The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

相対密度比を用いたマルチラベルロジスティック回帰について

Multi-label Logistic Regression Using Relative Density Ratio

岡部 格明 *1 Masaaki Okabe　土田 潤 *2 Jun Tsuchida　宿久 洋 *3 Hiroshi Yadohisa

*1同志社大学大学院文化情報学研究科
Graduate School of Cluture and Infomation Science, Doshisha University

*2東京理科大学工学部
Faculty of Engineering, Tokyo University of Science

*3同志社大学文化情報学部
Faculty of Culture and Information Science, Doshisha University

Multi-label classification is a supervised learning problem where multiple labels may be assigned to each instance. The main baseline for multi-label classification is binary relevance method, which is estimate the binary classification model for each label. In binary classification, there are cases where poor results are data when the class is imbalance.

In this paper, we propose a multi-label classification model used relative density ratio. In this model, we used relative F-measure by relative density ratio for weight of error function to solve the class imbalance problem.

1. はじめに

機械学習における分類問題のひとつにマルチラベル分類問題がある．マルチラベルとは，各対象に複数のラベルが付与されているような状況であり，マルチラベル分類問題は，対象ごとに付与されるであろうラベルを推定するような問題である．この問題は，文書や画像の分類などに応用されている．

マルチラベル分類問題を解決するための方法には，各ラベルを一つのクラスとしてみなし，マルチクラスの問題として解く，Label Power Set 法 [Tsoumakas 07] と各ラベルをそれぞれのクラスに属するかそうでないかを 2 値分類の問題として解く，Binary relevance 法 [Boutell 04] がある．Label Power Set 法は，データ内に存在しているラベルの組み合わせに対して，一つのラベルを割り当てるため，ラベル同士の関係性を反映した判別を実現することができる．一方で，データに存在しないラベルの組み合わせに対しては，対応できないという欠点がある．また，ラベルの種類が 2 の累乗の形になってしまうため，対象の数が少ないデータセットに適用したときに，過適合を起こす場合がある．Binary relevance 法は，それぞれのラベルのみを考慮するため，ラベル間の関係性に着目することができないが，計算コストが少ないという利点がある．

一般に，マルチラベル分類問題では，データの数に比べ，あるラベルが付与されている割合が極端に小さいという，ラベルのクラス不均衡の問題が起こりやすいと言われている [Sun 13]．このような場合に，評価指標として単純な正解率を用いると，全てのデータにラベルが付いていないという予測結果が得られたとしても，高い正解率を得てしまう場合がある．正解率のほかの評価指標として F-measure があり，この指標は，先ほどのような状況に対して，「悪い」判別であることを表すという性質がある．F-measure の算術平均が macro F-measure であり，ある特定のクラスの判別が「悪い」場合にも対応する指標である．

そこで，本発表では，Binary relebance 法による，近似された macro F-measure を重みとして用いた，ロジスティック回帰を提案する．この手法では，各ラベルの混同行列における関係から，相対 F-measure を相対密度比によって近似する．それを誤差関数の重みとして用いることで，正解率だけでなく，F-measure を考慮した最大化問題を解くことになり，ラベルのクラス不均衡の問題を解決する．

連絡先: 岡部格明，同志社大学大学院文化情報学研究科，京都府京田辺市多々良京都谷 1-3，0774-65-7657，mokab.0328@gmail.com

2. 問題設定と評価指標

データ行列 $\boldsymbol{X} = (\boldsymbol{x}_1, .., \boldsymbol{x}_i, .., \boldsymbol{x}_N)' \in \mathbb{R}^{N\times D}$，各対象のラベル行列 $\boldsymbol{Y} = (\boldsymbol{y}_1, .., \boldsymbol{y}_i, .., \boldsymbol{y}_N)' \in \{0,1\}^{N\times C}$ が与えられたとする．Binary relebance 法では，対象 i のデータ $\boldsymbol{x}_i$ が与えられたとき，その対象 i がラベル c を持つか否かを表す関数 $f_c : \mathbb{R}^D \to \{0,1\}$ $(c = 1, ..., C)$ を推定することを目的とする．

また，推定結果の評価指標として，各ラベル c に属するか否かを表した，混同行列 (表 1) を得る．混同行列の各要素 $\mathrm{TP}(c)$, $\mathrm{FN}(c)$, $\mathrm{FP}(c)$, $\mathrm{TN}(c)$ はそれぞれ，$y_{ic} = 1$ であり，$f_c(\boldsymbol{x}_i) = 1$ であった個数，$y_{ic} = 1$ であり，$f_c(\boldsymbol{x}_i) = 0$ であった個数，$y_{ic} = 0$ であり，$f_c(\boldsymbol{x}_i) = 1$ であった個数，$y_{ic} = 0$ であり，$f_c(\boldsymbol{x}_i) = 0$ であった個数を表している． このとき，ラベル c に対する

