近傍事例集合の分布密度を用いたMultiple-Instance学習

川村俊樹[†1] 上原邦昭[†1]

通常の教師あり学習では事例とラベルが1対1に対応付けられているが，現実のアプリケーションでは，1対1のラベル付けは不可能な場合がある．Multiple-Instance学習はこのような状況の問題を取り扱うために考案された学習手法である．Multiple-Instance学習では，ラベルは個々の事例にはつかず，事例集合にのみつく．このため，通常の教師あり学習より制約が弱く，多くの問題を扱えるようになる．逆に，学習は困難な問題となる．本論文では，「近傍事例集合密度による正事例らしさ」と「事例集合の領域定義」の2つを組み合わせた手法を提案する．具体的には，事例集合ごとに各事例の正事例らしさを求め，それらを用いた事例集合の領域の重ね合わせによって，正事例が多く含まれうる領域を求める．さらに，この領域の重なりから事例集合のラベル推定を行う手法を提案する．最後に，人工データとベンチマークデータセットによって提案手法の有効性を示す．

Multiple Instance Learning by Distribution Density of Neighbor Sets of Instances

Toshiki Kawamura[†1] and Kuniaki Uehara[†1]

Multiple-instance problems arise from the situations where training class labels are attached to sets of samples (named bags), instead of individual samples within each bag (called instances). Common single-instance learning algorithms can hardly good performance when being applied to multiple-instance problems directly. We present a new multiple-instance learning method that combines a measure of the intersection of the positive bags minus the union of the negative bags and weights by the density of neighbor positive bags. We present experimental results on artificial data and benchmark datasets.

1. はじめに

Multiple-Instance学習（MIL）は，事例を集合ごとに扱う学習手法である．通常の教師あり学習では，各事例にラベルが与えられるが，現実世界の問題では個々の事例にラベルが付けられず，事例の集合にのみラベル付けがされることがある．このような事例集合をバッグと呼ぶ．バッグのラベルは，含まれる事例がすべて負のとき負例バッグ，1つでも正のとき正例バッグとなる．このように，正例バッグ内には正と負の両方の事例が含まれているため，未知事例の分類は困難な問題と考えられている．

MILは，Dietterichらが定式化した問題[1]であり，薬効予測判定に用いられた．その後，多くの分野で適用されてきた．たとえば，画像をバッグと見なすことによるCBIR[2]や画像分類[3]，バッグと事例をそれぞれWebページとリンクと見なすことによるウェブマイニング[4]などがある．

現在までにMILを解くための様々な手法が提案されている．たとえば，事例が正となる領域（正領域）を超矩形と仮定したaxis-parallel rectangle[1]，負例バッグから離れており，正例バッグが密集している点が正であると仮定したDiverse Density（DD）[5]，DDとEMを組み合わせたEM-DD[6]，Support Vector Machineを Multiple-Instance問題に拡張したmi-SVMと MI-SVM[7]，制約付きの半教師あり学習と見なしたMISSL[8]，バッグごとのk-Nearest Neighborによる Citation-kNN[9]などがある．

本研究では，近傍バッグの密度から計算される正事例らしさと，正例バッグの重複領域を用いた，新たなMIL手法を提案する．MILの困難さは，正例バッグ内の事例は，どれが正事例で，どれが負事例であるか分からないことにある．そこで，近傍の正例バッグ数から，各事例の正事例らしさを求める．たとえば，ある正例バッグに2つの事例が含まれているとき，一方は近くに異なる正例バッグが3つあり，他方は異なる正例バッグが1つしかないとすると，前者のほうが正

†1 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University

事例である可能性が高いという考え方である．求めた正事例らしさは，各事例の重みとして与え，ラベル推定時に利用する．

さらに，以上の考え方を拡張して，ある事例のラベルを推定することを考える．ある事例が特徴空間上で異なる正例バッグの近くにあり，近くに負例バッグがなければ，正ラベルであると考えられる．逆に，近くに負例バッグがあれば負ラベルだと考えられる．また，ある事例が多くの正例バッグに含まれる可能性があるほど，その事例は正ラベルを持つと考えられる．そこで，本研究では正例バッグの領域を定義し，正事例らしさと正例バッグの共通部分領域という概念を組み合わせて，事例のラベルを推定する手法を提案する．

本論文では，まず 2 章で提案手法について具体的に示し，3 章では類似手法である DD と C-kNN について説明する．4 章では提案手法の評価と実験結果を考察し，最後に 5 章で結論と今後の課題について述べる．

