Vol.2015-CG-161 No.4
Vol.2015-CVIM-199 No.4
2015/11/6

不完全ラベル付きデータからのマルチラベル分類問題

兼平 篤志[1,a)]　原田 達也[1,b)]

概要：本研究は付与されたラベルが不完全なデータからのマルチラベル分類学習に焦点を当てる．画像アノテーションをはじめとして，マルチラベル問題においてはサンプルにラベルが完全に付与されていない状況は頻繁に起こり，このようなデータからの識別器学習は重要な問題である．不完全なラベル問題は (1) サンプルに付与されたラベルは信頼できる，(2) サンプルに本来付与されるべきだが，実際には付与されていないラベルが存在する，という性質を持つ．このような設定の問題は 2 値分類問題においては PU (Positive and unlabeled) 分類問題として研究されてきた．本研究ではマルチラベル分類問題における不完全なラベル付きのサンプルからの学習を，2 値 PU 分類問題を拡張したマルチラベル PU 分類問題として捉え，解析を行うことでマルチラベル PU 分類問題においてラベル欠損の影響をなくすために損失関数が満たすべき条件を示した．また複数のデータセットを用いた実験によってそれらの有効性を示した．

キーワード：PU 分類，マルチラベル分類，半教師付き学習

1. 序論

マルチラベル分類問題は 1 つのサンプルが複数のラベルを取り得るという設定を扱う問題である．最も単純なマルチラベルの学習手法として，クラスごとに独立に識別器を学習する方法が挙げられるが，これはラベル間の依存関係を無視しており，ラベル間に相関がある場合に精度が悪くなることが知られている [1]．そのためラベル間の依存関係を考慮したマルチラベル学習手法が必要となる．マルチラベル分類問題は機械学習において重要な問題として，従来多くの研究が成され [1], [2], [3]，また画像認識分野においても様々な応用先が存在する [4], [5], [6].

マルチラベル分類学習のためのデータセットの収集にはクラウドソーシングを用いる方法や半自動で行う方法 [7] などがあるが，多くの場合全てのサンプルに完全なラベルを付与することは困難であり得られるラベルは不完全なものとなる．

例えば，一般的な画像認識のデータセット作成では，**図 1** のように複数のクラウドワーカに同一の画像へのラベル付けを依頼し，信頼性確保の為に複数，あるいは全てに共通する回答をそのサンプルのラベルとみなす．この処理により，図 1 の”テーブル”のような誤ったラベルが付与されることを避けることができる一方，”ソファ”のような正解のラベルまでも排除してしまう可能性がある．その為，得られるデータは，(1) 付与されたラベルは信頼できる，(2) サンプルに本来付与されるべきだが実際には付与されていないラベルが存在する，という性質を持つ．このようなデータを用いて学習された識別器の精度は低下すると考えられ，不完全なラベル付けの問題はマルチラベル分類問題において本質的な問題であると言える．

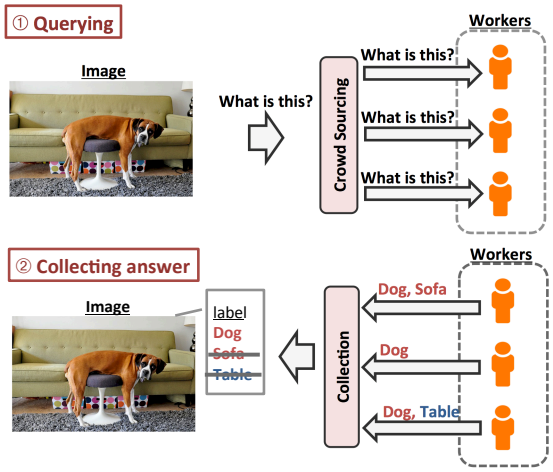

図 1 クラウドソーシングによるデータセット作成．複数の回答に共通する回答をラベルとみなす処理により得られるラベルは信頼のおけるものとなる一方で，不完全なものとなる．

Fig. 1 Construction of dataset using crowd sourcing. Obtained labels will be reliable while incomplete, due to the process assigning only the common labels from different answers.

[1] 東京大学
The Univ. of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113–8656, Japan
[a)] kanehira@mi.t.u-tokyo.ac.jp
[b)] harada@mi.t.u-tokyo.ac.jp

そこで，本研究の目的は，サンプルに付与されたラベルが不完全であっても頑強に学習が可能となる方法を提案することである．

本研究では以下の問題を扱う．

(1) サンプルに付与されたラベルは必ず正例ラベルである．
(2) サンプルに本来付与されるべきだが，実際には付与されていないラベルが存在する．
(3) サンプルが複数ラベルを取り得る．

(1)，(2) の問題設定は機械学習の分野では PU（Positive and Unlabeled）分類問題として研究されてきた [8], [9]．本稿では (1)，(2)，(3) を含めマルチラベル PU 分類問題と定義し，2 値 PU 分類問題の解析 [10] をマルチラベル分類問題に拡張することで不完全なラベル付けデータを用いた学習を頑強にするための損失関数の条件を示す．本研究の貢献は以下のようにまとめられる．

(1) 不完全なラベル付きサンプルからのマルチラベル分類の学習ために，マルチラベル PU 分類問題を定義した．
(2) マルチラベル PU 分類問題においてラベル欠損の影響をなくすために損失関数が満たすべき条件を示した．
(3) 複数のデータセットを用いた実験によりそれらの条件の有効性を示した．

本稿の構成は以下の通りとなる．1 章では，本研究の目的と貢献について述べた．2 章では，関連研究について述べる．3 章では，マルチラベル分類問題の設定を説明する．次に 4 章でマルチラベル問題の PU 分類問題への適用を考え，不完全ラベルデータを用いてロバストに学習を行うための損失関数の条件を述べる．5 章では人工データと画像アノテーションデータにおいて実験を行い，それらの条件の有効性を示した．6 章では考察を行い，7 章で本稿の結論を述べる．

