最小近傍法の平均的挙動の解明

5J－1

岡本 青史　湯上 伸弘
（株）富士通研究所

1. はじめに

平均的解析 [4] は，学習アルゴリズムの正答率に関する挙動を解析する理論的な枠組である．平均的解析では，事例空間上のある確率分布のもとで，目標概念に対する学習アルゴリズムの正答率を，訓練事例数や属性数などの領域パラメータの関数として理論的に導出する．この正答率関数を用いることで，学習アルゴリズムの挙動を解析することができる．

学習アルゴリズムの挙動を解析する他の理論的な枠組として，PAC学習が有名である．しかしながら，PAC学習の枠組における学習に必要な訓練事例数の見積り（sample complexity）は，実際の問題領域で得られる訓練事例数とあまりにもかけ離れている場合が多い．このように，PAC学習には，その解析結果を実際の実験結果と関連付けるのが困難であるという問題がある．

一方，平均的解析は，学習アルゴリズムの平均的な挙動を解析するため，その解析結果は実験結果と直接関連付けられるという特長を持つ．さらに，平均的解析では，正答率を領域パラメータの関数として導出するため，正答率と様々な領域パラメータの関係を解析できるという特長も持つ．

これらの特長を生かして，様々な学習アルゴリズムの平均的解析が行なわれているが（例えば，[2, 3]），従来の平均的解析には，以下の問題点がある．

1. 限定した概念クラスに対する解析を行なっているため，得られた解析結果を他の概念クラスに流用するのが困難である．
2. 正答率関数が複雑になり過ぎて，正答率の計算に非常に時間がかかるため，非常に小さな属性数や訓練事例数に対する解析しか行なうことができない．

近年，2項分布を正規分布で近似することにより，2番目の問題の解決をはかる研究が行なわれているが [1]，領域パラメータの値によっては十分な精度の近似が行なえないという問題がある．

本論文では，ブール属性で表現される任意の目標概念に対する最小近傍法の平均的解析を行なう．本解析では，任意の目標概念に対する正答率を，領域パラメータの単純な関数として導出することで，上述した従来の平均的解析の問題点を解決し，汎用性の高い結果を与える．さらに，クラスノイズが最小近傍法の正答率に与える影響を明らかにする．

2. 問題設定

本解析では，ブール属性を対象とした任意の目標概念を扱う．属性数を m，クラス数を k で表す時，目標概念 φ は，以下で表される．

$$\varphi : \{0,1\}^m \to \{1,2,\ldots,k\} . \qquad (1)$$

目標概念 φ によるクラスの割り当てと属性値とが独立な場合，その属性を φ に対する非関連属性と呼び，そうでない場合，関連属性と呼ぶ．

最小近傍法は，ある確率分布 $\mathcal{D}$ に従って独立に与えられた訓練事例を全てメモリ中に格納する．テスト事例が $\mathcal{D}$ に従って与えられた時，ある距離関数を用いて，テスト事例に対する最小近傍事例をメモリ中から選択し，最小近傍事例のクラスラベルをテスト事例のクラスとして分類する．ここで，最小近傍事例を選択する場合のタイブレイクはランダムに行なわれるとする．

3. 正答率関数の導出

本節では，n 個の訓練事例が与えられた時の，目標概念 φ に対する最小近傍法の正答率関数 $A(n)$ を理論的に導出する．本節の解析では，確率分布 $\mathcal{D}$ として一様分布を仮定し，ノイズは存在しないとする．また，最小近傍法は，距離関数としてハミング距離 h を用いるとする．

任意の目標概念 φ に対する正答率関数を導出するために，同じクラスに属し，距離 d をもつ事例のペアの数 $\psi_\varphi(d)$ の特徴つけを行なう．任意の事例のペアを $\langle s,t \rangle$ で表す時，$\psi_\varphi(d)$ は，次式で表現できる．

$$\psi_\varphi(d) = \sum_{\langle s,t \rangle} g_\varphi(s,t,d) . \qquad (2)$$

Extended Average-Case Analysis of Nearest Neighbor Classifier
Seishi Okamoto and Nobuhiro Yugami
Fujitsu Laboratories Ltd.
E-mail: {seishi, yugami}@flab.fujitsu.co.jp

ここで，$g_\varphi(s,t,d)$ は，次式で定義される．

$$g_\varphi(s,t,d) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } h(s,t)=d \text{ and } \varphi(s)=\varphi(t) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