2. 近傍バッグの密度分布による分類手法

本章では正事例らしさと正例バッグの領域表現を用いた分類手法を提案する．まず，正事例らしさの重みについて説明する．正例バッグは負例バッグとは違い，正事例と負事例の両方を含んでいるため，各事例が正であるか負であるか分からない．そこで，各事例の正事例らしさを計算すれば，事例ラベルのあいまいさを減らし，分類精度向上に貢献すると考えられる．この正事例らしさは，近傍にある正例バッグの密度から求める．これは，近くに異なる正例バッグの事例があり，負例バッグの事例がないほど，その事例は正である可能性が高いと考えられるためである．求められた正事例らしさは，各事例に重みとして与える．

図 1 に重み付けの例を示す．図 1 は，特徴空間上における事例の位置を示している．異なる記号は異なるバッグの事例を表し，黒星，黒四角，黒三角が負例バッグの事例を，白星と白四角，白三角，白五角が正例バッグの事例を示している．図 1 では，最左の白三角形は近くに白星形，白四角形，白五角形の 3 種類の異なる正例バッグの事例があるが，右下隅にある白三角形は近くには正例バッグの事例がない．このことから，本手法では両者の値をそれぞれ 3，0 とし，これらの値を正例バッグごとに [0, 1] に正規化して事例の重みとしている．

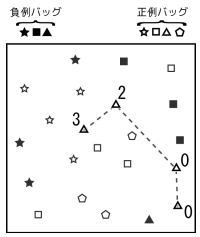


図 1 重み計算の例
Fig. 1 An example of weight calculation.

近傍バッグの決定に用いる距離として，事例間の距離にはユークリッド距離

$$dI(x,y) = \| x - y \| \tag{1}$$

を用い，事例とバッグの距離にはバッグ内で最も近い事例とのユークリッド距離

$$dB(x,b) = \min_j dI(x,b_j) \tag{2}$$

を用いている．なお，x，y は事例であり，b はバッグを示し，b_j は b の j 番目の事例を示している．

重みは，事例 x を中心とし，全負例バッグの事例のうち l 番目に近い事例との距離を半径とした超球形内に存在する正例バッグ数から計算するので，形式的には

$$\mathrm{W}(x) = \frac{1}{Z}\sum_i \sigma_l(x, B_i^+) \tag{3}$$

$$\text{where } \sigma_l(x, B_i^+) = \begin{cases} 1 \text{ if } dB(x, B_i^+) < \underset{y \in \cup_j B_j^-}{\mathrm{lth}} dI(x,y) \\ 0 \qquad \text{otherwise} \end{cases}$$

となる．なお，B_i^+ は i 番目の正例バッグであり，B_j^- は j 番目の負例バッグである．Z は結果が [0, 1] となるように正規化するための定数であり，$\sum_i \sigma_l(x, B_i^+)$ の最大値となる．lth は l 番目に小さな値をとる関数であるため，$\underset{y \in \cup_j B_j^-}{\mathrm{lth}} dI(x,y)$ は，全負例バッグの事例 $\cup_j B_j^-$ のうち x から l 番目に近い事例 y との距離を求めている．最近傍の負事例までででなく，l 番目の負事例までの正例バッグ数を数えているのは，少数の負事例が近傍にあるとしても，その影響を受けにくくするためである．

次に正例バッグの領域表現について定義する．ある正例バッグの事例すべてをあえて正事例と仮定し，負例バッグの事例を負事例として学習させた学習器によって決定できる領域を，その正例バッグの領域とする．この正例バッグの領域は，1 つだけではもちろん負事例も含まれるが，複数の異なる正例バッグの領域の共通部分は，正事例のみからなっている可能性が高い．つまり，この共通部分の領域にある事例は，正事例である可能性が高いと考えられる．このことは後に述べる．

学習器に 1-Nearest Neighbor を用いた場合の正領

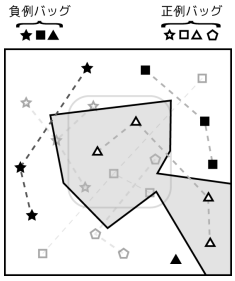


図 2 正例バッグの領域の例
Fig. 2 A example of a positive bag area.

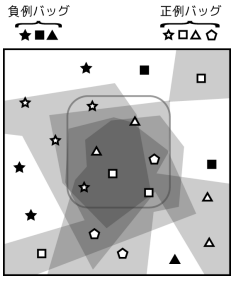


図 3 正例バッグの重なりの例
Fig. 3 An example of duplication positive bag area.