2. 関連研究

2.1 マルチラベル分類問題

従来マルチラベル分類問題について多くの研究が為されており，その中でも最も一般的なアプローチの 1 つとしてランキングベースの方法がある．ランキングベースの手法はランク損失 [11] を最小化することで全ての正例のクラスが負例のクラスよりも高くランク付けされることを目指す．[12] はこれを大規模データに拡張可能なように拘束条件を緩和した．コンピュータビジョンの分野においてもランク損失を用いた多くのモデルが提案されており，[5] はランク損失を拡張し，さらに特徴空間とラベル空間の共通空間を求める手法と統合したモデルである WSABIE を提案した．[13], [14] は物体認識への応用に用いた．[15] は画像アノテーションの為に，[5] が提案した損失関数を畳み込みニューラルネットの損失として用いた．しかし，これら全ての研究は完全にラベルが付与されていることを前提としており，ラベルが不完全であることを想定していない．

2.2 PU 分類問題

PU 分類問題は正例サンプルとラベルなしサンプルのみから識別器を学習する問題であり，従来いくつかの研究が行われてきた [9]．[8] は確率的な方法で，観測可能なサンプルを用いて学習した識別器から真の識別器を推定する方法を提案した．[10] は PU 分類問題の解析を行い，2 値の PU 分類問題は正例クラスとラベルなしクラスで損失関数の重みを変えるコスト考慮型学習 [16] として捉えることができ，対称な非凸の代理損失関数を用いることによってロバストな学習が可能となることを示した．しかし，これらの研究はすべて 2 値分類問題に焦点を当てたものであり，サンプルが複数ラベルを取り得ることを想定していない．

2.3 不完全ラベル付きデータからの学習

マルチラベル問題におけるラベルの欠損を取り扱った研究として，[17]，[18]，[19] がある．[17] は正例ラベルと負例ラベルの識別スコアの差分を正則化項としてランク損失に加え，正則化に group lasso を用いることで最適化の中で欠損ラベルの影響を除外することを試みた．また [18] は [8] の方法にラベル間の関係を取り入れることでマルチラベル問題へ拡張した．[19] は条件制約付きボルツマンマシンを用いた．しかしこれらの研究では不完全ラベル付きのデータからのマルチラベル PU 分類問題において，損失関数が満たすべき条件は述べられていない．

3. マルチラベル分類

本章では，マルチラベル分類問題の設定について説明する．$\mathcal{X}$ をサンプル空間，$\mathcal{Y} = \{0, 1\}^m$ をサンプルが取り得るラベルの集合とする．ここで m はクラス数である．$y_i = 1$ はクラス i がサンプルに対して正例ラベルであることを示し，$y_i = 0$ は負例であることを表す．サンプル数 N のデータセット $S = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_N, \mathbf{y}_N)\}$ は $\mathcal{X} \times \mathcal{Y}$ 上の未知の確率分布からサンプルされるとする．また各クラスらしさを表すスコア関数を $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_m(\mathbf{x})) : \mathcal{X} \to \mathbb{R}^m$ と定義する．マルチラベル分類問題では，損失関数の（サンプル，ラベル）空間上における期待値を最小化することを試みる．つまり，以下の値を最小化する $\mathbf{f}^* = \mathrm{argmin}\, \mathcal{L}(\mathbf{f})$ を求める．

$$\mathcal{L}(\mathbf{f}) = \mathbb{E}_{\mathbf{xy}}[L(\mathbf{f}(\mathbf{x}), \mathbf{y})]$$

$L(\mathbf{f}(\mathbf{x}), \mathbf{y})$ はサンプルに対するスコアとラベルによって定義される損失関数であり，損失関数の種類としてハミング損失やサブセット 0-1 損失などいくつか提案されているが，ここでは一般的に使われるランク損失について考える．

3.1 ランク損失

マルチラベル分類で一般的に使われるランク損失について説明する．マルチラベル問題で用いられるランク損失

は，2 つのラベルのペアに対してそれらが誤ってランク付けされる場合に損失を与えるものであり，i, j をクラスのインデックスとした時，以下の式で定義される．

$$L_{\mathrm{rank}}(\mathbf{f}(\mathbf{x}), \mathbf{y}) = \sum_{y_i=1, y_j=0} [\![f_i < f_j]\!] + \frac{1}{2}[\![f_i = f_j]\!]$$

サンプルの i クラスの識別器に対するスコアを表す $\mathbf{f}(\mathbf{x})$ の i 番目の要素を $\mathbf{x}$ を省略して f_i と書く．また $[\![\cdot]\!]$ は括弧の中の条件を満たすなら 1，そうでないなら 0 をとる指示関数である．最小化すべき期待損失は，

$$\begin{aligned}\mathcal{L}_{\mathrm{rank}} &= \mathbb{E}_{\mathbf{xy}}[L_{\mathrm{rank}}(\mathbf{f}(\mathbf{x}), \mathbf{y})] \\ &= \sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}) \mathbb{E}_{\mathbf{x}|\mathbf{y}}[L_{\mathrm{rank}}(\mathbf{f}(\mathbf{x}), \mathbf{y})] \\ &= \sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}) \sum_{y_i=1, y_j=0} \mathbb{E}_{\mathbf{x}|\mathbf{y}} \left[[\![f_i < f_j]\!] + \frac{1}{2}[\![f_i = f_j]\!] \right] \quad (1)\end{aligned}$$

となる．2 つの総和を入れ替えることにより，式 (1) は以下で書き直すことができる．

$$\begin{aligned}&\mathcal{L}_{\mathrm{rank}} \\ &= \sum_{y_i=1, y_j=0} \sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}) \mathbb{E}_{\mathbf{x}|\mathbf{y}} \left[[\![f_i < f_j]\!] + \frac{1}{2}[\![f_i = f_j]\!] \right] \\ &= \sum_{y_i=1, y_j=0} P(y_i = 1, y_j = 0) \mathbb{E}_{\mathbf{x}|y_i=1, y_j=0} \left[[\![f_i < f_j]\!] + \frac{1}{2}[\![f_i = f_j]\!] \right]\end{aligned}$$

