The 31st Symposium on Information Theory
and its Applications (SITA2008)
Kinugawa, Tochigi, Japan, Oct. 7–10, 2008

局所空間におけるナイーブベイズ法を用いた分類手法の改良

An Improved Classification Method of Locally Weighted Naive Bayes

餅原　道元*　　石田　崇†　　平澤　茂一*
Dogen MOCHIHARA　　Takashi ISHIDA　　Shigeichi HIRASAWA

Abstract— The K-Nearest Neighbor(K-NN) has been widely used as an effective classification model that decides class label with majority voting of K nearest neighbor samples. In this case class labels of data are equally treated. As an improved version of this algorithm, the Locally Weighted Naive Bayes(LWNB) method has been studied which selects the neighbors of the test instance using Euclidean metric, and then builds the Naive Bayes model in the local neighborhood without simple majority voting. In this paper, we propose an improved classification method of the Locally Weighted Naive Bayes.

Keywords— K-Nearest Neighbor, Locally Weighted Naive Bayes, Classification, Naive Bayes

1　はじめに

K-NN(K-Nearest Neighbor) 法は，距離関数により求めたデータ間の類似度を用いて近傍 K 個を選び，それら近傍 K 個の多数決によりテストデータのクラス決定を行う分類手法であり，分類問題において広く使われている [1].

しかし，K-NN 法ではクラスの確率推定に単純な多数決を用いているという問題点がある．より正確な確率推定を行うために，E.Frank らにより単純な多数決の代わりにナイーブベイズ法を用いる Locally Weighted Naive Bayes(LWNB)[2] が提案されている．

LWNB 法はユークリッド距離を用いてテストデータの近傍 K 個を選び，それらの近傍により形成される局所空間においてナイーブベイズ分類器を構築してテストデータを分類する手法である．ナイーブベイズ法では要素間に独立性を仮定しているため属性間の依存性を考慮できないが，局所空間においてナイーブベイズ分類器を構築することで，属性間の依存性の影響を弱めることができる．

本研究では，LWNB 法の局所空間におけるクラス推定を改良した手法を提案する．離散値・連続値・時系列データなど多種のデータを含む UCI データセット [3] による実験により，LWNB 法と比較して提案手法の精度が向上することを示す．

* 〒 169–8555 東京都新宿区大久保 3–4–1 早稲田大学創造理工学研究科経営システム工学専攻，Major in Industrial and Management Systems Enginnering, Graduate School of Creative Science and Engineering,Waseda University, 3–4–1, Okubo, Shinjuku-ku, Tokyo, 169–8555 Japan. E-mail: motihara@hirasa.mgmt.waseda.ac.jp

† 〒 169–8050 東京都新宿区西早稲田 1–6–1 早稲田大学メディアネットワークセンター，Media Network Center,Waseda University, 1-6-1, Nishiwaseda, Shinjuku-ku, Tokyo, 169–8050 Japan.

2　準備

2.1　K-NN 法

K-NN 法は距離関数により求めたデータ間の類似度を用いて近傍 K 個を選び，それら近傍 K 個の多数決によりテストデータのクラス決定を行う分類手法である．一般的には以下の式で定義されるユークリッド距離を用いる．ただし，$a_i, b_{i'}$ はデータ，m は属性数，a_{ij} はデータ a_i の j 番目の属性値である．

$$d(a_i, b_{i'}) = \sqrt{\sum_{j=1}^{m} (a_{ij} - b_{i'j})^2} \quad \text{for } i, i' \in N. \tag{1}$$

2.2　ナイーブベイズ法

ナイーブベイズ法は，学習データ集合を用いてクラスを推定することによってテストデータを分類する手法である．テストデータ集合を $\boldsymbol{A} = \{a_1, a_2, \ldots, a_t\}$，学習データ集合を $\boldsymbol{B} = \{b_1, b_2, \ldots, b_n\}$，データ $a_i (1 \leq i \leq t), b_{i'} (1 \leq i' \leq n)$ の特徴ベクトルを $\boldsymbol{a_i} = (a_{i1}, a_{i2}, \ldots, a_{im})$, $\boldsymbol{b_{i'}} = (b_{i'1}, b_{i'2}, \ldots, b_{i'm})$ とする．クラス $c_1, c_2, \ldots, c_l \in C$ があるとき，テストデータ a_i を以下の式により求まるクラスに分類する．

$$\arg\max_{1 \leq q \leq l} P(c_q|a_i) = \arg\max_{1 \leq q \leq l} \frac{P(c_q)P(a_i|c_q)}{P(a_i)}. \tag{2}$$