域の例を**図 2** に示す．黒星，黒四角，黒三角の事例を負事例，白三角の事例を正事例として 1-Nearest Neighbor に学習させたとき，正と判定される領域を灰色で示す．つまり，この灰色の部分が白三角の含まれる正例バッグの領域である．ここで，白三角の正例バッグの領域に属する事例は，その正例バッグの影響を受けて，正事例である可能性が高まると考えられる．

以上の正事例らしさの重みと正例バッグの領域により，ある正例バッグからの正事例らしさ（影響）を計算する．形式的には，事例 x が正例バッグ B_i^+ から受ける影響は

$$\mathrm{P}B_i(x) = \mathrm{W}(\operatorname*{argmin}_{y \in B_i^+} dI(x,y)) \cdot kNN(x; B_i^+, B^-) \quad (4)$$

とする．なお，B^- はすべての負例バッグの事例であり，y は B_i^+ 内で x に最も近い事例となる．また，$kNN(x; B_i^+, B^-)$ は，訓練事例 B_i^+，B^-，テスト事例 x，出力が正 (1) または負 (0) の k-Nearest Neighbor（kNN）学習器であり，近傍数 k はパラメータとして与えている．

上記計算式で計算される各正例バッグからの影響を足し合わせると，正領域である可能性が高いほど値が大きくなると考えられる．たとえば，正例バッグの領域を重ね合わせると**図 3** になる．図 3 の真の正領域（角丸四角）内部では，正例バッグの領域が多く重なり合い，各正例バッグからの影響を足し合わせた値が大きくなる．

事例 x における，すべての正例バッグ領域からの影響は，

$$E(x) = \frac{1}{i} \sum_i \mathrm{P}B_i(x) \quad (5)$$

と計算し，$E(x)$ が閾値 t 以上のときは事例 x が正，未満のときは事例 x が負と推定する．また，バッグのラベルは，バッグ内に正と推定した事例が 1 個でも存在すれば正例バッグ，そうでなければ負例バッグと推定する．なお，閾値の決定は以下の手順により求めている．

(1) 訓練データ内のあるバッグ b から $\max_j E(b_j)$ を計算する．
(2) 閾値 $t = \max_j E(b_j)$ と設定し，
  (a) 訓練データをバッグごとに分割する．
  (b) 1 つのバッグをテストデータ，残りを訓練データとし，バッグ内の事例 x が 1 個でも $E(x) > t$ となれば，そのバッグを正例バッグとする．
  (c) それぞれバッグを取り替えて (b) を実行し，正しく分類できたバッグ数を求める．
(3) 全訓練バッグで同様に (1)，(2) を行い，その中で正しく分類できたバッグが最も多かった t を閾値として用いる．

なお，本研究において，$E(x)$ を各正例バッグ領域の影響の積ではなく和によって計算しているのは，正例バッグに属さない領域は $\mathrm{P}B$ が 0 になるためである．積によって計算すると，すべての正例バッグの領域に属する部分領域のみが $E(x) > 0$ となり，それ以外の領域は $E(x) = 0$ となる．これは，すべての正バッグ領域の共通部分領域しか，正領域であると推定できないことを意味する．この共通部分領域は，正領域に比べて小さくなる可能性が高いため，影響の和によって $E(x)$ を求めている．

3. 関 連 研 究

現在までに MIL を解くための手法は数多く存在している．その中で提案手法と類似した手法として，Diverse Density，重み付き多数決による改良 Diverse Density，Citation-kNN について紹介する．

3.1 Diverse Density

Diverse Density（DD）は，ある座標がどれだけ多くの正例バッグの事例に近く，かつどれだけ負例バッグの事例に遠いかの指標を多様性密度として定義して

おり，この多様性密度が高い座標点が正であるという考え方である．

特徴空間の座標 x における多様性密度は，与えられた各バッグからの影響の積によって計算される．バッグから x への影響は，noisy-or model [10] によって評価する．noisy-or model とは，正と負のように否定の関係にあるすべてのデータから，最も確からしい結果を計算する方法である．多様性密度は，形式的には

$$\mathrm{D}D(x) \propto \Big(\prod_j \Pr(x|B_j^+)\prod_k \Pr(x|B_k^-)\Big) \quad (6)$$
$$\Pr(x|B_i^+) = (1-\textstyle\prod_j(1-\Pr(x|B_{ij}^+)))$$
$$\Pr(x|B_i^-) = \textstyle\prod_j(1-\Pr(x|B_{ij}^-))$$

となる．なお，各事例からの影響はガウス分布 $\Pr(x|B_{ij}) = \exp(-\sum_k s_k^2(B_{ijk}-x_k)^2)$ を用いて評価する．k は属性のインデクス，s はスケールファクタである．

