近傍に基づく類似事例検索の理論的解析

4AH－2

岡本 青史　　湯上 伸弘
（株）富士通研究所

1. はじめに

近年，事例ベース推論や例に基づく学習に代表されるように，類似事例検索に基づく推論 / 学習の研究が活発に行なわれている．これらの枠組では，現在の問題に類似した過去の事例を検索し，検索された類似事例を直接利用することで，問題に対する解を導出する．ここで，解の妥当性は検索される類似事例に強く依存するため，類似事例をどのように定義するかが大きな問題となる．

最も一般的な類似事例の定義では，類似事例の数 (k) を固定し，現在の問題に最も近い k 個の過去の事例を，問題に対する類似事例と定義する．この定義を用いた学習手法は，k-Nearest Neighbor Method(k-NN) と呼ばれている．また，問題からの距離 d を固定し，問題からの距離が d 以内の事例を類似事例と定義することも出来る．本論文では，この定義を用いた学習手法を，d-Nearest Neighborhood Method(d-NNh) と呼ぶことにする．図1は，2次元空間において，k-NN と d-NNh が形成する，現在の問題に対する近傍を示しており ($k = 5, d = 0.3$)，近傍内の事例が類似事例として検索され，現在の問題に対する解の導出に利用される．

本論文では，分類問題に対する k-NN と d-NNh の平均的解析を行なう．平均的解析とは，事例空間上の確率分布を固定することで，目標概念に対して学習アルゴリズムが正しい答を出力する確率（正答率）を導出し，アルゴリズムの平均的な挙動を解析する理論的枠組である．我々は，3種類のノイズ（関連属性ノイズ，非関連属性ノイズ，クラスノイズ）が扱えるように，平均的解析の枠組を拡張してきた [1]．ここでは，この拡張された平均的解析の枠組を用いて，ノイズを含む問題領域における k-NN と d-NNh の挙動を理論的に比較する．

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 はブール属性，重み w_i は $w_i \in \{0, 1\}$ であり，$\|\{a_i \mid w_i = 1\}\| = n$ である．$w_i = 1$ となる属性 a_i を関連属性，$w_i = 0$ となる a_i を非関連属性と呼ぶ．

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．事例空間上の確率分布は，これらの確率によって規定される．関連属性ノイズ，非関連属性ノイズはそれぞれ，関連属性，非関連属性の値を確率 σ_r, σ_i で反転させ，クラスノイズは，事例のクラスラベルを確率 σ_c で補のクラスに置き代えるとする．

本論文では，k-NN と d-NNh を，類似事例にラベル付けされたクラスの多数決により，現在の問題のクラス (C 又は $\overline{C}$) を決定する学習アルゴリズムとして扱う．ここで，事例間の距離はハミング距離で定義し，類似事例やクラスの決定に関するタイブレイクはランダムに行なうとする．

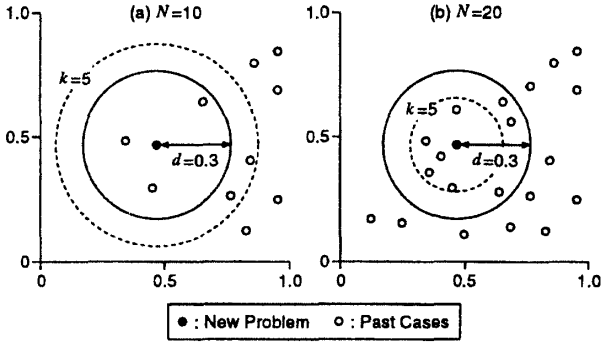

図 1: k-NN と d-NNh が形成する，問題に対する近傍

3. 正答率関数

本解析では，N 個の訓練事例が与えられた場合の k-NN と d-NNh の正答率を，領域パラメータ ($m, n, l, p, q, \sigma_r, \sigma_i, \sigma_c, N$ と，k 又は d) の関数として理論的に導出する．しかしながら，紙面の都合上，本論文では正答率関数の導出を割愛する．k-NN に対する導出は [2] を，d-NNh に対しては [3] を参照されたい．

4. アルゴリズムの挙動

理論的に導出された正答率関数を用いて，k-NN と d-NNh の学習曲線，並びに訓練事例数に対する最適な k と d の値の変化を解析する．ここで，最適な k と d の値とはそれぞれ，k-NN と d-NNh の正答率を最大とする

Theoretical Analysis of Case Retrieval Methods Based on Neighborhood
Seishi Okamoto and Nobuhiro Yugami
Fujitsu Laboratories Ltd.
E-mail: {seishi, yugami}@flab.fujitsu.co.jp

k と d の値のことをいう．本論文では，各ノイズは訓練事例だけに影響を及ぼすとし，$p = q = 1/2$ として，3-of-5/2 概念に対する k-NN と d-NNh の挙動を解析する．紙面の都合上，非関連属性ノイズに対する解析結果は割愛するが，以下の定理が成立する (証明は略)．