ここで，$P(a_i)$ は大小比較に影響を及ぼさないので，

$$\arg\max_{1 \leq q \leq l} P(c_q|a_i) = \arg\max_{1 \leq q \leq l} P(c_q)P(a_i|c_q). \tag{3}$$

となる．$P(c_q), P(a_i|c_q)$ は最尤推定するなどして求める．

なお，ナイーブベイズ法は要素間に独立性を仮定しているため属性間の依存性を考慮できない．また，パラメータの数が少ないため比較的少数の学習データで学習が可能である．

3　従来手法

3.1　LWNB 法

LWNB 法 [2] はユークリッド距離を用いてテストデータの近傍を選び，それらの近傍により形成される局所空間においてナイーブベイズ分類器を構築してテストデータを分類する手法である．LWNB 法ではテストデータか

らの距離に応じて学習データに重みが割り当てられる．各学習データに割り当てられる重み $w_i^{'}$ は以下の式 (4)〜式 (6) で表されるようにテストデータからの距離が遠いほど小さくなる．ただし，d_h はテストデータ a_i と，テストデータからのユークリッド距離の上位から h 番目の学習データとのユークリッド距離，n は学習データ全体の個数である．

$$w_i^{'} = \frac{w_i \times K}{\sum_{h=1}^{n} w_h}, \tag{4}$$

$$w_h = f(\frac{d_h}{d_k}), \tag{5}$$

$$f(y) = 1 - y \qquad \text{for } \ y \in [0,1]. \tag{6}$$

式 (4) の重み $w_i^{'}$ を用いてテストデータを以下の式 (7) により求まるクラスに分類する．ただし，$c(b_i^{'})$ はデータ $b_i^{'}$ のクラス，$I(x=y)=1, I(x \neq y)=0$ である．

$$\arg\max_{1\leq q\leq l} P(c_q|a_i) = \arg\max_{1\leq q\leq l} \frac{P(c_q)\Pi_{j=1}^{m} P(a_{ij}|c_q)}{\sum_{q=1}^{l}[P(c_q)\Pi_{j=1}^{m} P(a_{ij}|c_q)]}, \tag{7}$$

$$P(c_q) = \frac{1 + \sum_{i'=1}^{n} I(c(b_{i'}) = c_q) w_i^{'}}{l + \sum_{i=1}^{n} w_i^{'}}, \tag{8}$$

$$P(a_{ij}|c_q) = \frac{1 + \sum_{i'=1}^{n} I(a_{ij} = b_{i'j}) I(c(b_{i'}) = c_q) w_i^{'}}{m + \sum_{i'=1}^{n} I(a_{ij} = b_{i'j}) w_i^{'}}. \tag{9}$$

以下に従来手法のアルゴリズムを説明する．

[**従来アルゴリズム**]

1. 式 (1) のユークリッド距離 $d(a_i, b_{i'})$ を用いてテストデータからの距離の上位 K 個を選び近傍 K 個とする．

2. テストデータの近傍 K 個を用いて式 (7) により求まるクラスにテストデータを分類する．

3. テストデータ集合が空集合でなければ **1.** へ．テストデータ集合が空集合であればアルゴリズムを終了する．

□

LWNB 法は良好な性能を持ち，この研究分野において研究成果は数多く参照されている．本研究でもベースラインシステムとして比較の対象とする．

4 提案手法

4.1 提案の概要

LWNB 法では全てのテストデータにおいて K 近傍（K は定数）を選択し，それら K 近傍を用いてナイーブベイズ法によりテストデータを分類している．しかし，各テストデータでテストデータ周辺に分布している学習データは異なるため最適な K の値も各テストデータで異なると考えられる．そこで各テストデータで最適な近傍の個数 $r^*(r^* \leq K)$ を推測し，推測された r^* を用いてナイーブベイズ分類器を構築してテストデータを分類する r^*-LWNB 法を提案する．

r^* を推測する手法として，事後確率による距離関数である SF2[4] を利用する．SF2 は R.D.Short 及び K.Fukunaga ら [5] により提案された距離関数 SF を，J.P.Myles[4] らがマルチクラスに拡張した距離関数である．SF 及び SF2 を式 (10)(11) に示す．

$$SF(a_i, b_{i'}) = |P(c_1|a_i) - P(c_1|b_{i'})|, \tag{10}$$

$$SF2(a_i|b_{i'}) = \sum_{q=1}^{l} |P(c_q|a_i) - P(c_q|b_{i'})|. \tag{11}$$

事後確率を用いたマルチクラスでの分類の場合，式 (11) が最小となる場合に分類誤り率が最小となることが示されている [4][5]．そこでユークリッド距離を用いて K 近傍を選んだ後，更に SF2 の値が大きいデータを取り除くことで r^* 近傍に絞る手法を考える．

