社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－66 (17)
2007／9／4

近傍事例集合の分布密度を用いたMultiple-Instance学習

川村俊樹† 上原邦昭†

通常の教師あり学習では事例とラベルが一対一に対応づけられているが，現実のアプリケーションでは，一対一のラベル付けが不可能な場合がある．このような問題に対処するために，事例の集合にのみラベルを付与する学習問題が Multiple-Instance 学習である．Multiple-Instance 学習は，個々の事例にラベルが付与されていないため，通常の教師あり学習よりあいまいな表現が可能である，逆に分類は困難な問題となる．本稿では，この問題を解くために既存の Multiple-Instance 学習手法を組み合わせて，重み付けとアンサンブルを用いた手法を提案する．重みは近傍の正例集合密度から計算し，アンサンブルは，各正例集合ごとに，正例集合の事例と全ての負例集合の事例から作られる弱学習器によって行う．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特徴を示す．

Multiple Instance Learning by distribution density of neighbor sets of instances.

Toshiki Kawamura† and Kuniaki Uehara†

In traditional supervised learning, a learning algorithm receives a training set which consists of individually labeled examples. But in real application, the teacher cannot always label an individual instance. In Multiple-Instance learning, the teacher labels examples that are sets of instances. Although this learning problem is harder than traditional supervised learning, Multiple-Instance problems is lager than supervised ones. In this paper, we propose an ensemble of weighted instance learners. Our method calculates these weights by the density of neighbor positive sets and constructs an ensemble by applying weak learners trained from each positive set. We present experimental results on artificial data and benchmark datasets.

1. はじめに

通常の教師あり学習問題では，学習データの各事例にラベルが与えられる．しかし，現実世界の問題では個々の事例にラベルが付けられない場合がある．このような問題では，個々の事例にはラベルを付けず，事例の集合にラベルを付ける Multiple-Instance 学習（MIL）が利用される．

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．特に Diverse Density[1]（DD）は，直観的で単純な仮説にも関わらず，高い分類精度を実現しており，多くの応用例において適用されている．しかし，DD のアプローチでは，事例が正となる領域（正領域）の形状が単純であることを前提としているため，正領域が複雑な場合に対応できないという問題がある．

また，Citation-kNN[2]（C-kNN）は事例集合間の距離関数を新たに定義することにより，k-NN を MIL に拡張した手法である．この手法では，正領域が複雑でも対応可能である．しかし，常に集合単位で分類をするため，個々の事例の情報を十分には利用していないと考えられる．たとえば，個々の事例を分類に利用する手法に比べて精度が劣る可能性があると考えられる．

そこで，本研究では DD の性能と C-kNN の柔軟性に注目し，それらを組み合わせた手法を提案する．本手法は，正例集合（正例バッグ）ごとに学習器を用意してアンサンブルすることで推論を行うものである．同時に弱学習器の精度向上のために，正例バッグの各事例に重みを付与する．重みは，近傍にある正例バッグの密度によって計算される．これによって，正例バッグの各事例が正事例である確からしさを得ることができる．

本論文では，まず 2 章で Multiple-Instance 学習の概要を与えたあと，既存手法である DD と C-kNN について説明する．3 章では，提案手法について具体的に示し，4 章で提案手法の評価と実験結果を考察する．最後に 5 章で，結論と今後の課題について述べる．

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

2. Multiple-Instance 学習 (MIL) とは

MIL とは，通常の教師あり学習よりもあいまいなデータを扱える手法である．具体的には，個々の事例にラベルが与えられないデータでも，バッグ（任意の数の事例集合）にラベルを与えることで MIL として扱える．

バッグのラベルは，バッグに含まれる事例によって決まる．バッグ内に正の事例を 1 つでも含むとき，バッグのラベルは正となり，逆にバッグ内が負の事例のみであるとき，バッグのラベルは負となる．

以下では MIL アルゴリズムのうち，特に DD と C-kNN の手法について説明を行う．まず以降で使用するバッグの記号表現を定義する．正例バッグ全体を $\mathbf{B}^+$，負例バッグ全体を $\mathbf{B}^-$ とする．i 番目のバッグを $\mathbf{B_i}$，ラベルが分かっている場合は $\mathbf{B_i^+}, \mathbf{B_i^-}$ と書くことにする．また，バッグの j 番目の事例を $\mathbf{B_{ij}}$，その事例の k 番目の属性を $\mathbf{B_{ijk}}$ とする．