ここで，誤ランク率を

$$\begin{aligned}&R(i, j) \\ &= \mathbb{E}_{\mathbf{x}|y_i=1, y_j=0} \left[[\![f_i < f_j]\!] + \frac{1}{2}[\![f_i = f_j]\!] \right] \\ &= P(f_i < f_j \mid y_i = 1,\ y_j = 0) + \frac{1}{2} P(f_i = f_j \mid y_i = 1, y_j = 0)\end{aligned}$$

と定義する．誤ランク率 $R(i, j)$ はサンプルの正例ラベルの識別器のスコアが負例ラベルの識別器のスコアよりも小さくなるような確率である．この誤ランク率を用いて，ランク損失 $\mathcal{L}_{\mathrm{rank}}$ は，

$$\begin{aligned}&\mathcal{L}_{\mathrm{rank}} \\ &= \sum_{y_i=1, y_j=0} P(y_i = 1, y_j = 0) R(i, j) \\ &= \sum_{1\le i<j\le m} P(y_i = 1, y_j = 0) R(i, j) + P(y_i = 0, y_j = 1) R(j, i)\end{aligned} \quad (2)$$

と書ける．ランク損失 $\mathcal{L}_{\mathrm{rank}}$ は誤ランク率 $R(i, j)$ を全てのラベルの組に対して期待値を取ったものと考えることができる．本研究では，PU 問題においてこの損失関数を最小化することを考える．

4. マルチラベル PU 分類

本研究ではマルチラベル問題において，サンプルに付与されたラベルは必ず正例ラベルであり，かつサンプルに付与されていないラベルが必ずしも負例ラベルであるとは限らないという 2 つの性質を持つデータから識別器を学習する問題をマルチラベル PU 分類問題と定義した．本章では PU 分類問題の解析 [10] をマルチラベル分類問題に拡張することで，マルチラベル PU 分類問題においても 2 値 PU 分類問題と同様に，

(1) マルチラベル PU 分類問題は，正例とラベルなしサンプルを用いたコスト考慮型学習として書き表すことができる．
(2) ラベル欠損が存在するデータに対して代理損失関数を使って学習を行った際にラベル欠損が存在しない場合の値から誤差が生じてしまうが，対称な代理損失関数を選択することでこの誤差がキャンセルされる．

ことを説明する．

4.1 損失関数の適切な重み付け

本節ではマルチラベル PU 分類問題はコスト考慮型学習で書き表すことができる，つまり損失関数に対して適切な重み付けが必要であることを説明する．

コスト考慮型学習 [16] とは誤分類に対する損失の重みをクラスごとに変化させる学習方法である．これをランク損失に適用した際，以下の式で表せる．

$$\begin{aligned}&\mathcal{L}_{\mathrm{rank}} \\ &= \sum_{1\le i<j\le m} c_{ij} P(y_i = 1, y_j = 0) R(i, j) + c_{ji} P(y_i = 0, y_j = 1) R(j, i)\end{aligned}$$

c_{ij} は $y_i = 1, y_j = 0$ であるラベル組への誤りに対する損失の重みを表す．

不完全にラベル付けされたサンプルを用いて，式 (2) を最小化することを考える．本研究の設定では学習データに負例ラベルが存在しないため，式 (2) における誤ランク率 $R(i, j)$ を直接学習データから推定することはできない．そこで，学習データから推定可能な擬似誤ランク率 $R_X(i, j)$ を以下のように定義する．

$$\begin{aligned}&R_X(i, j) \\ &= P(f_i < f_j \mid s_i = 1, s_j = 0) + \frac{1}{2} P(f_i = f_j \mid s_i = 1, s_j = 0)\end{aligned}$$

$s_i \in \{0, 1\}$ はラベルが付いているかどうかの状態を表し，$s_i = 0$ はラベル無し，$s_i = 1$ はラベル有りの状態を表す．擬似誤ランク率 $R_X(i, j)$ は付与されたラベルのクラスに対する識別器のスコアがラベルが付与されていないクラスのスコアよりも小さくなる確率であり，不完全なラベル付きのデータから推定可能である．ラベル付けが不完全であることを考慮せずに，完全にラベルが付いているものとみなした場合，最小化しているのは式（2）ではなく，

$$\begin{aligned}&\hat{\mathcal{L}}_{\text{rank}}\\ &= \sum_{s_i=1,s_j=0} P(s_i=1,s_j=0)R_X(i,j)\\ &= \sum_{1\le i<j\le m} P(s_i=1,s_j=0)R_X(i,j)+P(s_i=0,s_j=1)R_X(j,i)\end{aligned} \tag{3}$$

となる．

ここで，データの生成される分布に対して [10] と同様に，ラベル無しサンプルは周辺分布から生成されると想定する．この設定は PU 問題では”case-controlled” [20] と呼ばれる．この条件は $P(*|s_i=0)=P(*)$ と書くことができる．更に，本来正例であるラベルの欠損はサンプルに偏りなく起こるという仮定をとる．この条件は，$P(*|s_i=1)=P(*|y_i=1)$ と書ける．これらの仮定を用いることで，擬似誤ランク率 $R_X(i,j)$ は誤ランク率 $R(i,j)$ を用いて次式の様に書き表すことができる (付録 A.1)．

$$R_X(i,j)=(1-\pi_{ij})R(i,j)+\pi_{ij}R_{-X}(i,j) \tag{4}$$

ここで，

$$\begin{aligned}&R_{-X}(i,j)\\ &= P(f_i<f_j\,|\,y_i=1,y_j=1)+\frac{1}{2}P(f_i=f_j\,|\,y_i=1,y_j=1)\end{aligned}$$