4.2 距離関数の性質

テストデータの近傍 K 個それぞれの学習データに対する SF2 の値を昇順に並び替えた場合の SF2 の r 番目 $(1 \leq r \leq K)$ の値を $SF2(r)$ とする．r を 1 から順に増やしていったとき，$SF2(r)$ の増加率が最小となることは $SF2(r+1)$ の値が大きくなることを表すと考えられる．SF2 の値が大きいデータはテストデータに近い性質を持っているとは考えにくいため近傍の候補から除く．すなわち，$SF2(r)$ の増加率が最小となる場合には r 番目までのデータを用いてテストデータを分類することを考える．

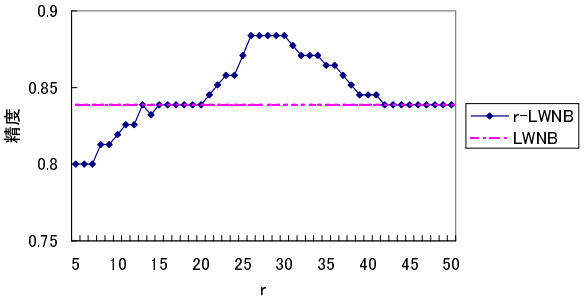
SF2 との増加率の最小値を求める際に以下の式 (12) を用いる．式 (12) は横軸に r，縦軸に $SF2(r)$ をとったグラフにおいて，点 $(1, SF2(1))$ との傾きが最も小さくなる点 $(r^*, SF2(r^*))$ を求める式である．

$$r^* = \arg\max_{1\leq r\leq K} \frac{SF2(r) - SF2(1)}{r-1}. \tag{12}$$

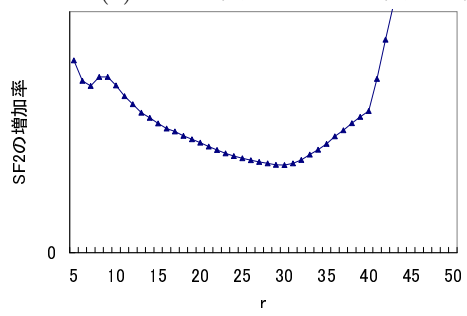
4.3 手法の実験による検証

以上の仮定を実証するため，UCI データセット [3] に含まれる hepatitis データセットを用いて予備実験を行った．図 1 の (a) は LWNB 法 (K=50) と，K 近傍内にて更に SF2 により r 近傍に絞った後，ナイーブベイズ分類器を構築して分類する手法 (r-LWNB 法) との精度の比較を表すグラフである．また，(b) は全てのテストデータの $SF2(r)$ の平均を求めて r で割った値を，全ての r について求めたグラフである．

図 1 より，(a) のグラフにおける r-LWNB 法の最適な r は，全てのテストデータにおいて式 (12) で得られる r^* の平均値に近いことが (b) のグラフからわかる．また，他のデータセットについても同様の結果が得られた．よって図 1 より，精度と SF2 の増加率に関する仮定が成り立つと考えられる．



(a)LWNB 法と r-LWNB 法の比較



(b)SF2 の増加率

図 1: 精度と SF2 の増加率の関係

4.4 提案アルゴリズム

以上を踏まえて以下に提案手法のアルゴリズムを説明する．

[**提案アルゴリズム**]

1. 式 (1) のユークリッド距離 $d(a_i, b_{i'})$ を用いてテストデータからの距離の上位 K 個を選び近傍 K 個とする．

2. テストデータの近傍 K 個のデータに対して SF2 を計算してこれを昇順に並び替え，r 番目までの SF2 の和 $SF2(r)$ の増加率が最小となる r^* を式 (12) により求める．

3. テストデータの近傍 r^* 個を用いて式 (7) によりナイーブベイズ分類器を構築してテストデータのクラスを決定する．

4. テストデータ集合が空集合でなければ **1.** へ．テストデータ集合が空集合であればアルゴリズムを終了する．

□

5 実験結果及び考察

5.1 実験条件

実験は UCI データセット [3] の中から選んだ 29 のデータセットを用いて評価した．これら 29 のデータセットは広く共通して使われているデータマイニングツール Weka [6] に組み込まれているデータセットである．また，精度は一つ抜き法 (データ集合からデータを 1 つ選んでテストデータに使用し，残りを学習データに使用する．この手順を全てのデータがテストデータとして使用されるまで繰り返す手法) により求め，近傍の個数を表す K は LWNB 法における最適な値を用いた [2]．なお，本論文での提案手法は精度と SF2 の増加率の関係を大まかに捉える手法であるため，近傍の個数が K のときのみ LWNB 法の精度が最適となるようなデータセット (sonar，vowel，balance-scale，waveform-5000) は除いた．

