最近隣法の正答率に対する理論的解析

3J－2

岡本 青史　佐藤 健
（株）富士通研究所

1. はじめに

現存する多くの事例ベース推論システムにおいて，事例間の類似性は問題領域に強く依存した形で定義されている．このことは，類似性の定義を再利用することや，その妥当性を評価することが困難であるといった問題を引き起こしている．この問題を解決するために，我々は事例間の類似性に関する理論的解析が重要であると考え，PAC学習や平均的解析（Average-Case Analysis）の枠組を用いて類似性に関する研究を行なっている[2, 3]．本論文では，類似性の定義で最も基本的な最近隣法に対する平均的解析を行なう．

最近隣法に関する平均的解析は，関連属性の連言で定義される目標概念を対象とした研究が存在する[1]．その枠組における主な解析結果は，非関連属性数が正答率に顕著な影響を与えるというものである．しかし，クラスの特徴を表しているはずの関連属性に関する解析は十分でない．そこで，本論文では[1]とは異なった問題定義を与え，関連属性数と正答率の関係を中心に最近隣法に関する平均的解析を行なう．

2. 問題定義

$F_1 = \{f_1, \ldots, f_{r_1}\}, F_2 = \{f_{r_1+1}, \ldots, f_{r_1+r_2}\}$ を属性の集合とし，$F_1 \cup F_2 \neq \emptyset, F_1 \cap F_2 = \emptyset$ を仮定する．また，f_i の属性値を a_i で表し，$a_i \in \{0, 1\}$ とする．

$I = (a_1, \ldots, a_{r_1+r_2})$ として，例空間 $\mathcal{H}$ を次のように定義する．

$$\mathcal{H} = \{\, I \mid (a_1 \cdots a_{r_1}) + (a_{r_1+1} \cdots a_{r_1+r_2}) = 1 \,\}.$$

ここで，この例空間 $\mathcal{H}$ 上の確率分布 $\mathcal{F}$ は一様であると仮定する．

クラス数は2とし，それぞれのクラス C_1，C_2 は，$I \in \mathcal{H}$ として次のように定義される．

$$C_1 = \{\, I \mid a_1 \cdots a_{r_1} = 1 \bigwedge a_{r_1+1} \cdots a_{r_1+r_2} = 0 \},$$
$$C_2 = \{\, I \mid a_1 \cdots a_{r_1} = 0 \bigwedge a_{r_1+1} \cdots a_{r_1+r_2} = 1 \}.$$

このとき，F_1, F_2 の要素をそれぞれ C_1, C_2 に対する関連属性という．

本論文では，最近隣法を以下のように定義する．

確率分布 $\mathcal{F}$ に従って独立に得られた n 個の訓練事例を全て，事例ベース $\mathcal{CB}$ 中に格納する．
この時，テスト例 T とのハミング距離が最も小さい訓練事例の集合 $\mathcal{N} \subseteq \mathcal{CB}$ に対し，$S \in \mathcal{N}$ をランダムに選び，T を S と同じクラスに分類する．

上述の仮定のもとで，最近隣法がテスト例 T を正しく分類する確率である正答率に関する平均的解析を行なう．

3. 解析

本論文では紙面の都合上，正答率の理論的導出結果に関する記述を省略する．この理論的正答率の導出過程及び導出結果は[2]を参照されたい．[2]において，正答率 A は r_1, r_2, n の関数として表され，これらの変数に値を代入することで解析を行なう．

3.1. $r_1 = r_2$ の場合

図1は $r_1 = r_2$ という仮定のもとで，関連属性数と訓練事例数が正答率に及ぼす影響を示している．図1中の横軸は関連属性数 $r_1 (= r_2)$ を表しており，縦軸は正答率を表している．図1から以下の解析結果を得ることが出来る．

- 訓練事例数が多いほど正答率が高いことが，いずれの関連属性数に対しても成立する．
- いずれの訓練事例数に対しても，関連属性数の増加に伴って最初は正答率が単調に低下するが，ある関連属性数を谷として正答率は単調に上昇する．
- 谷となる関連属性数は，訓練事例数が多いほど大きくなっている．

今，テスト事例 T の属すクラスを C，属さないクラスを $\overline{C}$ とする．この時，正答率 A は $\mathcal{N}$ 中に C に属する訓練例しか存在しない場合の正答率 A_1 と，C に属する訓練例と $\overline{C}$ に属する訓練例の両方が存在する場合の正答率 A_2 の和で表される．ここで，$A_2 = 0$ とした場合の正答率を図2で表す．図2から以下の解析結果を得ることが出来る．

- いずれの訓練事例数に対しても，関連属性数の増加に伴って正答率が単調に減少する．

これらの解析結果から，関連属性数の増加に伴う最初の正答率 A の低下は，A_1 の低下が A_2 の上昇より大きいことから，また谷を境とした A の上昇は，その逆が成り立つことから説明出来る．

A Theoretical Analysis of Predictive Accuracy for the Nearest Neighbor Classifier
Seishi Okamoto, Ken Satoh
Fujitsu Laboratories Ltd.
E-mail: {seishi,ksatoh}@flab.fujitsu.co.jp

3.2. r_1 と r_2 の関係

図3は98%以上の正答率を得るために必要となる訓練事例数を表している．図3中の横軸は r_1 を，縦軸は訓練事例数を表しており，$Attr2$ は r_2 を表している．図3から以下の解析結果を得ることが出来る．