であり，また

$$\pi_{ij}=P(y_j=1\,|\,y_i=1)$$

とおいた．$R_{-X}(i,j)$ は付与されていないラベルの中に正例ラベルが含まれることにより誤って与えられる損失を表す．つまり，不完全なラベル付きデータを用いた学習においては，ランク損失は（正例ラベル，負例ラベル）のペアではなく，（ラベル有り，ラベル無し）のペアに対して損失を与える．しかし，このペアの中には本来は損失に含まれるべきではない（正例ラベル，正例ラベル）のペアが含まれている．ここでは，この誤って過剰に与えれた損失を $R_{-X}(i,j)$ と表し，またその割合を π_{ij} と表す．従って（4）式は本来与えられるべきであった損失と与えられるべきではなかった損失に分解していると解釈することができる．(4) 式を変形することにより，

$$R(i,j)=\frac{1}{1-\pi_{ij}}\left(R_X(i,j)-\pi_{ij}R_{-X}(i,j)\right)$$

が得られる．これを式 (2) に代入し，

$$\begin{aligned}&R_{-X}(i,j)+R_{-X}(j,i)\\ &= P(f_i>f_j \text{ or } f_i=f_j \text{ or } f_i<f_j\,|\,y_i=1,y_j=1)\\ &= 1\end{aligned}$$

の関係を使うことで，以下の式が得られる (付録 A.2)．

$$\begin{aligned}&\mathcal{L}_{\text{rank}}\\ &= \sum P(y_i=1)R_X(i,j)+P(y_j=1)R_X(j,i)-P(y_i=1,y_j=1)\\ &= \sum c_{ij}P(s_i=1,s_j=0)R_X(i,j)+c_{ji}P(s_i=0,s_j=1)R_X(j,i)\\ &\quad -P(y_i=1,y_j=1)\end{aligned} \tag{5}$$

ここで

$$c_{ij}=\frac{P(y_i=1)}{P(s_i=1,s_j=0)}$$

である．式 (3) と比較することで，マルチラベル PU 分類問題はコスト考慮型学習として捉えることができ，ラベル欠損を含むデータを用いて学習を行う際，各項を c_{ij} によって重み付けを行った損失関数を最小化することで，本来最小化すべき損失関数を最小化できることが分かる．$P(s_i=1,s_j=0)$ は学習データセットに存在する $s_i=1,s_j=0$ を満たすサンプルの割合で学習データセットから推定することができる．また，$P(y_i=1)$ は [21], [22] 等を用いて推定することができる．

4.2 対称な代理損失関数の使用

本節では最適化の際に使われるべき代理損失関数の条件を導く．誤ランク率 R は以下の 0-1 関数の期待値として書くことができる．

$$R(i,j)=\mathbb{E}_{\mathbf{x}|y_i=1,y_j=0}[l_{0-1}(f_i-f_j)],\ \ l_{0-1}(x)=\begin{cases}1\ (\text{if}\,x<0)\\ \frac{1}{2}\ (\text{if}\,x=0)\\ 0\ (\text{otherwise})\end{cases}$$

同様に

$$R_X(i,j)=\mathbb{E}_{\mathbf{x}|s_i=1,s_j=0}[l_{0-1}(f_i-f_j)]$$

と書ける．これらの $l_{0-1}(x)$ を用いた損失関数の直接的な最適化は組み合わせ最適化問題となり困難である．そこで最適化の際に一般的に代理損失 $l'(x)$ によって 0-1 損失の近似を行う．例えば，SVM 等で使われるヒンジ損失関数 l'_{hin} を用いると擬似誤ランク率は以下のように近似することができる．

$$R_X(i,j)\approx\mathbb{E}_{\mathbf{x}|s_i=1,s_j=0}[l'_{\text{hin}}(f_i-f_j)],\ \ l'_{\text{hin}}(x)=\begin{cases}1-x\ (\text{if}\,x<1)\\ 0\ (\text{otherwise})\end{cases}$$

ラベルが完全に付与されたデータを用いた場合，最小化すべきランク損失は式（2）に代理損失を適用すると以下で表せる．

$$\begin{aligned}&\mathcal{L}'_{\text{rank}}\\ &= \sum P(y_i=1,y_j=0)E_{\mathbf{x}\,|\,y_i=1,y_j=0}[l'(f_i-f_j)]\\ &\quad +P(y_i=0,y_j=1)E_{\mathbf{x}\,|\,y_i=0,y_j=1}[l'(f_j-f_i)]\end{aligned}$$

一方で，ラベルの欠損があるデータに対して代理損失を適

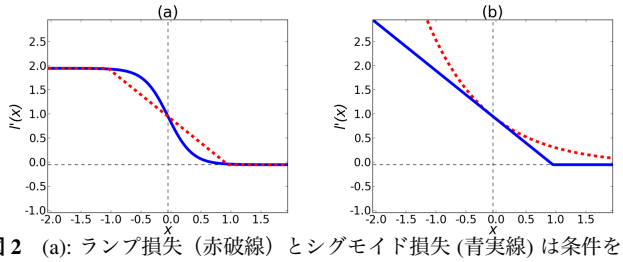

図 2 (a): ランプ損失（赤破線）とシグモイド損失 (青実線) は条件を満たす．(b):指数損失（赤破線）とヒンジ損失（青実線）は条件を満たさない．

Fig. 2 (a): Ramp loss (red dashed line) and sigmoid loss (blue solid line) meet the described condition. (b): Exponential loss (red dashed line) and hinge loss (blue solid line) don't meet the described condition.