2.1 Diverse Density

DD は，属性空間上で正例バッグが多く重なりあい，負例バッグが含まれていないところが正領域であるという考え方の手法である．このため，正例バッグと負例バッグの集合が与えられたとき，全ての負例バッグ内の全ての事例から遠く，正例バッグ内の少なくとも 1 つの事例に近い部分領域を獲得する．また，獲得すべき部分領域は，正例バッグに含まれる事例の密度が高いだけでなく，正例バッグの密度が高い部分領域である．このような部分領域を獲得する際の指標を多様性密度という．

形式的には，属性空間の座標 x における多様性密度 DD(x) は，与えられた i 番目の正例バッグ B_i^+ と，負例バッグ B_i^- からの寄与 $\Pr(t|B_i)$ の積として定義される．x への各バッグ i の寄与は，noisy-or model の関数によって評価する．

$$\mathrm{DD(x)} \propto \left(\prod_j \Pr(\mathrm{x}|B_j^+) \prod_k \Pr(\mathrm{x}|B_k^-)\right) \quad (1)$$

$$\Pr(t|B_i^+) = (1 - \textstyle\prod_j (1 - \Pr(x|B_{ij}^+)))$$

$$\Pr(t|B_i^-) = \textstyle\prod_j (1 - \Pr(x|B_{ij}^-))$$

なお，各事例の寄与はガウス分布 $\Pr(x|B_{ij}) = \exp(-\sum_k s_k^2 (B_{ijk} - x_k)^2)$ 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ーである．

バッグの多様性密度は，そのバッグに含まれる事例毎の多様性密度の最大値とする．

$$\mathrm{DD}(B_\mathrm{x}) = \max_m \mathrm{DD}(\mathrm{B_{xm}}) \quad (2)$$

この値が大きければバッグが正例であり，小さければ負例であると判定する．

DD では，noisy-or model を用いて，正事例も負事例も含む正例バッグから正となる領域（正領域）を取り出しているが，正領域は 1 つであると仮定している．このため，正領域が複数の場合は性能が低くなると考えられる．

2.2 Citation-kNN

C-kNN は，バッグ単位での k-NN を用いた lazy-learning 手法である．ただし，通常の k-NN 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 c 近傍バッグであり，citer とは，未知バッグが c 近傍になる訓練バッグである．

各近傍バッグは，ハウスドルフ距離 [3] を拡張した距離関数によって計算する．拡張ハウスドルフ距離は，集合 $A = \{a_1, \ldots, a_m\}$，$B = \{b_1, \ldots, b_n\}$ 間において以下のように定義され，k 番目に短い距離を採用する．

$$H_k(A, B) = \max\{h_k(A, B), h_k(B, A)\} \quad (3)$$

$$\text{where } h_k(A, B) = \operatorname*{kth}_{a \in A} \min_{b \in B} \| a - b \| \quad (4)$$

reference の関係になる正例バッグ数を 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，負例バッグであると分類する．

C-kNN の問題として，すべてのデータをバッグ単位で扱っているため，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，負例バッグの事例は各事例のラベルが明らかに負であるにも関わらず，その情報を利用していない．そのため，負例バッグの情報を利用している手法に比べて性能が劣る可能性がある．

3. 近傍事例集合の密度分布による分類手法

本章では DD の多様性密度の考え方と C-kNN の lazy-learning アプローチを組み合わせた手法を提案する．本手法のアプローチは，アンサンブルによって多様性密度を利用し，弱学習器として k-NN を用いた lazy-learning の考え方を利用する．この特徴によって DD と C-kNN の弱点を補い合うことが出来る．

本研究では，各正例バッグを学習器で表現し，アンサンブルによって，属性空間の座標 x のラベルを推定する．正例バッグを学習器で表現するとは，1 つの正例バッグと全ての負例バッグを用いて学習させて，属

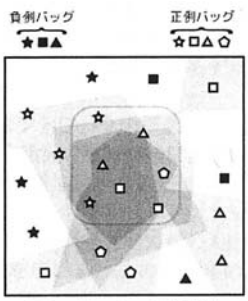

図 1 アンサンブルの例.
Fig. 1 Example of Ensemble.

性空間上での正例バッグの領域を表現することである.

アンサンブルした例を**図 1** に示す. 図 1 は, 弱学習器に Nearest Neighbor を使った結果を示している. 異なる記号は異なるバッグを示しており, 色の濃い部分は, 正例バッグの領域が多く重なり合っている部分であり, 正領域である可能性が高いことを示している.