- いずれの r_2 に対しても，$r_1 = r_2$ となる r_1 から遠ざかるに伴って，必要な訓練事例数は単調に増加し，ある山を越えると単調に減少する．
- 山となる訓練事例数は，r_2 の増加に伴って少なくなる．
- いずれの r_2 に対しても，必要な訓練事例数は $r_1 = r_2$ となる r_1 を軸としてほぼ対称となる．

図4は $r_2 = 16$ の場合の谷と山にあたる $r_1 = 16, 20$ の場合についての正答率を示している．図4中の横軸は訓練事例数を，縦軸は正答率を表している．さらに，$A(i,j), AN(i,j)$ はそれぞれ，$r_1 = j, r_2 = i$ の場合の正答率 A, A_1 を表している．図4から以下の解析結果を得ることが出来る．

- $A(16,16)$ と $A(16,20)$ に関して，訓練事例数が4以下の場合は前者が後者より小さく，そうでない場合には逆が成立する．

これらの解析結果から，r_2 を固定した場合，非常に少ない事例数では $r_1 = r_2$ となる問題が最も難しく，ある程度の事例数が与えられると，$r_1 = r_2$ となる問題よりも r_2 から少し離れた r_1 をとる問題の方が難しくなることが分かる．

4. おわりに

本論文では，正答率の理論的導出に基づいた最近隣法の平均的解析を行い，関連属性数と正答率の関係についての有用な解析結果を得ることが出来た．

今後の課題としては，まず現在の平均的解析の枠組をノイズが扱えるように拡張することが挙げられる．さらに，属性の重みを用いて定義される類似度関数に対する平均的解析を行なっていきたい．

参考文献

[1] P.Langley and W.Iba. Average-Case Analysis of a Nearest Neighbor Algorithm. *Proceedings of IJCAI'93*, pp.889–894, 1993.

[2] S.Okamoto and K.Satoh. A Mathematical Predictive Accuracy for the Nearest Neighbor Classifier. *Proceedings of EWCBR'94*, to appear, 1994.

[3] K.Satoh and S.Okamoto. Toward PAC-Learning of Weights from Qualitative Distance Information. *Proceedings of AAAI'94 Workshop on CBR*, pp.128–132, 1994.

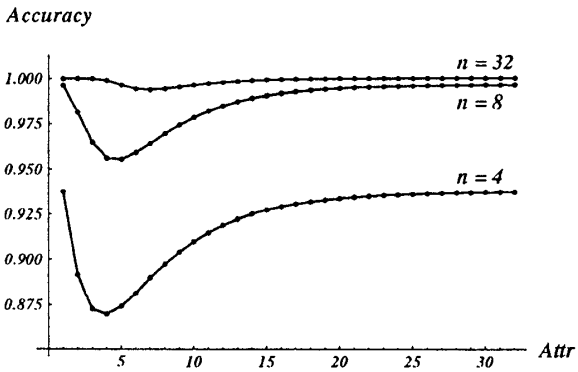

図 1: $r_1 = r_2$ の場合の正答率

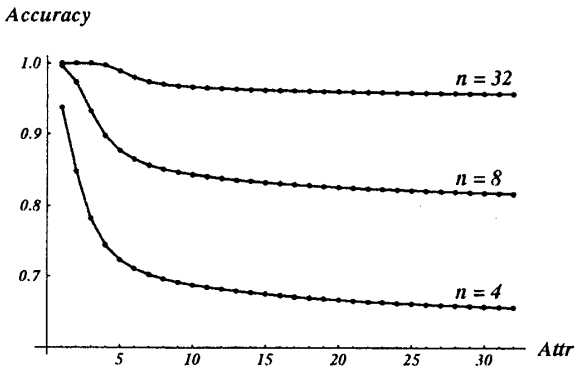

図 2: $A_2 = 0$ とした場合の正答率

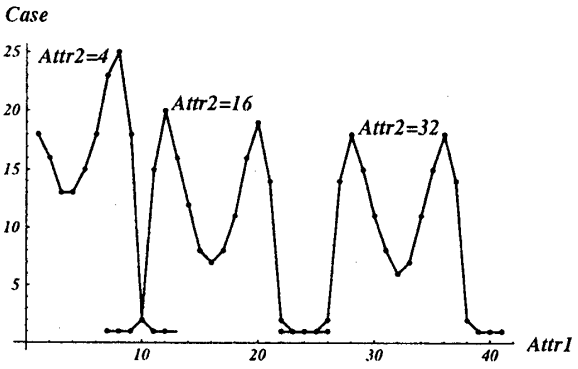

図 3: 正答率 98% を得るために必要な訓練事例数

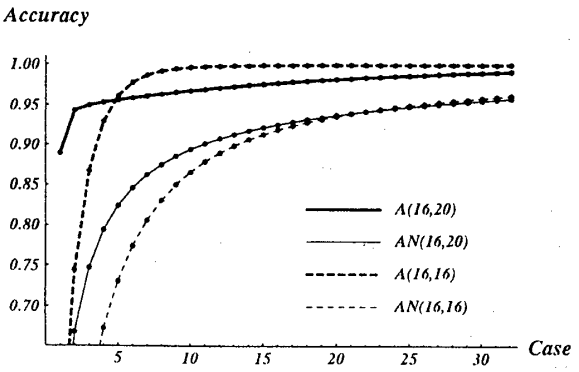

図 4: $r_2 = 16$ の場合の正答率