定理 1. $q = 1/2$ の場合，m-of-n/l 概念クラスに対する k-NN と d-NNh の正答率は，非関連属性ノイズの発生確率とは独立である．

図 2は，最適な k と d の値をとった場合の，k-NN と d-NNh の学習曲線を示している．図中のエラーバーは，d-NNh に対するモンテカルロシミュレーションの実験結果 (100 個の訓練セットに対する正答率の 95% の信頼区間) を表している．理論的結果は，実際の実験結果と良く一致している．また，k-NN と d-NNh は，関連属性ノイズ，クラスノイズの発生確率に拘らず，ほぼ同じ学習曲線を示している．このことは，k-NN と d-NNh がほぼ同じ学習能力を有することを示唆している．

図 3は，訓練事例数に対する最適な k と d の値の変化を示している．いずれのノイズに対しても，訓練事例数の増加に伴って，最適な k の値は，ほぼ線形的に大きくなっているのに対し．最適な d の値は，殆んど一定の値をとっている．すなわち，k-NN における最適な k の値を得るためには，事例数の増加に伴って，k の値を線形的に大きくしていく必要がある．一方，d-NNh の場合には，事例数が小さい時に d の値を適切に選択することによって，事例数の増加に伴う d の値を殆んど変更する必要がない．このことは，k-NN と d-NNh を実際のシステムに適用する場合，事例数の増加に伴う k と d の値の管理に関して，d-NNh が k-NN に対する大きな利点を持つことを示唆している．

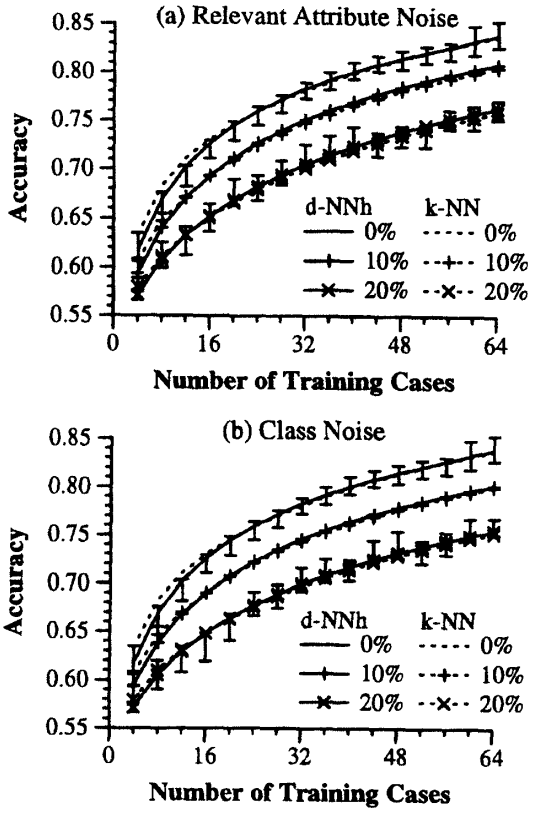

図 2: 学習曲線

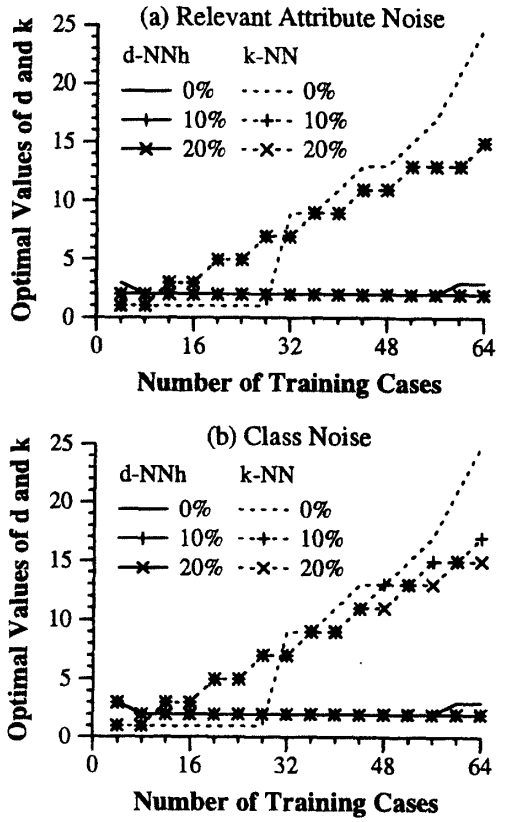

図 3: 訓練事例数に対する最適な k と d の値

参考文献

[1] 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.

[2] Okamoto, S., and Yugami, N. An Average-Case Analysis of the k-Nearest Neighbor Classifier for Noisy Domains. to appear in *Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI-97)*, 1997.

[3] Okamoto, S., and Yugami, N. Theoretical Analysis of Case Retrieval Method Based on Neighborhood of a New Problem, to appear in *Proc. of Int. Conf. on Case-Based Reasoning*, 1997.