バッグの多様性密度は，そのバッグに含まれる事例ごとの多様性密度の最大値

$$\mathrm{D}D(b_j) = \max_j \mathrm{D}D(b_j) \quad (7)$$

とし，この値が大きければバッグは正であり，小さければ負であると判定する．

この手法では，noisy-or model を用いて多様性密度を計算している．noisy-or model は，すべてのデータを用いて，その影響が最大となる点を導き出すため，すべての正例バッグは正領域に属する事例を持つ必要がある．これは，正領域が 1 つであれば問題はない．しかし，正領域が複数ある場合，すべての正例バッグが各正領域に属する事例を持たなければ，正しく推定できない．たとえば正領域が 2 つあり，全正例バッグの半数は一方の正領域に属する正事例のみ持ち，もう半数が他方の正領域に属する正事例のみ持つとき，多様性密度は 2 つの正領域の中央で最大になる可能性が高い．これは，DD は限られた場合のみしか，複数の正領域を推定できないことを意味する．このため，正領域が複数ある場合，精度が低下する可能性がある．

3.2 重み付き多数決としての改良 Diverse Density

この手法は，タンパク質相互作用を知るために，山川ら [11] が DD を改良した手法である．この手法では，$\mathrm{D}D(x)$ におけるバッグからの影響を積から和に変更している．

DD における多様性密度の値は，座標 x の近傍に 1 つでも負例バッグの事例が存在すると強く抑制される．しかし，山川らの研究では，特徴空間内の多くの範囲で負例バッグの事例が散在する場合，そのままでは負例からの影響が強くなりすぎることが分かった．そこで，周辺に少数の負例バッグの事例が存在しても，正例バッグの事例が多数存在する部分領域では正と推定するため，バッグの影響の和によって $\mathrm{D}D(x)$ を求める．形式的には以下のように表せる．

$$\tilde{\mathrm{D}D}(x) \propto \sum_i \mathrm{sign}_i \left[1 - \prod_j (1 - \Pr(x|B_{ij}^{\pm}))\right] \quad (8)$$

ここで sign_i はバッグ i が正例のとき $+$，負例のとき $-$ である．

山川らの手法は，各バッグからの影響の和によってラベル推定を行う点と重みを用いる点で提案手法と類似している．しかしながら，提案手法では正例バッグの密度による重みに対して，山川らの手法では事例からの距離にガウス関数による重みである．このため，山川らの手法では事例間の距離によっては，スケールファクタを変更する必要があるが，提案手法ではパラメータ変更をせずにすむと考えられる．たとえば，事例間の距離が離れているスパースな状況では，山川らの手法ではスケールファクタを小さくしなければならないが，提案手法ではパラメータの変更は必要ない．

3.3 Citation-kNN

C-kNN は，バッグ単位での kNN を用いた lazy-learning 手法である．ただし，通常の kNN とは異なり，reference と citer という関係を用いている．reference とは，未知バッグから見た近傍バッグであり，citer とは，未知バッグが近傍になるバッグである．

各近傍バッグは，ハウスドルフ距離 [12] を拡張した距離関数によって計算する．ハウスドルフ距離とは，距離空間における部分集合の測定関数であり，集合 $A = \{a_1, \ldots, a_m\}$，$B = \{b_1, \ldots, b_n\}$ 間の距離は，以下のように定義される．

$$H(A,B) = \max\{h(A,B), h(B,A)\} \quad (9)$$
$$\text{where } h(A,B) = \max_{a\in A}\min_{b\in B} \| a-b \|$$

ただし，ハウスドルフ距離をそのまま使用すると，集合内に存在する 1 つの外れ値により，結果が大きく変わることがあるため，少数の外れ値による結果の変動を抑えるためにハウスドルフ距離を拡張している．拡張したハウスドルフ距離は，$h(A,B)$ を以下のように再定義している．

$$h_k(A,B) = \mathrm{kth}_{a\in A}\min_{b\in B} \| a-b \| \quad (10)$$

なお，kth は k 番目の最大値を示す．つまり，$k = 1$ のとき最大値をとるので，ハウスドルフ距離と等しくなる．

C-kNN では，テストバッグから reference と citer の関係になるバッグのラベルから，ラベルを推定している．reference の関係になる近傍の c 個のバッグのうち，正例バッグ数を R_p，負例バッグ数を R_n とする．同様に citer の関係になる近傍の c 個のバッグのうち，正例バッグ数を C_p，負例バッグ数を C_n とする．このとき，$R_p + C_p > R_n + C_n$ ならば，正例バッグであると分類し，$R_p + C_p \leq R_n + C_n$ ならば，負例バッグと分類する．近傍数 c はパラメータとして与える．