次に，ある事例が与えられた時に，その事例からの距離が d よりも大きい事例が，事例空間上の一様分布に従って出現する確率 $P_{\text{out}}(d)$ を求める．この確率 $P_{\text{out}}(d)$ は，与えられる事例とは独立に，次式で表現できる．

$$P_{\text{out}}(d) = \frac{1}{2^m}\sum_{e=d+1}^{m}\binom{m}{e}. \quad (4)$$

テスト事例が与えられた時，テスト事例と最小近傍事例の距離が d である確率 $P_{\text{nn}}(d)$ は，確率 $P_{\text{out}}(d)$ を用いて，以下で与えられる．

$$P_{\text{nn}}(d) = P_{\text{out}}(d-1)^n - P_{\text{out}}(d)^n. \quad (5)$$

目標概念 φ の特徴つけ $\psi_\varphi(d)$ と，確率 $P_{\text{nn}}(d)$ を用いて，最小近傍法の正答率関数 $A(n)$ は，次の定理1で与えられる．

定理 1.

$$A(n) = \sum_{d=0}^{m} P_{\text{nn}}(d)\frac{\psi_\varphi(d)}{2^m\binom{m}{d}}. \quad (6)$$

定理1から，正答率 $A(n)$ の計算量は，訓練事例数 n とほぼ独立であることが分かる．すなわち，定理1は，非常に大きな訓練事例数に対しても，最小近傍法の平均的挙動を解析できることを示している．

次に，非関連属性数も正答率 $A(n)$ の計算量とほぼ独立であることを示す．ここで，関連属性数を r，非関連属性数を i で表し（$m=r+i$），任意の事例から全ての非関連属性を取り除いた事例の集合を $\mathcal{S}^r$ で表す．また，同じクラスに属し，距離 d をもつ $\mathcal{S}^r$ 中の事例のペアの数を $\psi_\varphi^r(d)$ で表す．非関連属性は，目標概念 φ によるクラスの割り当てとは独立であることから，$\psi_\varphi(d)$ と $\psi_\varphi^r(d)$ の関係は，次式で与えられる．

$$\psi_\varphi(d) = 2^i\sum_{j=0}^{i}\binom{i}{j}\psi_\varphi^r(d-j). \quad (7)$$

すなわち，次の定理2が成立する．

定理 2.

$$A(n) = \sum_{d=0}^{r+i} P_{\text{nn}}(d)\sum_{j=0}^{i}\binom{i}{j}\frac{\psi_\varphi^r(d-j)}{2^r\binom{r+i}{d}}. \quad (8)$$

4. クラスノイズの影響

本節では，クラスノイズの影響を受ける前の最小近傍法の正答率 A と，受けた後の正答率 A_c の関係を示し，クラスノイズが正答率に与える影響を解析する．ここで，クラスノイズは，事例のクラスラベルを c の確率で，他のクラスラベルにランダムに置き換えるとする．

まず，クラスノイズが訓練事例にだけ影響を及ぼし，テスト事例には影響を及ぼさない場合を考える．この場合，A と A_c の関係は，次の定理3で与えられる．

定理 3.

$$A_c = \left(1-\frac{kc}{k-1}\right)A + \frac{c}{k-1}. \quad (9)$$

定理3から，最小近傍法の正答率は，クラスノイズの発生確率の増加に伴って，傾き $(1-kA)/(k-1)$ で線形変化することが分かる．特に，$kA=1$ の場合は，クラスノイズは，最小近傍法の正答率に全く影響を及ぼさないことが分かる．

クラスノイズが訓練事例とテスト事例の両方に影響を及ぼす場合は，次の定理4が成立する．

定理 4.

$$A'_c = \left(1-\frac{kc}{k-1}\right)^2 A + \frac{(2k-kc-2)c}{(k-1)^2}. \quad (10)$$

定理3，定理4は，任意の確率分布，任意の目標概念，任意の訓練事例数 / 属性数，そして任意の距離関数に対して成立するため，ブール属性の問題領域に対する汎用的な結果を示している．

参考文献

[1] Langley, P., and Sage, S. Tractable Average-Case Analysis of Naive Bayesian Classifiers. In *Proc. of 16th ICML*, 1999.

[2] Okamoto, S., and Yugami, N. Theoretical Analysis of the Nearest Neighbor Classifier in Noisy Domains. In *Proc. of 13th ICML*, pp. 355–368, 1996.

[3] Okamoto, S., and Yugami, N. An Average-Case Analysis of the k-Nearest Neighbor Classifier for Noisy Domains. In *Proc. of 15th IJCAI*, pp. 238–243, 1997.

[4] Pazzani, M., and Sarrett, W. A Framefork for Average Case Analysis of Conjunctive Learning Algorithms. *Machine Learning*, *9*, pp. 349–372, 1992.