属性空間の座標 x におけるラベルは, 正バッグごとに全ての負バッグとの寄与の和から計算される.

$$f(\mathrm{x}) = \sum_i \Pr(\mathrm{x}|\mathrm{B}_i^+, \mathrm{B}^-) \tag{5}$$

この値が大きければ, 正例であると判定する. バッグの評価には, 事例ごとの評価の最大値 $f(\mathrm{B_x}) = \max_j f(\mathrm{B_{x,j}})$ を用いている.

弱学習器の精度向上のために, 正例バッグの各事例に正である確からしさを重みとして与えている. なお, 事例の重みは近傍の正例バッグ密度から計算している. 具体的には, 近傍バッグを調べるために事例とバッグの距離を以下のように定義している.

$$d(\mathrm{x}, B_i) = \min_l \| x - B_{il} \| \tag{6}$$

定義した距離を用いて, 事例 x に近い l 個のバッグを $C_1, \cdots, C_l$ とすると, 重みは次の式によって計算している.

$$\mathrm{Weight}(\mathrm{x}) = \frac{1}{l} \sum_l \sigma(C_l) \tag{7}$$

$$\sigma(C_i) = \begin{cases} 0 & \text{if } C_i \in B^- \\ 1 & \text{otherwise} \end{cases}$$

正例バッグごとに作られる弱学習器 $\Pr(\mathrm{x}|\mathrm{B}_i^+, \mathrm{B}^-)$ は, 重みを使った kNN によって計算している. 事例 $c = B_{ij}^+, B_{pq}^-$ のうち, ユークリッド距離から事例 x に近い k 個の事例を $c_1, \cdots, c_k$ とすると, $\Pr(\mathrm{x}|\mathrm{B}_i^+, \mathrm{B}^-)$ は次の式となる.

$$\Pr(\mathrm{x}|\mathrm{B}_i^+, \mathrm{B}^-) = \frac{1}{k} \sum_k \mathrm{Weight}(c_k) \tag{8}$$

バッグの評価は, 含まれる事例ごとの評価の最大値としている. したがって, $f(\mathrm{B_x}) = \max_m f(\mathrm{B_{x}}_m)$ の値が大きければ, バッグ $\mathrm{B_x}$ は正例であり, 小さければ負例となる.

4. 性能評価および検討

人工データおよびベンチマークデータセットによる性能評価を行う.

4.1 人工データ

人工データの条件は, バッグ数が 50, バッグ内の事例数が平均 10, 属性数が 2 次元であり, 変化させる条件は, 事例の分布, 正例バッグの比率, バッグラベルのノイズ, 正領域の数である. それぞれ, 事例全てが混合正規分布に従う場合とバッグごとの正規分布に従う場合, 正例バッグと負例バッグの比率は 1 : 5, 1 : 1, 5 : 1 の 3 通り, ノイズがある場合 (10%) とない場合, 正領域が 1 つの場合と 2 つの場合について実験する.

テストデータは属性空間上を等間隔に取り出した 2401 点を用いている. 分類器の性能は, AUC[4] を用いて計算した. 本実験におけるパラメータは, 提案手法では $k = 2, l = 2$, DD では $s = 1$, C-kNN では $k = 2, c = 4$ としている.

実験結果を **表 1** に示す. 表 1 に, distribution は事例の分布が全体の混合分布に従う場合 (Instance) とバッグごとの分布に従う場合 (Bag), rate は正例バッグと負例バッグの比率, noise はバッグラベルの有無, one area と two area はそれぞれ正領域が 1 つの場合と 2 つの場合である. また, 数値はすべて AUC (%) を示し, 特に最も高かった利得を太字で強調している.

DD は正領域が 2 つ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. C-kNN は事例の分布がバッグごとの正規分布の場合に高い利得を得ているが, 事例の分布が全体の混合分布の場合には利得が低下している. 一方, 提案手法はどの場合においても安定した結果を示している.

提案手法は, 正例バッグと負例バッグの比率による利得の変化は見られない. また, 事例の分布の違いによる利得の変化も見られない. よって, 安定した手法だと考えられる. さらに, 正領域が 2 つの場合はどの条件においても既存手法を上回っている. これらから, 本手法は既存手法に比べ, 特に正領域が分割しているときに高い精度が得られると考えられる.