C-kNN は，バッグ間の距離を拡張したハウスドルフ距離によって求めている．この距離関数は，事例がバッグごとにクラスタ化されている場合は，バッグ間の距離を正しく測定できるが，クラスタ化されていない場合は，正しくは測定できないという問題がある．たとえば，すべてのバッグが特徴空間全体に分散している場合，任意のバッグ間の距離に差はなくなり，バッグ間の距離としては役に立たない．このため，バッグ内の事例が分散している場合には性能が低下する可能性がある．また，負例バッグの各事例のラベルは明らかに負であるにもかかわらず，その情報を利用していない．つまり，負例バッグもバックごとに扱うため，負例バッグ内の少数の事例のみを分類に利用している．このため，負例バッグが少ない状況では，負例バッグ内の全事例を分類に利用する他手法に比べて性能が劣る可能性がある．

4. 性能評価および検討

提案手法の性能評価のため，人工データおよびベンチマークデータセットによる実験を行う．まず，人工データによって，正領域が複数ある場合とバッグ数が少ない場合では，提案手法が優位であることを示す．次に，ベンチマークデータセットによって，既存手法との性能を比較する．また，パラメータを変更したときの性能の変化についても評価する．

4.1 人工データ

人工データにより，類似手法と提案手法の性質の違いを示す．人工データによる比較は，上記の類似手法との性質の違いを明確にするため，現実のデータを考慮して，

(1) 正領域が 1 つ，2 つ，3 つの場合
(2) バッグ内の事例が，バッグごとに集まる場合，バッグに依存しない場合

表 1 人工データセットの条件
Table 1 Assumptions of artificial dataset.

# of bags	50
# of instances per bag	10
# of attributes	2
Distribution	mixture gaussian
# of positive areas	{1,2,3}
Rate of label	{1:5,1:1,5:1}
Label miss	{0%,10%}
Instance relation in bag	{neighborhood, no}

(3) 正例バッグと負例バッグの比率が変化した場合
(4) 誤ったバッグラベルがある場合

について行う．

人工データの条件を**表 1** に示す．固定した条件は，すべてのバッグが数 120，バッグ内の事例数が 10，属性数が 2 次元，事例の分布が混合正規分布である．変化させる条件は，正領域の数が 1，2，3，バッグ内の事例の関係がない場合と互いに近い場合，正例バッグと負例バッグの比率が 1 : 5，1 : 1，5 : 1 の 3 通り[⋆1]，バッグラベルの誤りがない場合とある（10%）場合である．以上の条件を変化させた人工データを作成し，それぞれについて実験を行う．正領域の範囲は，正領域の数により変化させ以下のように設定している．

- 1 つの場合
  - $-0.2 < a_1 < 0.2$, $-0.2 < a_2 < 0.2$
- 2 つの場合
  - $0.2 < a_1 < 0.5$, $0.2 < a_2 < 0.5$
  - $-0.5 < a_1 < -0.2$, $-0.5 < a_2 < -0.2$
- 3 つの場合
  - $-0.45 < a_1 < -0.15$, $-0.4 < a_2 < -0.2$
  - $0.15 < a_1 < 0.45$, $-0.4 < a_2 < -0.2$
  - $-0.45 < a_1 < -0.15$, $0.2 < a_2 < 0.4$

なお，a_1，a_2 は事例の属性値である．

事例の分布は，5 つの二次元正規分布をランダムな割合で混合している．各正規分布の条件は，共分散は 0，平均は $[-0.5, 0.5]$，分散は $[0, 0.4]$ からランダムに決定している．バッグ内の事例の関係については，バッグに依存しない場合は，生成された事例からランダムに選択しバッグを作成しており，バッグごとに集まる場合は，近傍事例によってバッグを作成する．

テストデータとして，訓練データと同じ分布から生成される 1,000 個のデータを用いている．本実験では，正領域の狭さから，テストデータは正事例に比べて負事例が多いので，すべて負と推定しても精度が高くな

⋆1 バッグ数では，それぞれ正 20 負 100，正 60 負 60，正 100 負 60 となる．

表 2 人工データセットによる結果
Table 2 The experimental result: artifical dataset.