5.2 実験結果と考察

表 1 に正しく分類された割合の実験結果を示す．また，図 1(a) と同様の実験を行った場合に r の値が 40 から 49 のときに LWNB 法 (K=50) の精度を上回っている場合があるデータセットのみの実験結果を表 2 に示す．

表 1: 分類精度（K=50）

データセット	K-NN	NB	LWNB	r^*-LWNB $r^* \geq 40$
anneal	0.903	0.867	0.934	**0.947**
audiology	0.770	0.841	0.827	0.823
autos	0.532	0.580	0.605	0.610
balance-cancer	0.727	0.699	0.703	**0.717**
breast-w	0.936	0.956	0.938	0.938
colic	0.815	0.611	0.712	0.707
credit-a	0.852	0.746	0.832	0.836
credit-g	0.715	0.701	0.709	0.717
diabetes	0.669	0.660	0.715	0.715
glass	0.533	0.533	0.551	**0.589**
heart-c	0.825	0.785	0.759	0.752
heart-h	0.793	0.670	0.741	0.741
heart-statlog	0.830	0.767	0.793	0.800
hepatitis	0.800	0.819	0.839	0.839
hypothyroid	0.927	0.925	0.915	0.915
ionosphere	0.849	0.880	0.795	0.798
iris	0.820	0.907	0.927	0.933
kr-vs-kp	0.864	0.630	0.896	**0.941**
labor	0.684	0.772	0.772	**0.807**
lymph	0.804	0.709	0.777	0.784
mushroom	0.998	0.999	0.999	1.000
primary-tumor	0.407	0.248	0.442	0.442
segment	0.865	0.909	0.910	0.913
sick	0.964	0.937	0.961	0.961
soybean	0.659	0.933	0.900	0.896
splice	0.817	0.519	0.944	0.946
vehicle	0.655	0.616	0.675	0.668
vote	0.908	0.901	0.940	0.940
zoo	0.535	0.436	0.644	0.644
平均	0.775	0.743	0.799	**0.804**

1. 実験結果について

SF2 を利用して各テストデータ毎に近傍のデータの個数を K 個から r^* 個に絞ることを考慮した結果，提案 r^*-LWNB 法の分類精度の平均値は K-NN 法及びナイーブベイズ法を上回り，LWNB 法をわずかに上回った．また 29 のデータセットのう

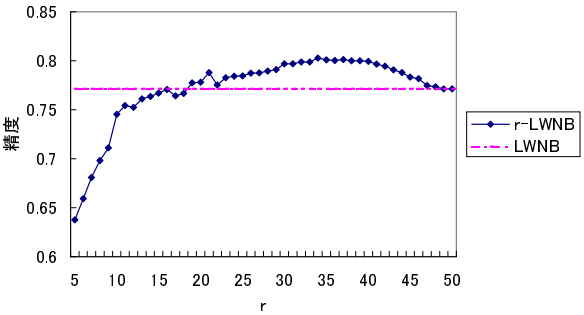
表 2: 分類精度（K=50）

データセット	K-NN	NB	LWNB	r^*-LWNB $r^* \geq 40$
anneal	0.903	0.867	0.934	**0.947**
glass	0.533	0.533	0.551	**0.589**
hepatitis	0.800	0.819	0.839	0.839
kr-vs-kp	0.864	0.630	0.896	**0.941**
labor	0.684	0.772	0.772	**0.807**
平均	0.757	0.724	0.798	**0.825**

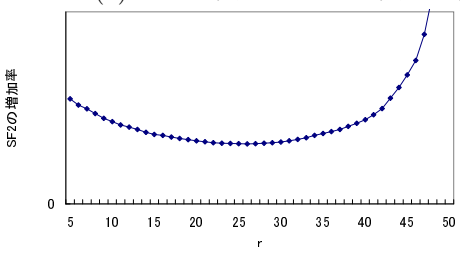
ち，5 つのデータセット (anneal，balance-cancer，glass，kr-vs-kp，labor) で LWNB 法の精度を大きく上回った．