表 1　人工データセットによる結果.
Table 1　The experimental result: artifical dataset.

distribution	rate	noise	one area			two area		
			DD	C-kNN	提案手法	DD	C-kNN	提案手法
Instance	1:5	なし	57.01	81.78	**91.31**	52.34	66.77	**77.02**
		あり	45.58	57.38	**67.85**	48.24	53.23	**61.20**
	1:1	なし	87.53	70.26	**95.98**	70.04	67.81	**86.69**
		あり	**62.37**	51.72	61.68	64.64	60.10	**73.92**
	5:1	なし	92.77	58.59	**96.71**	69.11	57.64	**83.07**
		あり	81.06	54.03	**84.24**	62.97	53.32	**66.30**
Bag	1:5	なし	48.64	91.87	**96.85**	30.31	84.55	**86.81**
		あり	35.84	82.65	**90.57**	31.77	75.22	**81.95**
	1:1	なし	98.11	89.61	**98.79**	74.45	81.96	**90.48**
		あり	95.57	82.83	**96.02**	69.79	79.65	**87.47**
	5:1	なし	**97.73**	76.87	96.11	60.70	71.98	**83.55**
		あり	94.20	69.04	**94.31**	63.72	68.63	**78.05**

表 2　ベンチマークデータセットによる結果.
Table 2　The experimental result: benchmark dataset.

	Musk1	Musk2	Elephant	Fox	Tiger
提案手法	**92.4**	85.3	**88.4**	**67.8**	**80.4**
C-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

4.2　ベンチマークデータセット

既存手法との性能比較のために，MIL 用データセットの麝香芳香予測データと画像分類データによる実験を行う．麝香芳香予測データ（Musk1，Musk2）は，UCI Machine Learning Repository[5] から入手した．また，画像分類データ（Elephant，Fox，Tiger）は，Andrew ら [6] が生成したデータを使用した．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 $k=3, l=1$，それ以外は $k=4, l=4$ 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り引用したが，EM-DD の分類精度は Andrews らによる結果 [6] を使用しており，C-kNN の画像分類データセットの結果は，Jun Yang の行った結果 [7] を参照している．また，APRs[8]，DD，EM-DD[9]，MI-SVM，mi-SVM[6] は ten-fold corss validation，C-kNN は leave-one-out により検定が行われている．そのため，提案手法は両方の検定を利用している．

実験結果を**表 2** に示す．表 2 にて，上が leave-one-out であり，下が ten-fold corss validation である．また，数値はすべて分類精度 (%) を示し，特にもっとも高かった分類精度を太字で強調している．

C-kNN との比較は，Musk1，Musk2，Tiger ではほぼ同等，Elephant，Fox では高い精度を示している．

他の既存手法との比較では，Musk1，Musk2 で APRs には劣るものの，DD とほぼ同等の精度を示している．画像分類のデータセットにおいては，Tiger では，MI-SVM に劣るものの，ほぼ最良である．

これらの結果から，DD と C-kNN を組み合わせた本手法は，データセットによらず高い性能を示す手法であると分かった．

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 学習の分類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，推論時にはバッグのラベルと重みから分類する手法を提案した．提案手法では，正バッグごとに kNN 学習器を生成し，属性空間上でのバッグの領域を表現して，正バッグの多様性密度を調べている．この結果，DD の仮説の優秀さと C-kNN の柔軟性の両方をもつ手法を実現できた．

今後は，各種の条件を伴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 kNN を弱学習器として用いたが，これから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考察する予定である．

参考文献

1) Maron, O. and Lozano-Pérez, T.: A framework for multiple-instance learning, *Proc. of NIPS*, MIT Press, pp. 570 576 (1998).
2) Wang, J. and Zucker, J.-D.: Solving the Multiple-Instance Problem: A Lazy Learning Approach, *Proc. of 17th ICML*, pp. 1119 1125 (2000).
3) Edgar, G. A.: *Measure, topology, and fractal geometry*, Springer (1995).
4) 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).
5) A.Asuncion, D.N.: UCI Machine Learning Repository (2007).
6) Andrews, S., Tsochantaridis, I. and Hofmann, T.: Support Vector Machines for Multiple-Instance Learning., *Proc. of NIPS*, pp. 561 568 (2002).
7) Yang, J.: A Toolkit for Multiple-Instance Learning and its Experiments with Information Retrieval.
8) 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).
9) Zhang, Q. and Goldman, S. A.: EM-DD: An Improved Multiple-Instance Learning Technique., *Proc. of NIPS*, pp. 1073 1080 (2001).