relation	rate	noise	one area			two area			three area		
			DD	C-kNN	提案手法	DD	C-kNN	提案手法	DD	C-kNN	提案手法
neighborhood	1:5	なし	66.83	78.60	**95.67**	51.54	74.89	**88.33**	50.31	65.18	**91.67**
		あり	58.42	65.09	**67.11**	36.56	56.50	**63.22**	39.59	56.35	**63.22**
	1:1	なし	61.99	84.38	**95.11**	56.70	74.60	**92.22**	53.79	67.02	**88.89**
		あり	60.40	**69.95**	64.67	58.39	60.24	**62.11**	51.52	**63.32**	62.00
	5:1	なし	76.66	75.83	**92.00**	68.33	69.46	**89.33**	61.49	59.39	**82.33**
		あり	**70.03**	64.08	53.67	49.78	**50.23**	47.33	50.32	**57.82**	56.11
no relation	1:5	なし	67.80	57.52	**91.67**	52.47	49.57	**89.00**	60.53	48.45	**87.67**
		あり	66.83	54.81	**82.89**	45.48	51.74	**80.44**	60.84	48.54	**78.67**
	1:1	なし	63.56	54.14	**92.44**	54.29	54.41	**91.22**	59.47	47.21	**89.78**
		あり	62.91	52.84	**84.33**	47.07	50.89	**81.78**	60.19	49.28	**82.33**
	1:5	なし	75.22	54.08	**76.11**	49.94	52.40	**73.78**	66.97	51.06	**74.67**
		あり	54.49	52.87	**69.89**	51.60	53.47	**68.89**	59.73	46.74	**71.11**

る．このため，分類器の良さを示す Area Under the Curve（AUC）[13] により利得を計算し，性能を比較している．AUC は，Receiver Operating Characteristic 曲線[⋆1]の良さを要約したもので，最大で 1 となり，ランダムな場合は 0.5 となる．形式的には，

$$AUC = \frac{\sum_i r_i - n_p(n_p+1)/2}{n_p n_n} \qquad (11)$$

である．なお，n_p，n_n はテスト事例の正事例と負事例の数であり，r_i は正である確率順に並べたとき i 番目までの正事例数である．本実験における提案手法のパラメータとして，$k=3$，$l=2$ としている．DD のパラメータであるスケールファクタは 1，C-kNN のパラメータは $k=2$，$c=4$ としている．

実験結果を**表 2** に示す．表 2 の relation はバッグ内の事例が互いに近い場合（neighborhood）とバッグ内の事例間に関係がない場合（no relation），rate は正例バッグ数と負例バッグ数の比率，noise はバッグラベルのノイズの有無，one area は正領域が 1 つの場合，two area は正領域が 2 つの場合，three area は正領域が 3 つの場合である．また，数値はすべて AUC（%）を示し，特に最も高かった値を太字で強調している．

DD は正領域が複数の場合に利得が低下しているが，その他の場合はノイズに対する高い耐性が見られる．このことから，DD は正領域が複数の場合は苦手とするが，正領域が 1 つであればノイズへの耐性が高いといえる．このため，あらかじめ正領域が 1 つまたは，すべての正例バッグがそれぞれの正領域に含まれる事例を持つと分かっているとき，高い性能を示すと考えられる．

また，正例バッグ数の比率が高いと，利得が高い傾向にある．これは，多様性密度は負例バッグから遠く，正例バッグが密集する度合いを計算しており，正例バッグが多いほど密集点を計算しやすいため，利得が高くなると考えられる．

C-kNN は，バッグ内の事例が互いに近い場合に高い利得を得ているが，バッグ内の事例がバッグに依存しない場合，利得が大きく低下している．また，正例バッグ数と負例バッグ数の比率が異なると性能が低下する傾向にあり，負例バッグ数が少ない方が低下しやすい傾向が見てとれる．

提案手法は，ほぼすべての場合で最大の性能を示しているが，バッグ内の事例が互いに近く，かつ誤っているバッグラベルが存在すると性能が低下している．しかし，性能が低下している場合でも，他手法とほぼ同等の性能を示している．さらに，他の条件に対しては安定した性能を示している．これらから，本手法は既存手法に比べて優れた性能を示しているといえる．

4.2 ベンチマークデータセット

既存手法との比較を行うため，MIL 用データセットである麝香芳香予測データセットと画像分類データセットによる実験を行う．麝香芳香予測データセット（Musk1，Musk2）は UCI Machine Learning Repository [14] から入手した．Musk2 は Musk1 よりもバッグに含まれる事例数が大きい．また，画像分類として Andrews ら[7] が生成したデータのうち，Elephant，Fox，Tiger を使用した．なお，本実験における提案手法のパラメータは，Musk1 では $k=2$，$l=1$，Musk2 では $k=3$，$l=1$，画像分類データでは $k=4$，$l=4$ とした．

既存手法の結果については各手法を提案した論文よ

⋆1 分類器のパラメータを変化させながら，縦軸に TruePositive/(TruePositive + FalseNegative)，横軸に FalsePositive/(FalsePositive + TrueNegative) をとった曲線．