用した場合，式（5）の R_X を代理損失の期待値によって置き換えることにより，

$$\begin{aligned}
&\mathcal{L}''_{\text{rank}} \\
&= \sum P(y_i=1)E_{\mathbf{x}\mid s_i=1,s_j=0}[l'(f_i-f_j)] \\
&\quad +P(y_j=1)E_{\mathbf{x}\mid s_i=0,s_j=1}[l'(f_j-f_i)] - P(y_i=1,y_j=1) \\
&= \sum P(y_i=1,y_j=0)E_{\mathbf{x}\mid y_i=1,y_j=0}[l'(f_i-f_j)] \\
&\quad +P(y_i=0,y_j=1)E_{\mathbf{x}\mid y_i=0,y_j=1}[l'(f_j-f_i)] \\
&\quad -P(y_i=1,y_j=1)\Big(1-E_{\mathbf{x}\mid y_i=1,y_j=1}[l'(f_i-f_j)+l'(f_j-f_i)]\Big)
\end{aligned}$$

が得られる．上の 2 式から $\mathcal{L}''_{\text{rank}} = \mathcal{L}'_{\text{rank}} + (\text{Error})$ の関係にあり，代理損失を用いた場合，不完全なラベル付けがなされたデータに対する損失関数は完全にラベルが付いているデータに対する損失関数から誤差が生じることが分かる．ここで，代理損失関数 $l'(\cdot)$ を $l'(f_i-f_j)+l'(f_j-f_i)=1$ を満たすように選択すると誤差項はキャンセルされ，ラベル欠損がない場合の値と一致する．ヒンジ損失や指数損失のような代理損失関数 (**図 2**(b)) はこの条件を満たさず，ランプ損失やシグモイド損失のような対称な非凸の代理損失関数 (図 2(a)) はこの条件を満たす．

これは例えば，データが完全にラベル付けされており，かつ完全に分離可能な場合，つまり $\min \mathcal{L}_{\text{rank}} = 0$ の場合を考えると，

$$\operatorname{argmin} \mathcal{L}'_{\text{rank-hinge}} = \operatorname{argmin} \mathcal{L}'_{\text{rank-ramp}}$$

でありヒンジ損失を使った場合もランプ損失を使った場合も得られる最適な識別平面は同じである．PU 問題の設定では上の議論から，

$$\operatorname{argmin} \mathcal{L}''_{\text{rank-ramp}} = \operatorname{argmin} \mathcal{L}'_{\text{rank-ramp}}$$

である一方，

$$\operatorname{argmin} \mathcal{L}''_{\text{rank-hinge}} \neq \operatorname{argmin} \mathcal{L}'_{\text{rank-hinge}}$$

でありヒンジ損失のような非対称な関数を代理損失として用いた場合，得られる識別平面は最適なものから誤差が生じてしまうことがわかる．

表 1 実験で比較した手法．それぞれ条件 1（適切に重み付けられた損失関数を使用），条件 2（対称の損失関数を使用）のあるなしに対応する

Table 1 Methods used in experiments. Each method corresponds to whether cond.1 (use of appropriately weighted loss function) and cond.2 (use of symmetric loss function) are met or not.

	Baseline	Method1	Method2	Method3 (proposed)
cond.1 (weighted loss)		✓		✓
cond.2 (symmetric loss)			✓	✓

5. 実験

本研究で導いた条件の有効性を示すために人工データセット，MSCOCO [23]，NUS-WIDE [24] の 3 つのデータセットを用い実験を行った．

5.1 設定

サンプルに付いている正例ラベルを 0% ~ 80% の割合で欠損させ，そのデータを用いて学習し，精度評価を行った．欠損は以下の様にして与えられた．

(1) 全正例ラベル数に欠損率を掛けることで全体の欠損ラベル数 N_{noise} を決定．
(2) 総数が N_{noise} となるように，クラス c に対する欠損ラベル数 N^c_{noise} を多項分布から決定．
(3) クラス c のラベルが付与されているサンプルを N^c_{noise} 個をランダムに選択し負例とした．

評価指標は Average precision のサンプル平均を用いた．条件 1（適切に重み付けられた損失関数を使用），条件 2（対称の損失関数を使用）を適用するかどうか (図では cond.1, cond.2 と表記) で以下の 4 通りの手法を比較した．全ての実験を通して条件 2 の非対称の関数としてヒンジ損失を，対称な関数としてランプ損失を用いた．スコア関数として線形な関数を用い，確率的勾配降下法により更新した．つまり，$\mathbf{f}(\mathbf{x}) = \mathbf{W}^T\mathbf{x}$ であり，$\frac{1}{N}\sum_{n=1}^{N} l(\mathbf{f}(\mathbf{x}_n), \mathbf{y}_n)$ なる損失関数に対して，

$$\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta^{(\tau)}\frac{\partial l(\mathbf{f}(\mathbf{x}_n), \mathbf{y_n})}{\partial \mathbf{W}}$$

のように更新を行う．N は学習サンプル数，$\eta^{(\tau)}$ は学習係数，τ はイテレーション数を表し本実験では $\eta^{(\tau)} = \eta^{(0)}/\sqrt{\tau}$ とした．

5.2 人工データ

次元数 100 のサンプル，クラス数 50 のマルチラベルデータセットを以下の処理によって生成した．

(1) ラベル数 n をポアソン分布からサンプリングする．
(2) 属するクラス c_i を n 回多項分布からサンプリングする．
(3) サンプルしたクラス $\{c_1, c_2, ..., c_n\}$ をラベルとする．

