平 成 15 年 度　　卒 業 論 文 概 要

入学年度 平成 12 年度	**学生番号** 12117660	**氏名** 小酒井 一稔
卒業研究題目 事例間類似度の帰納に基づく分類手法		和田・犬塚 **研究室**

1. はじめに

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．この方法では，問題領域での類似性尺度が重要となる．

データ形式に依存しない定義を持つ類似性尺度として，与えられた 2 つの事例の分類が一致する確率で定義される類似性尺度 [1] が提案されている．この類似度は，2 つの事例の関係が記述された事例「**結合事例**」と，それらの事例が同分類に属するかどうかから学習することで得られるものである．しかし，既存の結合事例生成手法では，扱うことのできるデータ形式が限られているという問題がある．

そこで本論文では**学習対象の特徴に合わせた柔軟な記述ができ，連続的な値を扱うことができる**手法を提案する．この手法は，属性値が分類に与える影響を距離として算出する手法 [2] を適用して，事例間関係を記述する手法である．そして，ある一定の距離ごとに汎化を行ない同一視することで，共通部分の抽出が可能である．

2. 類似度に基づく分類

本論文で用いる，確率論を用いた類似性尺度による分類法とは，式 (1) によって定義される類似性尺度を用い，k-NN 法と類似した分類を行なう方法である．

まず事例間の類似度を，2 つ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 1 (類似性尺度) 事例 $x, y \in U$ 間の類似度 $sim(x, y)$ を次式で定義する．

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

次に分類投票を定義する．式 (1) で定義される類似性尺度を利用した分類投票は以下のように定義される．

定義 2 (分類投票) 事例 $x \in U$ の分類 $c \in C$ に対する分類投票を次式で定義する．

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

ここで，[1] では以下の定理が証明されている．

定理 1 (類似度による事後確率導出定理) 決定的分類を持つ事例が，独立かつ同一の分布 *(iid)* にしたがって生起するとき，事例 $x \in U$ の分類 $c \in C$ 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．

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

以上の結果から，事例 $x \in U$ に対する分類は式 (5) で予測できる．そして，このような最大事後確率 (MAP) 分類法は予測エラーの期待値を最小化するものである．

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

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

以上の方法を用いて分類する手法を，sim_{MAP} 法と呼ぶ．

3. 結合事例生成手法

sim_{MAP} 法では，事例を結合することで事例間関係を記述する．具体的には，事例 $x, y \in U$ から，結合事例 $cmb(x, y)$ を生成する．結合事例から学習することによって類似度が算出される．

提案手法では，事例結合に，事例間における「距離」を用いる．分類に与える影響の大きさを「距離」として算出し，値が近いものを，分類への影響が近いとみなし，**汎化**する．この汎化された距離の値を属性値として持つ，結合事例を生成する．汎化の度合を変化させることにより，より柔軟に事例間関係を記述することが可能となる．

定義 3 (提案手法：距離汎化結合法) 事例 $x, y \in U$ が $x = (x.A, x.B, \cdots)$，$y = (y.A, y.B, \cdots)$ であり，適切な距離関数 $distance(\cdot, \cdot)$ によって，属性値間の距離が定義されるとき，以下のように事例を結合する．ただし， g は汎化度合であり，この値と得られた距離値から，新たな属性値を生成する．

$$cmb(x, y) = (generalize(distance(x.A, y.A)), \\ generalize(distance(x.B, y.B), \cdots)) \tag{6}$$

$$generalize(d) = \lfloor d/g \rfloor \tag{7}$$

このように，式 *(7)* のような汎化関数 $generalize(\cdot)$ によって汎化された距離値を属性値に持つ結合事例を生成する方法を，**距離汎化結合法**と呼ぶ．

4. 実験および考察

実験データには，UCI の機械学習データベースを用い，評価方法には k-fold cross validation 法を用いる．

実験では，訓練事例集合から分類に対する事後確率を学習する，確率モデル学習器 (PPL) が用意されているものとし，ナイーブベイズ学習器を PPL として使用した．

事例間距離を導入したことで，k-NN 法との比較実験が可能となった．また比較実験として，決定木生成システム C4.5 を用いた分類を行なった．

表 1 ：実験結果 (g は汎化度合)

データベース名	C4.5	5-NN	距離汎化結合法			
			$g = 0.001$	$g = 0.01$	$g = 0.1$	$g = 1$
balance	0.797	0.386	**0.851**	0.845	0.845	0.778
flare1	0.576	0.344	0.576	0.582	**0.607**	0.517
hayes-roth	**0.742**	0.598	0.659	0.659	0.644	0.394
promoters	0.811	0.500	0.632	0.642	0.802	**0.896**
wine	0.865	0.472	**0.949**	0.933	0.921	0.888

sim_{MAP} 法は，事例数や属性数が少ないと不安定な結果となったが，k-NN 法と比較すると優れているといえる．C4.5 に対しては，精度を下回ることもあった．

距離汎化結合法を用いた場合の分類精度は，汎化度合の変化に対して単峰性があると考えられる．これは，本手法が学習対象に合わせた学習が可能であることを示している．

5. 結論と今後の課題

本論文では，事例間距離を用いた結合事例生成手法を提案した．この手法は，属性値間の距離を算出し，汎化することで，属性値間の隔たりの度合を属性値に持つ事例を生成する．汎化の度合を変化させることにより，学習対象の特徴に応じた学習が可能となる．また本手法を，既存の手法では扱うことのできなかった，連続的な値をとる属性に対して適用し，実験より有効性を確認できた．

今後の応用問題としては，より実際に分類を必要とする条件を想定した，本手法の有効性に関する実験が必要である．

参考文献

[1] 山田 泰大，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，修士論文，名古屋工業大学，2003

[2]D.Randall Wilson and Tony R.Martinez, Improved Heterogeneous Distance Functions, Jounal of Artificial Intelligence Research,1997