性尺度の定義とそれを用いた分類法を提案した．この定義では分類という観点によって類似度が変化する性質を備え持つ．さらに，一定の条件を下であるが，分類投票の結果を分類の事前確率で修正することにより，分類に対する事後確率が計算できることを示した．このことは，提案した類似性尺度が妥当なものであることを示している．また，本論文では比較実験により本提案手法の有効性を確認した．定理は不確実性をともなう場合は明確ではないが，重み付き k-NN 法はノイズに対して耐性が強いことから本手法もある程度の不確定性に対しても，対応できると考えられる．人工データにおける実験は提案手法が真に不確定な領域に対しても十分適用できる可能性を示している．今後は応用問題として，実際のハイパーテキストの分類問題のような，事例間に明示的な関連がある問題領域で本手法の有効性を確認する必要がある．

参考文献 Yasuhiro Yamada, Nobuhiro Inuzuka, Hirohisa Seki, MAP Classification with a Similarity Measure, in Proceedings of The IASTED International Conference on Artificial and Computational Intelligence, pp. 155-160, 2002.