ノイズを考慮した最小近傍法の理論的解析

4M－5

岡本 青史
（株）富士通研究所

1. はじめに

最小近傍法は，パターン認識の分野に起源を持つ分類手法であり，機械学習や情報検索等の広い分野に応用されている．最小近傍法は理論的にもよく研究されているが（例えば[1]），最小近傍法の正答率がノイズによってどのような影響を受けるかは明らかになっていない．

本論文では，平均的解析の枠組[4]を用いることにより，最小近傍法に対するノイズの影響を解明する．対象とするノイズは，関連属性ノイズ，非関連属性ノイズ，クラスノイズの３つのタイプである．本解析ではまず，既存の平均的解析の枠組を３つのタイプのノイズが扱えるように拡張し，最小近傍法の正答率を理論的に導出する．次に，この導出結果を用いて，各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

目標概念として，以下の m-of-n/l 概念クラスを扱う．

$$C = \{ (a_1, \ldots, a_{n+l}) \mid w_1 a_1 + \ldots + w_{n+l} a_{n+l} \geq m \}$$

ここで，$a_i, w_i \in \{0,1\}$，$\|\{a_i \mid w_i = 1\}\| = n$ であり，$w_i = 1$ である属性 a_i を関連属性といい，$w_i = 0$ である属性 a_i を非関連属性という．

任意の関連属性，非関連属性はそれぞれ確率 p, q で値１をとるとする．任意の例はこれらの確率に基づいて例空間上から独立に与えられる．

本論文では，以下の特徴を持つ最小近傍法を扱う．

- 訓練例は重複を許して全て格納
- 例間の距離はハミング距離で定義
- 最小近傍例の選択におけるタイブレイクはランダム

3. 正答率関数

N 個の訓練事例が与えられた場合，最小近傍法がテスト例を正しく分類する確率（正答率）を理論的に導出する．正答率は，ドメインを定義する表１中のパラメータの関数として表現される．本解析では，ノイズが訓練例だけに影響を及ぼす場合と，訓練例とテスト例の両方に影響を及ぼす場合の２通りの正答率関数を導出する．前者の場合の正答率関数は以下のように表現できる．

$$A = \sum_{y=0}^{l} \left\{ \sum_{x=0}^{m-1} P_{\mathrm{occ}}(x,y)\,(1 - P_{\mathrm{pos}}(N,x,y)) + \sum_{x=m}^{n} P_{\mathrm{occ}}(x,y) P_{\mathrm{pos}}(N,x,y) \right\} \quad (1)$$

Theoretical analysis of nearest neighbor method in noisy domains, Seishi Okamoto, Fujitsu Laboratories Ltd., e-mail: seishi@flab.fujitsu.co.jp

N	:	訓練事例数
n	:	関連属性数
l	:	非関連属性数
m	:	閾値
p	:	関連属性が値１を持つ確率
q	:	非関連属性が値１を持つ確率
σ_r	:	関連属性ノイズの発生確率
σ_i	:	関連属性ノイズの発生確率
σ_c	:	クラスノイズの発生確率

表 1: パラメータ

ここで，$P_{\mathrm{occ}}(x,y)$ は，１の立っている関連属性数，非関連属性数がそれぞれ x, y 個であるテスト例の出現確率を表しており，以下で求められる．

$$P_{\mathrm{occ}}(x,y) = \binom{n}{x}\binom{l}{y} p^x (1-p)^{n-x} q^y (1-q)^{l-y} \quad (2)$$

また，式(1)中の $P_{\mathrm{pos}}(N,x,y)$ は，N 個の訓練例が与えられた場合，１の立っている関連属性数，非関連属性数がそれぞれ x, y であるテスト例が正のクラスに分類される確率を表している．我々の解析では，既存の平均的解析の枠組を３つのタイプのノイズが扱えるように拡張することで，$P_{\mathrm{pos}}(N,x,y)$ を計算する．紙面の都合上，正答率関数の導出については割愛するが，導出の詳細については，[3]を参照されたい．

4. ノイズの影響

正答率関数中の各ノイズの発生確率を変化させることで，各ノイズが最小近傍法の正答率に及ぼす影響を解析する．ここで，各ノイズは訓練例だけに影響を及ぼすとし，あるノイズの影響を解析する場合に，他のノイズは発生しないこととする．また，関連属性数 n は５に固定して，解析を行なう．

4.1. 関連属性ノイズ

図１は，$N = 32, p = 1/2, q = 1/2$ と固定した場合の，関連属性ノイズの発生確率に対する最小近傍法の正答率の変化を表している．

図１から，$m = 1$ の場合の正答率は殆んどノイズの影響を受けず，$m = 3$ の場合の正答率は大きく影響を受けることが分かる．すなわち，正答率に対する関連属性ノイズの影響は，閾値 m に強く依存し，非関連属性数 l に殆んど独立である．特に，$\sigma_r = 0.5$ の場合には，同じ m の値を持つ概念に対して，最小近傍法は l の値に