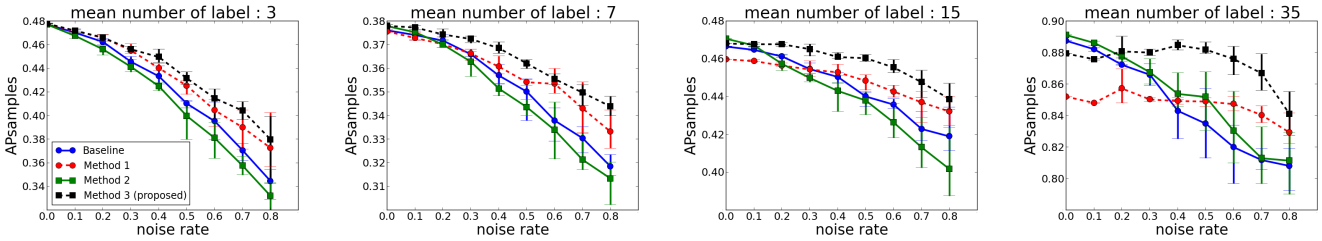

図 3 ラベル欠損を含む人工データを用いた実験結果．左からそれぞれ平均ラベル数 (3, 7, 15, 35) の結果．提案した条件を両方満たす手法が最もラベルの欠損に対して頑強となる．

Fig. 3 Experimental results on synthetic dataset. Each figure from left corresponds to different mean number of label (3, 7, 15, 35). The Method which meets both conditions has robustness to label deficit.

(4) 各クラスからのサンプリング回数 k をポアソン分布からサンプリングする．
(5) k 回多項分布からサンプリングを行い $\mathbf{w}_j$ とする．
(6) サンプリングした $\mathbf{w}_j$ の和 $\sum_j \mathbf{w}_j$ を特徴量とする．

多項分布のパラメータは一様分布からサンプリングされ決定される．サンプルは合計で 10000 サンプル生成し，学習に 8000 サンプル，評価に 2000 サンプルを用いた．サンプル回数の平均値を 50 とした．またサンプルに付与されるラベル数の平均値を $(3, 7, 15, 35)$ と変化させてデータセットを生成し，それぞれに対して学習，評価を行った．

実験結果は**図 3** のようになった．図の左からサンプルに付与されるラベル数の平均値を $(3, 7, 15, 35)$ とした時の結果である．図 3 から全ての条件において本研究で導出した条件を両方満たす損失関数を最小化する方法 (Method3) が最も頑強に学習できていることが分かる．また，サンプルに付与されるラベル数が多くなればなるほど，ラベルを欠損させた時に提案する条件を両方満たす手法 (Method3) と他手法との精度の差が大きくなることが分かる．

5.3 画像アノテーションデータセット

MSCOCO : MSCOCO は，画像に存在する物体のラベルに加えセグメンテーションされた物体領域の情報やキャプション等を含むデータセットである．本実験においては画像と付随する物体ラベルの情報のみを用いた．また 1 枚の画像に同一物体が複数存在する場合は重複を除き 1 つの物体とした．学習に 82,783 サンプル，評価に 40,504 サンプル用い，クラス数は 80 である．また画像の特徴量として，ILSVRC2012 のデータセットで学習済みの Alex-Net [25] の第 7 層目の中間出力を用いた．

NUS-WIDE : NUS-WIDE は Flickr の画像にタグ付けを行った画像アノテーション用のデータセットで，269,648 枚の画像とそれに対応する 5,018 種類のタグを含む．本実験ではデータセット内で指定されている 81 クラスを用いた．学習に 161,789 サンプル，評価に 107,859 サンプルを用い，画像特徴量としてデータセットに含まれている SIFT [26]+BoVW を用いた．特徴量の次元は 500 次元である．

MSCOCO と NUS-WIDE における実験結果はそれぞれ**図 4**，**図 5** のようになった．両データセットにおいて条件を両方満たす損失関数を最小化する方法 (Method3) が最もラベルの欠損に対して頑強に学習が行えていることが分かる．

6. 考察

人工データにおける実験結果から，(A) 欠損率が大きくなるほど，(B) サンプルに付与されるラベル数が多くなるほど条件 1 のある手法（Method1，Method3）とない手法 (Baseline，Method2) の精度の差が大きくなる結果が得られた．この結果は以下の式から説明することができる．条件 1 のように重みを変えなかった場合の損失関数を $\hat{\mathcal{L}}_{\mathrm{rank}}$ とすれば，真の損失関数との差 $\mathcal{L}_{\mathrm{rank}} - \hat{\mathcal{L}}_{\mathrm{rank}}$ は（a）実際は正例であるにも関わらずラベルが付いていないサンプルの割合に比例する項（b）2 つのクラスの両方にラベルが付いているサンプルの割合に比例する項に分解することができる．（補足 A.3）(a) の項は（A）の結果に対応し，（b）の項は（B）の結果に対応する．

更に条件 1 を満たす場合（C）ラベル数が増えれば条件 2 を満たす手法（Method3）と満たさない手法（Method1）の差が大きくなる結果が得られた．これは，PU 設定のデータに対して代理損失を用いることによって生じる誤差 $\mathcal{L}''_{\mathrm{rank}} - \mathcal{L}'_{\mathrm{rank}}$ が 2 つのクラスが同時に正例となる確率に比例しているからであると考えられる．

ラベル数が増えた場合，Method3 では欠損率が低い場合においても精度が悪くなる理由は，本研究では条件を導く過程で”case-controlled”の仮定をおいているため，この条件を満たさない場合は誤差が生じるためであると考えられる． 特に欠損率が 0%の場合，つまり完全にラベルが付与されている場合において，この誤差は非対称な代理損失関数を用いた場合はキャンセルすることができる．（補足 A.4）

7. 結論

本稿では，マルチラベル分類の本質的な問題である不完

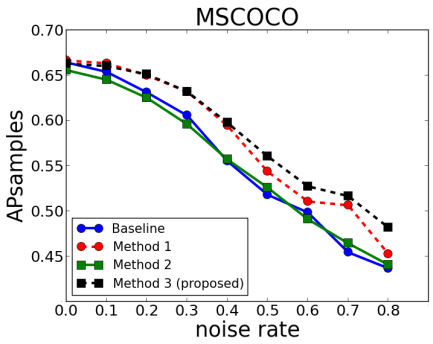

図 4 MSCOCO データセットにおける実験結果.
Fig. 4 Experimental result on MSCOCO dataset.