2. r^*-LWNB 法について

(1) まず LWNB 法の精度を上回った 5 つのデータセットについて考察する．これら 5 つのデータセットを用いて図 1 のように精度と SF2 の増加率の関係についての実験を行ったところ，図 2 の (a) において r の値が 40〜49 の間で r-LWNB 法の精度が LWNB 法の精度を上回っており，かつ (b) において r の値が 45 以下で最小となっている点で共通していた．よって多数のテストデータにおいて，推測された r^* の値が最適となったため精度が向上したと考えられる．図 2 は 5 つのデータセットにおける精度と SF2 の増加率の関係についての平均値を求めたグラフである．



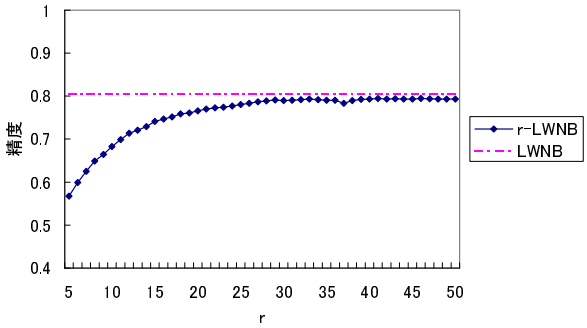
(a)LWNB 法と r-LWNB 法の比較



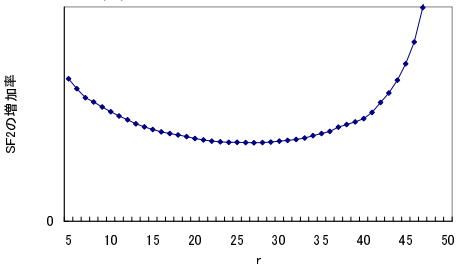
(b)SF2 の増加率

図 2: 精度と SF2 の増加率の関係 (5 つのデータセット)

(2) 次に LWNB 法と精度がほぼ等しくなった残り 24 のデータセットについて考察する．図 1 のように精度と SF2 の増加率の関係についての実験を行った．図 3 の (a) において r の値が 40〜49 の間で r-LWNB 法の精度と LWNB 法の精度とほぼ等しくなる場合では，r の値が 40〜49 のどの値になっても r^*-LWNB 法の精度と LWNB 法の精度がほぼ等しくなるため精度は変わらないと考えられる．実験をして検証した結果，24 のデータセットは上記のパターンに当てはまったために精度が向上しなかったことがわかった．図 3 は 24 のデータセットにおける精度と SF2 の増加率の関係についての平均値を求めたグラフである．



(a)LWNB 法と r-LWNB 法の比較



(b)SF2 の増加率

図 3: 精度と SF2 の増加率の関係 (24 のデータセット)

3. 各テストデータの r^* の値について

各テストデータでの最適な r の値である r^* を推測する際の r^* の探索範囲を 40 以上としたときに精度が最も良くなった．これは r の値が 1〜20 のような小さい値になるとナイーブベイズ法を用いた確率推定が十分にできなくなること，また r の値が 20〜40 では SF2 の増加率がほぼ横ばいとなる場合があり，精度と SF2 の増加率の関係を正確に表すことができなくなったからと考えられる．

6 まとめと今後

本研究では LWNB 法における近傍を選択する手法を改良した r^*-LWNB 法を提案した．テストデータ毎に近傍の学習データ K 個との距離 SF2 を計算する必要があ

るが，これにより従来手法である LWNB 法よりもよい結果を示すことができた．

今後の課題として，より正確に最適な r を選択する手法の考案，例えば，精度と SF2 の増加率の関係をより正確に求めることができる手法の考案などが挙げられる．

参考文献

[1] D.Aha, *Lazy learning*, Kluwer Academic Publishers, 1997.

[2] Frank, E., Hall, M., Pfahringer, B, *Locally Weighted Naive Bayes*, Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp.249-256, 2003.

[3] Merz, C., Murphy, P., Aha, D., *UCI Repository of Machine Learning Databases*, Dept of ICS, University of California, Irvine, 1997. http:www.ics.uci.edu/mlearn/MLRepository.html

[4] Myles, J. P., Hand, D. J, *The multi-class metric problem in nearest neighbour discrimination rules*, Pattern Recognition, 23(11), pp.1291-1297, 1990.

[5] Short, R. D., Fukunaga, K., *The Optimal Dsitance Measure for Nearest Neighbour Classification*, IEEE Transactions on Information Theory, Vol.27, pp.622-627, 1981.

[6] I.H.Witten and E.Frank., *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2005. http:www.ics.uci.edu/mlearn/MLRepository.html