依存することなく等しい正答率を持つことが見れる．このことは，正答率関数によって理論的に証明することが出来る．

4.2. 非関連属性ノイズ

図2は，$N = 64, p = 1/2, q = 1/3$とした場合の，非関連属性ノイズの発生確率に対する最小近傍法の正答率の変化を表している．

図2から，各概念に対する最小近傍法の正答率は，非関連属性ノイズの影響を殆んど受けないことが分かる．特に，$q = 1/2$の場合には，σ_iとqを除く任意の固定された他のパラメータに対して，最小近傍法はσ_iに依存することなく等しい正答率を持つことが，理論的に証明される．

4.3. クラスノイズ

図3は，$N = 32, p = 1/2, q = 1/2$とした場合の，クラスノイズに対する正答率の変化を表している．

図3から，各概念に対する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い，線形的に減少することが分かる．すなわち，最小近傍法はクラスノイズの影響を強く受ける．特に，$\sigma_c = 0.5$の場合，各概念に対する最小近傍法の正答率は，等しく0.5となっていることが見れる．より一般的に，σ_cを除く任意の固定されたパラメータに対して，σ_cに対する正答率を$\mathcal{A}$とする時，$1-\sigma_c$に対する正答率は$1-\mathcal{A}$となることが，正答率関数から理論的に導かれる．

5. おわりに

本論文では，3つのタイプのノイズを考慮し，最小近傍法の正答率に対する各ノイズの影響を平均的解析の枠組を用いて解析した．今後は，本研究と[2]による研究とを融合することにより，k-最小近傍法のノイズによる影響を解明する予定である．

参考文献

[1] Dasarathy. B. (Ed.). *Nearest neighbor (NN) norms: NN pattern classification techniques.* IEEE Computer Society Press, 1991.

[2] Okamoto, S., and Satoh, K. An Average-Case Analysis of k-Nearest Neighbor Classifier. In *Proc. of Int. Conf. on Case-Based Reasoning (LNAI, 1010)*, pp. 253–264, 1995.

[3] Okamoto, S., and Yugami, N. Theoretical Analysis of the Nearest Neighbor Classifier in Noisy Domains. In *Proc. of Int. Conf. on Machine Learning*, pp. 355–368, 1996.

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

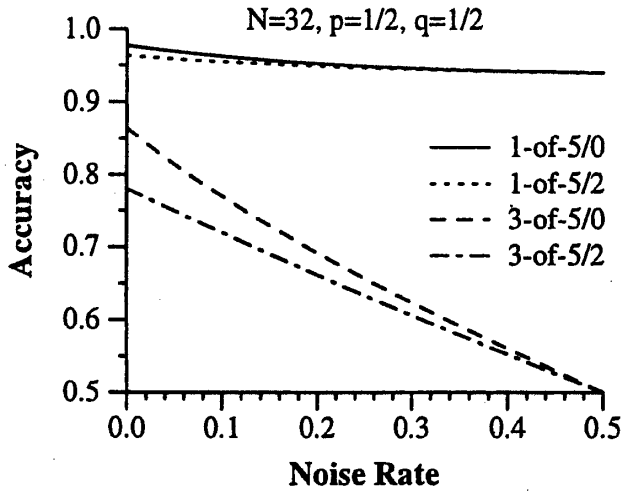

図 1: 関連属性ノイズの影響

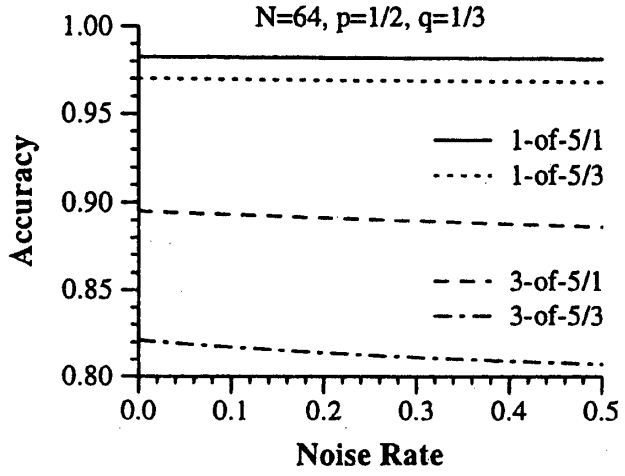

図 2: 非関連属性ノイズの影響

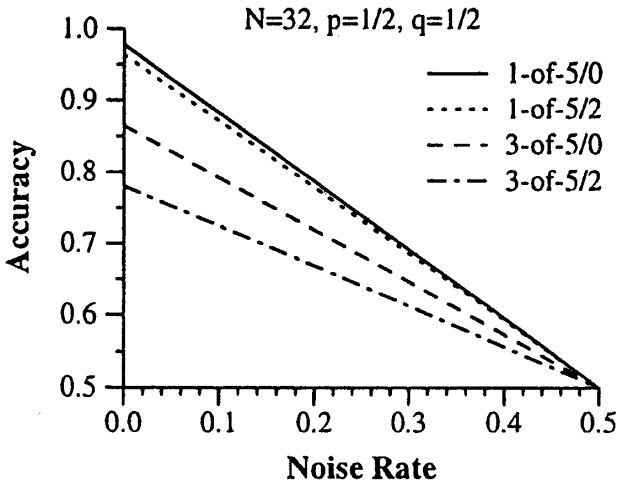

図 3: クラスノイズの影響