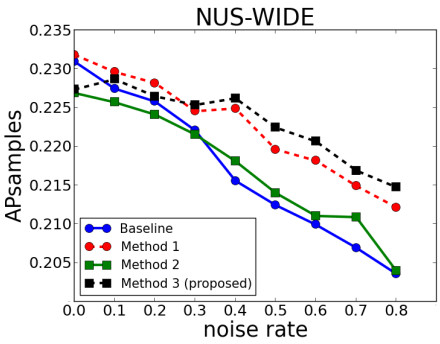

図 5 NUS-WIDE データセットにおける実験結果.
Fig. 5 Experimental result on NUS-WIDE dataset.

全なラベル付けデータからの学習に焦点を当てた．不完全なラベル付けされたデータからでも頑強に学習可能な方法を導くため，この問題を 2 値の PU 分類問題を拡張したマルチラベルの PU 分類問題として捉え，2 値の PU 分類問題に対する解析を拡張することによりマルチラベル PU 分類問題において学習を頑強に行うための損失関数の条件を示した．更に，複数のデータセットを用いた実験によってこれらの条件の有効性を示した．

謝辞 本研究は，JST CREST における研究領域「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」の研究課題「膨大なマルチメディアデータの理解・要約・検索基盤の構築」の支援により行ったものである．また本論文を完成させるにあたって，東京大学大学院 新領域創成科学研究科 杉山将教授には大変有意義かつ的確なご意見を頂きましたことを，この場をお借りして感謝申し上げます．

参考文献

[1] Cheng, W., Hüllermeier, E. and Dembczynski, K. J.: Bayes optimal multilabel classification via probabilistic classifier chains, *ICML, 2010*.
[2] Kapoor, A., Viswanathan, R. and Jain, P.: Multilabel classification using bayesian compressed sensing, *NIPS, 2012*.
[3] Chen, Y.-N. and Lin, H.-T.: Feature-aware label space dimension reduction for multi-label classification, *NIPS, 2012*.
[4] Carneiro, G., Chan, A. B., Moreno, P. J. and Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 29, No. 3, pp. 394–410 (2007).
[5] Weston, J., Bengio, S. and Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation, *IJCAI, 2011* (2011).
[6] Guillaumin, M., Mensink, T., Verbeek, J. and Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation, *ICCV, 2009*.
[7] Von Ahn, L. and Dabbish, L.: Labeling images with a computer game, *CHI, 2004*.
[8] Elkan, C. and Noto, K.: Learning classifiers from only positive and unlabeled data, *SIGKDD, 2008*.
[9] Lee, W. S. and Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression, *ICML, 2003*.
[10] Du Plessis, M. C., Niu, G. and Sugiyama, M.: Analysis of Learning from Positive and Unlabeled Data, *NIPS, 2014*.
[11] Elisseeff, A. and Weston, J.: A kernel method for multi-labelled classification, *NIPS, 2001*.
[12] Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S. and Singer, Y.: Online passive-aggressive algorithms, *The Journal of Machine Learning Research*, Vol. 7, pp. 551–585 (2006).
[13] Bucak, S. S., Mallapragada, P. K., Jin, R. and Jain, A. K.: Efficient multi-label ranking for multi-class learning: application to object recognition, *ICCV, 2009*.
[14] Akata, Z., Perronnin, F., Harchaoui, Z. and Schmid, C.: Label-embedding for attribute-based classification, *CVPR, 2013*.
[15] Gong, Y., Jia, Y., Leung, T., Toshev, A. and Ioffe, S.: Deep convolutional ranking for multilabel image annotation, *arXiv preprint arXiv:1312.4894* (2013).
[16] Elkan, C.: The foundations of cost-sensitive learning, *International joint conference on artificial intelligence*, Vol. 17, No. 1, Citeseer, pp. 973–978 (2001).
[17] Bucak, S. S., Jin, R. and Jain, A. K.: Multi-label learning with incomplete class assignments, *CVPR, 2011*.
[18] Kong, X., Wu, Z., Li, L.-J., Zhang, R., Yu, P. S., Wu, H. and Fan, W.: Large-Scale Multi-Label Learning with Incomplete Label Assignments.
[19] Li, X., Zhao, F. and Guo, Y.: Conditional Restricted Boltzmann Machines for Multi-label Learning with Incomplete Labels, *AISTATS, 2015*.
[20] Menon, A., Rooyen, B. V., Ong, C. S. and Williamson, B.: Learning from Corrupted Binary Labels via Class-Probability Estimation, *ICML, 2015*.
[21] Du Plessis, M. C. and Sugiyama, M.: Class prior estimation from positive and unlabeled data, *IEICE TRANSACTIONS on Information and Systems*, Vol. 97, No. 5, pp. 1358–1362 (2014).
[22] Blanchard, G., Lee, G. and Scott, C.: Semi-supervised novelty detection, *The Journal of Machine Learning Research*, Vol. 11, pp. 2973–3009 (2010).
[23] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft COCO: Common objects in context, *ECCV, 2014*.
[24] Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z. and Zheng, Y.: NUS-WIDE: a real-world web image database from National University of Singapore, *CIVR, 2009*.
[25] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, *NIPS, 2012*.
[26] Lowe, D. G.: Distinctive image features from scale-invariant keypoints, *International journal of computer vision*, Vol. 60, No. 2, pp. 91–110 (2004).