表 1: ラベル c に対する混同行列

		予測ラベル		合計
		c	c でない	
正解ラベル	c	$\mathrm{TP}(c)$	$\mathrm{FN}(c)$	n_c
	c でない	$\mathrm{FP}(c)$	$\mathrm{TN}(c)$	$N - n_c$
合計		m_c	$N - m_c$	N

Precision(再現率) と Recall(適合率) は次のように定義される．

$$\mathrm{Precision}(c) = \frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FP}(c)}, \quad \mathrm{Recall}(c) = \frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FN}(c)}.$$

一般に，Precision と Recall はトレードオフの関係にあるため，総合指標として，Precision と Recall に対して重みをつけた調和平均である，相対 F-measure という指標がつぎのように定義される．

$$\mathrm{F\text{-}measure}(\alpha, c) = \left(\alpha \frac{1}{\mathrm{Recall}(c)} + (1-\alpha)\frac{1}{\mathrm{Precision}(c)}\right)^{-1} \tag{1}$$

ただし，$0 \leq \alpha \leq 1$ である．

マルチラベル問題としての総合的な判別の評価の指標はそれぞれの指標の平均が用いられる．中でも，(1) 式の算術平均をとったものが macro F-measure であり，つぎのように定義される．

$$\text{F-measure}_{\text{macro}}(\boldsymbol{\alpha}) = \frac{1}{C}\sum_{c=1}^{C} \text{F-measure}(\alpha_c, c)$$

macro F-measure は，出現頻度の少ないラベルがあるときに，小さな値を取りやすいため，判別全体の「よさ」をうまく測ることができる．

3. 提案手法

データ行列 $\boldsymbol{X} \in \mathbb{R}^{N\times D}$，ラベル行列 $\boldsymbol{Y} = (\boldsymbol{y}_i) \in \{0,1\}^{N\times C}$ が与えられたとする．

このとき，各ラベルに対して，ロジスティック回帰モデルをつくるため，対象 i にラベル c が付く確率を次のように表す．

$$\mathrm{P}(y_{ic} = 1|\boldsymbol{x}_i) = p_{ic} = \frac{1}{1+\exp(-\boldsymbol{\beta}_c'\boldsymbol{x}_i)}. \tag{2}$$

ここで，$\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_C) \in \mathbb{R}^{D\times C}$，とすると，$\boldsymbol{B}$ について最小にするべき目的関数は以下のように表される．

$$L_{\boldsymbol{\beta}}(\boldsymbol{B}|\boldsymbol{\Theta}) = -\sum_{c=1}^{C}\sum_{i=1}^{N} w_{\alpha}^{c}(p_{ic}|\hat{\boldsymbol{\theta}}_c) \{y_{ic}\log p_{ic} + (1-y_{ic})\log(1-p_{ic})\} + \lambda_{\boldsymbol{\beta}}\sum_{k=1}^{C}||\boldsymbol{\beta}_c||^2. \tag{3}$$

ここで，$w_{\alpha}^{c}(p_{ic}|\hat{\boldsymbol{\theta}}_c)$ は相対密度比 [Yamada 13]，$\lambda_{\boldsymbol{\beta}} > 0$ は正則化パラメータ，$||\cdot||$ はフロベニウスノルムである．各クラスへの所属を表す y_{ic}，予測された各クラスへの所属確率 p_{ic} がそれぞれ連続型確率分布 P_{y_c}，P_{p_c} に独立同一に従うと仮定すると，確率分布 P_{y_c}，P_{p_c} の相対密度比は，それぞれの確率密度関数を f_{y_c}, f_{p_c} としたとき，以下のように定義される．

$$w_{\alpha}^{c}(p) = \frac{f_{p_c}(p)}{\alpha f_{y_c}(p) + (1-\alpha) f_{y_c}(p)}.$$

クラス c に対して，「理想的な判別」が行われた状況を考えると，混同行列における $\mathrm{FN(c)} = 0$ となることから，このときの相対 F-measure を以下のように表すことができる．

$$\text{F-measure}(\alpha, c) = \frac{n_c}{\alpha n_c + (1-\alpha) m_c} \tag{4}$$

n_c，m_c をそれぞれ f_{y_c}，f_{p_c} と置く．すると，それぞれ各対象 1 つ 1 つに対して「分類に寄与する度合い」のようなものと解釈することができ，これは相対 F-measure の形になる．目的関数 (3) 式の $w_{\alpha}^{c}(p_{ic}|\hat{\boldsymbol{\theta}}_c)$ の c に関する和の部分が全体での macro F-measure に対応している．