り引用しているが，EM-DD による結果は，初期値を恣意的にとったために分類精度が向上している可能性があるといわれている[7]．このため，表の分類精度では Andrews らによる結果[7] を使用している．APRs，DD，EM-DD，MI-SVM，mi-SVM は ten-fold cross validation，C-kNN は leave-one-out により検定している．このため，提案手法では両方の検定を利用している．

表 3 に実験結果を示す．アルゴリズムは，検定方法の違いから上下に分けている．上が leave-one-out，下が ten-fold corss validation である．また，数値はすべて分類精度（%）を示し，特に最も高かった分類精度を太字で強調している．

C-kNN との比較では，Musk1，Musk2，Tiger でほぼ同等の精度を示し，Elephant，Fox では高い精度を示している．このため，画像分類データは芳香予測データに比べて，バッグ内の事例間の距離が遠く，クラスタを作っていないと考えられる．また，これらのデータセットは正例バッグ数と負例バッグ数が等しいため，C-kNN に適したデータであると考えられる．このため，C-kNN は高い精度を示していると考えられる．

他の既存手法との比較では，Musk1，Musk2 で APRs に劣るものの，DD とはほぼ同等の精度を示している．APRs は薬効予測問題のために考案されたアルゴリズムであるため，データセットの特性を考慮したアルゴリズムとなっている．このため，データセットとの相性が良く，高い性能を得ていると考えられる．また，DD が他手法に比べて高い精度を示していることから，芳香予測データの正領域は 1 つである可能性がある．一方，Musk2 では精度低下が見られる．このデータセットは 1 つのバッグ内の事例数が多いという特徴を持つため，バッグ内の事例数が多い場合に性能が低下したと考えられる．画像分類のデータセットにおいて，Tiger では MI-SVM に劣るものの，ほぼ最良である．Tiger では正例バッグごとの重なりが少ないため，正例バッグの重なりを調べる手法ではうまく分類できず，精度が低下したと考えられる．これは，C-kNN も MI-SVM に比べて低い精度であることからも考えられる．また，人工データによる実験結果を考慮すると，実世界データはバッグ内の事例がクラスタ化しており，ラベル誤りのあるデータと考えられる．

これらの結果から，本手法の改良点として，バッグ内の事例が多い場合も分類精度が低下しないように改善すべきことが分かった．また，次元数が大きい場合に NN は精度が低下するので，正領域を定義する学習器を SVM に変更して，高次元の問題への対応も検討する必要がある．

4.3 パラメータ変化による実験結果

パラメータ k，l を変化させたときの提案手法の性能変化を調べる．実験方法は，一方のパラメータを 1 に固定し，他方のパラメータを 1 から 7 に変化させている．データ条件は，正例バッグ数と負例バッグ数がそれぞれ 30，バッグごとの事例数が 10，属性数が 2 であり．事例の分布は，5 つの二次元正規分布をランダムな割合で混合している．各正規分布の条件は，共分散は 0，平均は $[-0.5, 0.5]$，分散は $[0, 0.4]$ からランダムに決定している．生成された事例からランダムに選択しバッグを作成している．正領域の範囲は，正領域の数により変化させ以下のとおりに設定している．

- 1 つの場合
  - $-0.2 < a_1 < 0.2$，$-0.2 < a_2 < 0.2$
- 2 つの場合
  - $0.2 < a_1 < 0.5$，$0.2 < a_2 < 0.5$
  - $-0.5 < a_1 < -0.2$，$-0.5 < a_2 < -0.2$

なお，a_1，a_2 は事例の属性値である．

AUC の変化を**図 4** に示す．fix l one，fix l two は，パラメータ l を固定して，パラメータ k を変化させており，それぞれ正領域が 1 つの場合，正領域が 2 つの場合である．また，fix k one，fix k two は，パラ

表 3 ベンチマークデータセットによる結果
Table 3 The experimental result: benchmark dataset.

Algorithm	Musk1	Musk2	Elephant	Fox	Tiger
提案手法	**92.4**	85.3	**88.4**	**67.8**	**80.4**
Citation-kNN	**92.4**	**86.3**	80.5	60.0	78.0
提案手法	90.3	81.6	**83.0**	**63.1**	79.2
APRs	**92.4**	**89.2**	-	-	-
DD	88.9	82.5	-	-	-
EM-DD	84.8	85.8	78.3	56.1	72.1
MI-SVM	81.4	59.4	81.4	59.4	**84.0**
mi-SVM	87.4	83.6	82.2	58.2	78.9

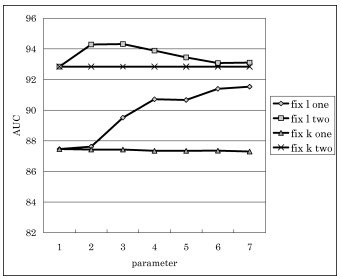


図 4 パラメータ変更時の AUC 変化
Fig. 4 Validation of AUC on parameter change.