付　　録

A.1

2 つの条件 $P(*|s_i = 0) = P(*)$, $P(*|s_i = 1) = P(*|y_i = 1)$ から

$$\begin{aligned}
&P(f_i < f_j \mid s_i = 1, s_j = 0)\\
=\ &P(f_i < f_j \mid y_i = 1, s_j = 0)\\
=\ &P(y_j = 0 \mid y_i = 1, s_j = 0)P(f_i < f_j \mid y_i = 1, y_j = 0, s_j = 0)\\
+\ &P(y_j = 1 \mid y_i = 1, s_j = 0)P(f_i < f_j \mid y_i = 1, y_j = 1, s_j = 0)\\
=\ &P(y_j = 0 \mid y_i = 1)P(f_i < f_j \mid y_i = 1, y_j = 0)\\
+\ &P(y_j = 1 \mid y_i = 1)P(f_i < f_j \mid y_i = 1, y_j = 1).
\end{aligned}$$

同様に,

$$\begin{aligned}
&\frac{1}{2}P(f_i = f_j \mid s_i = 1, s_j = 0)\\
=\ &\frac{1}{2}P(y_j = 0 \mid y_i = 1)P(f_i = f_j \mid y_i = 1, y_j = 0)\\
+\ &\frac{1}{2}P(y_j = 1 \mid y_i = 1)P(f_i = f_j \mid y_i = 1, y_j = 1).
\end{aligned}$$

合わせて

$$\begin{aligned}
&R_X(i, j)\\
=\ &P(f_i < f_j \mid s_i = 1, s_j = 0) + \frac{1}{2}P(f_i = f_j \mid s_i = 1, s_j = 0)\\
=\ &(1 - \pi_{ij})R(i, j) + \pi_{ij}R_{-X}(i, j)
\end{aligned}$$

A.2

式 (5) から,

$$\begin{aligned}
&P(y_i = 1, y_j = 0)R(i, j)\\
=\ &\frac{P(y_i = 1, y_j = 0)}{P(y_j = 0|y_i = 1)}(R_X(i, j) - P(y_j = 1|y_i = 1)R_{-X}(i, j))\\
=\ &P(y_i = 1)R_X(i, j) - P(y_i = 1, y_j = 1)R_{-X}(i, j).
\end{aligned}$$

これと，式（2）より,

$$\begin{aligned}
&\mathcal{L}_{\text{rank}}\\
=\ &\sum_{1\le i<j\le m} P(y_i = 1, y_j = 0)R(i, j) + P(y_i = 0, y_j = 1)R(j, i)\\
=\ &\sum_{1\le i<j\le m} P(y_i = 1)R_X(i, j) + P(y_j = 1)R_X(j, i)\\
&-P(y_i = 1, y_j = 1)\Big\{R_{-X}(i, j) + R_{-X}(j, i)\Big\}\\
=\ &\sum_{1\le i<j\le m} P(y_i = 1)R_X(i, j) + P(y_j = 1)R_X(j, i)\\
&-P(y_i = 1, y_j = 1)
\end{aligned}$$

となる.

A.3

$$\begin{aligned}
&P(y_i = 1) - P(s_i = 1, s_j = 0)\\
=\ &\Big\{P(y_i = 1, s_i = 0) + P(y_i = 1, s_i = 1)\Big\}\\
&-P(s_i = 1, y_i = 1, s_j = 0)\\
=\ &P(y_i = 1, s_i = 0) + P(y_i = 1, s_i = 1, y_j = 1, s_j = 1)\\
=\ &P(y_i = 1, s_i = 0) + P(s_i = 1, s_j = 1)
\end{aligned}$$

であるので,

$$\begin{aligned}
&\mathcal{L}_{\text{rank}} - \hat{\mathcal{L}}_{\text{rank}}\\
=\ &\sum_{1\le i<j\le m} P(y_i = 1)R_X(i, j) + P(y_j = 1)R_X(j, i) - \text{const}\\
-\ &\sum_{1\le i<j\le m} P(s_i = 1, s_j = 0)R_X(i, j) + P(s_i = 0, s_j = 1)R_X(j, i)\\
=\ &\sum_{1\le i<j\le m} P(s_i = 1, s_j = 1)\Big\{R_X(i, j) + R_X(j, i)\Big\}\\
&+P(y_i = 1, s_i = 0)R_X(i, j) + P(y_j = 1, s_j = 0)R_X(j, i)\\
&+\text{const}
\end{aligned}$$

第 1 項目はクラス i, j の両方のラベルが付いているサンプルの割合に，2, 3 項目はそれぞれのクラスで正例かつラベルが付いていないサンプルの割合に比例する.

A.4

完全にラベルが付与されている場合，最小化すべき損失関数は式（2）より

$$\begin{aligned}
&\mathcal{L}_{\text{rank}}\\
=\ &\sum_{1\le i<j\le m} P(y_i = 1, y_j = 0)R(i, j) + P(y_i = 0, y_j = 1)R(j, i)
\end{aligned}$$

である．一方，”case-controlled”の仮定の基で導かれた不完全ラベル付きのデータを用いた損失関数は,

$$\begin{aligned}
&\mathcal{L}_{\text{rank}}\\
=\ &\sum P(y_i = 1)R_X(i, j) + P(y_j = 1)R_X(j, i) - P(y_i = 1, y_j = 1)
\end{aligned}$$

である．実際に欠損が存在しなかった場合，$R(i, j) = R_X(i, j)$ となるため,

$$\begin{aligned}
&\mathcal{L}_{\text{rank-false}}\\
=\ &\sum P(y_i = 1)R(i, j) + P(y_j = 1)R(j, i) - P(y_i = 1, y_j = 1)
\end{aligned}$$

となる．これらの差は,

$$\begin{aligned}
&\mathcal{L}_{\text{rank-false}} - \mathcal{L}_{\text{rank}}\\
=\ &\sum_{1\le i<j\le m} P(y_i = 1, y_j = 1)\Big\{R(i, j) + R(j, i) - 1\Big\}
\end{aligned}$$

となり，このクラス i, j の同時確率に比例した誤差が生じる．用いる代理損失関数として対称な関数を選択すればこの誤差はキャンセルされる.