次に，相対密度比の推定については，[Kanamori 09] で提案されている Unconstrained Least-Squares Importance Fitting (uLSIF) に基づく方法で推定する．いま，相対密度比を $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_N)' \in \mathbb{R}_{+}^{N}$ として以下の通りにモデル化する．

$$w(p|\boldsymbol{\theta}) = \sum_{i=1}^{N} \theta_i K(p, p_i).$$

ここで，$K(p, p_i)$ は，カーネル関数であり，通常は RBF カーネルが使われる [Kanamori 09]．$\boldsymbol{\theta}_c$ の推定値は，

$$L_{\boldsymbol{\theta}}(\boldsymbol{\theta}_c|\boldsymbol{\beta}_c) = \frac{1}{2}\boldsymbol{\theta}_c'\hat{\boldsymbol{H}}_c\boldsymbol{\theta}_c - \hat{\boldsymbol{h}}_c'\boldsymbol{\theta}_c + \frac{\lambda_{\boldsymbol{\theta}}}{2}||\boldsymbol{\theta}_c||^2,$$

最小にすることで得られ，その解は $\hat{\boldsymbol{\theta}}_c = \max(\boldsymbol{0}, (\hat{\boldsymbol{H}}_c + \lambda_{\boldsymbol{\theta}}\boldsymbol{I}_N)^{-1}\hat{\boldsymbol{h}}_c)$ によって相対密度比の条件を満たす形で得られる．ここで，$\lambda_{\boldsymbol{\theta}}$ (≥ 0) は正則化パラメータであり，$\boldsymbol{I}_N$ は N 次単位行列である．さらに，$\hat{\boldsymbol{H}}_c \in \mathbb{R}^{N\times N}$ は第 (ℓ, m) 要素 $\hat{H}_{c\ell m}$ が $\hat{H}_{c\ell m} = \frac{\alpha}{N}\sum_{i=1}^{N} K(p_{ic}, p_{\ell c})K(p_{ic}, p_{mc}) + \frac{(1-\alpha)}{N}\sum_{i=1}^{N} K(y_{ic}, p_{\ell c})K(y_{ic}, p_{mc})$ である行列で，$\hat{\boldsymbol{h}}_c \in \mathbb{R}^N$ は，第 ℓ 要素に $\hat{h}_{c\ell} = \frac{1}{N}\sum_{i=1}^{N} K(p_{ic}, p_{\ell c})$ を持つベクトルである．正則化パラメータ $\lambda_{\boldsymbol{\theta}}$，RBF カーネルのバンド幅はクロスバリデーションによって定められる．

推定を行うためのアルゴリズムはつぎの通りである．

入力 データ行列 $\boldsymbol{X} \in \mathbb{R}^{N\times D}$，ラベル行列 $\boldsymbol{Y} = (\boldsymbol{y}_i) \in \{0,1\}^{N\times C}$，相対密度比に対するパラメータ α

出力 パラメータ行列 $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_C) \in \mathbb{R}^{P\times C}$

STEP0 各クラス $y_c \in \{0,1\}^D$ を目的関数としたロジスティック回帰のパラメータベクトル $\boldsymbol{\beta}_c$ $(c = 1, ..., C)$ を求め，$\boldsymbol{B}_0 = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_C)$ とする．

STEP1 得られたパラメータ行列 $\boldsymbol{B}_0$ を用いたモデル ((2) 式) による予測結果 $\hat{\boldsymbol{P}} = (p_{ic}) = (1/1+\exp(-\boldsymbol{\beta}_c'\boldsymbol{x}_i))$ を求める．

STEP2 各クラス $(c = 1, ..., C)$ に対するラベルベクトル $\boldsymbol{y}_c$ と各対象のクラス c への所属確率のベクトル $\boldsymbol{p}_c$ の α-相対密度比に関するパラメータ $\boldsymbol{\theta}_c$ を uLSIF によって推定し，$\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_C)$ とする．

STEP3 前の STEP で推定したパラメータ $\boldsymbol{\Theta}$ を用いて，(3) 式の $L(\boldsymbol{B}|\boldsymbol{\Theta})$ を最小にするパラメータ行列 $\boldsymbol{B}$ を求める．

STEP4 $||\boldsymbol{B}_0 - \boldsymbol{B}||^2 < \varepsilon$ ならば終了する．そうでない場合は $\boldsymbol{B}_0 = \boldsymbol{B}$ として，STEP1 に戻る．

参考文献

[Boutell 04] Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. (2004). Learning multi-label scene classification. Pattern recognition, **37**, 1757-1771.

[Kanamori 09] Kanamori, T., Hido, S., and Sugiyama, M.(2009). A Least-squares Approach to Direct Importance Estimation, *Journal of Machine Learning Research*, **10**, 1391–1445.

[Sun 13] Sun, L., Ji, S., and Ye, J. (2013). Multi-label dimensionality reduction. CRC Press.

[Tsoumakas 07] Tsoumakas, G., Vlahavas, I.(2007). Random k-labelsets: An ensemble method for multilabel classification., Proceedings of the 18th European Conference on Machine Learning (ECML 2007), Warsaw, Poland (2007) 406–417.

[Yamada 13] Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M.. (2013). Relative density-ratio estimation for robust distribution comparison. *Neural computation*, **25**, 1324–1370.