メータ k を固定して，パラメータ l を変化させており，それぞれ正領域が 1 つの場合，正領域が 2 つの場合である．

パラメータ k は，正領域が 1 つのときは増やすほど性能が向上しているが，正領域が 2 つのときは 3 で最大となっている．また，パラメータ l は，ほとんど変化がない．このことから，パラメータ l よりもパラメータ k の決定が重要と考えられる．

これらの結果から，近傍正例バッグ密度のパラメータよりも学習器のパラメータが重要であることが分かる．したがって，別の学習器を用いれば，性能向上も期待できる．また，パラメータ l の影響は少ないため，多くの場合に固定して利用できると考えられる．

5. おわりに

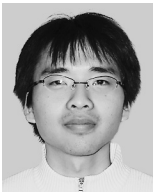
本研究では，バッグにラベルが付与された Multiple-Instance 学習アルゴリズムとして，各事例に近傍バッグから計算した重みを付与し，事例の分類時には正例バッグの領域と重みから分類する手法を提案した．この結果，類似手法である DD や C-kNN より高い性能を示す手法を実現できた．

今後は，各種の条件をともなう人工データを作成して実験を行う予定である．また，今回は正バッグ領域を定義する学習器として kNN を用いたが，今後は異なる学習器を用いることによる特性の変化についても考察する予定である．

参考文献

1) Dietterich, T.G., Lathrop, R.H. and Lozano-Perez, T.: Solving the Multiple Instance Problem with Axis-Parallel Rectangles, *Artificial Intelligence*, Vol.89, No.1-2, pp.31–71 (1997).
2) Yang, C. and Lozano-Perez, T.: Image Database Retrieval with Multiple-Instance Learning Techniques, *Proc. ICDE*, pp.233–243 (2000).
3) Maron, O. and Ratan, A.L.: Multiple-Instance Learning for Natural Scene Classification, *Proc. 15th ICML*, pp.341–349 (1998).
4) Zhou, Z.-H., Jiang, K. and Li, M.: Multi-Instance Learning Based Web Mining, *Appl. Intell.*, Vol.22, No.2, pp.135–147 (2005).
5) Maron, O. and Lozano-Pérez, T.: A Framework for Multiple-Instance Learning, *Proc. NIPS*, pp.570–576 (1998).
6) Zhang, Q. and Goldman, S.A.: EM-DD: An Improved Multiple-Instance Learning Technique, *Proc. NIPS*, pp.1073–1080 (2001).
7) Andrews, S., Tsochantaridis, I. and Hofmann, T.: Support Vector Machines for Multiple-Instance Learning, *Proc. NIPS*, pp.561–568 (2002).
8) Rahmani, R. and Goldman, S.A.: MISSL: Multiple-Instance Semi-Supervised Learning, *Proc. 23th ICML*, pp.705–712 (2006).
9) Wang, J. and Zucker, J.-D.: Solving the Multiple-Instance Problem: A Lazy Learning Approach, *Proc. 17th ICML*, pp.1119–1125 (2000).
10) Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann (1988).
11) 山川　宏，仲尾由雄，丸橋弘治：タンパク質相互作用属性の出現解析とその予測，人工知能学会全国大会（第 20 回）(JSAI-2006) 論文集 (2006).
12) Edgar, G.A.: *Measure, Topology and Fractal Geometry*, Springer (1995).
13) Hand, D.J. and Till, R.J.: A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems, *Mach. Learn.*, Vol.45, No.2, pp.171–186 (2001).
14) Newman, D.J., H.S.B.C. and Merz, C.J.: UCI Machine Learning Repository (2007). http://www.ics.uci.edu/∼mlearn/MLRepository.html

(平成 19 年 8 月 8 日受付)
(平成 19 年 9 月 26 日再受付)
(平成 19 年 11 月 28 日採録)



川村　俊樹

昭和 58 年生．平成 19 年神戸大学工学部情報知能工学科卒業．現在，同大学院自然科学研究科博士前期課程在学中．

上原　邦昭（正会員）

昭和 29 年生．昭和 53 年大阪大学基礎工学部情報工学科卒業．昭和 58 年同大学院博士後期課程単位取得退学．大阪大学産業科学研究所助手，講師，神戸大学工学部情報知能工学科助教授，同都市安全研究センター教授を経て，現在，同大学院工学研究科教授．工学博士．人工知能，特に機械学習，マルチメディア処理の研究に従事．人工知能学会，電子情報通信学会，計量国語学会，日本ソフトウェア科学会，AAAI 各会員．