F値を最大化するテキストの多重ラベリング法

賀沢秀人　　泉谷知範　　平博順　　前田英作
{kazawa,izumi,taira,maeda}@cslab.kecl.ntt.co.jp
ＮＴＴコミュニケーション科学基礎研究所

概 要

テキストに対して複数のラベルを付与する多重ラベリング問題において，高いF値 (=ラベリングの適合率と再現率の調和平均) を達成し，かつ，ノイズに対しても頑健なラベリング学習アルゴリズムを提案する．

本アルゴリズムでは，ラベル集合を F-space と呼ぶベクトル空間に埋め込んだのち，間違ったラベリングにたいするマージンが大きくなるように線形写像を推定することで，ラベリングの学習を行う．

実際の医学生物学文献と Web ページとを用いた実験で，SVM や最近傍法などの 識別学習器に比べてF値が向上することがわかった．

1　テキスト多重ラベリング問題

本研究では，ラベルが付与されたテキストがサンプルとして与えられたときに，ラベリング規則を自動的に学習する方法を提案する．なお，ここでは一つのテキストに同時に複数のラベルを付与することを想定しており，通常一つのラベルのみ付与するテキスト分類 [1] と比較すると，ラベル間の相関情報を用いることができる点が特徴である．以下では，テキスト分類との区別のため，本研究の問題設定を「テキスト多重ラベリング」もしくは単に「ラベリング」と呼ぶことにする．

テキスト多重ラベリングの研究が盛んになってきたのは，比較的最近である [2, 3, 4]．Schapire と Tingaud は，与えられたサンプルにおいて，ラベルの有無を正しく判定した割合 (accuracy) が最大になるように学習する手法を提案している [2, 4]．しかし，上田が [3] において指摘しているように，ラベルの種類が多い場合には,「常にラベルを付与しない」という規則を用いることで，容易に高い accuracy が得られるという問題がある．そこで，[3] では評価手法として，ラベリングのF値が用いられているが，そこで提案された手法自体はF値を最大化するような学習となっていない．

このような状況を踏まえ，本研究では，テキスト多重ラベリングにおいて，F値を最大化するような学習方法を提案する．

2　F値によるラベリング評価

F値は，ある正解集合 A とそれにたいする予測集合 B が与えられたときに，その「差」を評価する尺度の一つで次のように定義される．

$$F(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

F値は，情報検索で良く用いられる尺度である適合率と再現率の調和平均に等しい [5]．

本研究のラベリング問題においては，A をあるテキストに実際に付与されたラベルの集合とし，B をそのテキストにシステムが付与したラベルの集合として，F値を用いることにする [3]．この点，情報検索で用いられるF値とは計測する対象が異なっているので，注意して頂きたい．[1]

一方，[2, 4] で用いられている正解率 (accuracy) は，次のように定義される．

$$Acc(A,B) = \frac{|A \cap B| + |A^c \cap B^c|}{|\mathcal{L}|} \tag{2}$$

ここで，$\mathcal{L}$ はラベリングに用いるラベル全体の集合である．式（2）では，実際には付与されなかったラベル（A^c）にたいする精度と，付与されたラベル（A）にたいする精度が，同じ重みで正解率に寄与している．そのため，テキスト多重ラベリングでしばしば起こるように，テキストに付与されるラベルのほうが遥かに少ない場合（$|A| \ll |A^c|$）,「常にラベルを付与しない」という規則（$B = \phi$）を用いるだけで，大きな正解率が得られてしまう．

それに対し，高いF値を達成できるような自明な規則は存在しない．そのため，ラベリング問題においては正解率よりも適切な評価尺度であると考えられる．

[1] 情報検索でF値を用いる場合は，A は検索要求に適合するテキスト集合であり，B は検索システムが出力するテキスト集合である．

3　最大マージンラベリング法

本節では，F値を最大化する学習アルゴリズムとして，**最大マージンラベリング法**を提案する．

いま，テキストは適当な方法により $\mathbb{R}^n$ のベクトル $\mathbf{x}$ に変換されているとする．さらに，Ω は全てのラベル組合せからなる集合とする．また，テキストへのラベリング規則を $f : \mathbb{R}^n \mapsto \Omega$ と書く．すると，サンプル $(\mathbf{x}_1, L_1), \ldots, (\mathbf{x}_m, L_m)$ （$\mathbf{x}_i \in \mathbb{R}^n, L_i \in \Omega$）が与えられたときの f のF値は次のように書くことができる．

$$\sum_{i=1}^{m} \frac{2|L_i \cap f(\mathbf{x}_i)|}{|L_i| + |f(\mathbf{x}_i)|}$$

この式は分母と分子の両方に f を含んでいるため，最大化する f を直接求めることは困難である．そこで，本研究では次のような二段階のアプローチを取ることとする．

1. F値が内積に相当するような空間を構成し，その中に全てのラベル組合せを埋め込む．
2. テキストベクトル $\mathbf{x}^i$ を上記空間中で正しいラベリング L_i に近い点に写すように写像 f を求める．

第一段階については，次のような性質を持つベクトル空間 $\mathcal{H}_F$（**F-space** と呼ぶ）の存在が証明できる．[2]

$$\langle \phi_F(A), \phi_F(B) \rangle_F = F(A, B) + O(\epsilon) \qquad (3)$$

ここで，ϕ_F はラベリングを $\mathcal{H}_F$ に埋め込む写像であり，$\langle \cdot, \cdot \rangle_F$ はF-spaceの内積である．また，ϵ は $F(A, B)$ を行列と見たときの最小の固有値であり，$O(\epsilon)$ は ϵ のオーダーの量を表す．数値計算によると ϵ は通常非常に小さいため，式（3）は，F-spaceの内積がF値とほとんど等しいことを表している．

アプローチの第二段階においては，テキスト・ベクトル空間 $\mathbb{R}^n$ から F-space H_F への**線形写像** W のうち，（1）$\mathbf{x}_i$ の像 $W\mathbf{x}_i$ が正しいラベリング L_i の近くに位置し，かつ，（2）L_i 以外の間違ったラベリング ω との境界との間にできるだけ大きなマージン（$\langle \cdots \rangle$ で表される量）を持つような写像を求める．（図 1）

最大マージンラベリング法（Maximal Margin Labeling，MML）　線形写像 $W : \mathbb{R}^n \mapsto \mathcal{H}_F$ で，次の

[2]紙面の都合上，本稿では証明についての詳細は記載しない．基本的には（1）$F(A, B)$ を要素とする行列 $\mathbf{F}$ の大きさは高々有限であり，その固有値は有限の最小値を持つ，（2）適当な正定数倍した単位行列を $\mathbf{F}$ に加えた行列は，正定値行列となる，（3）その行列要素が内積の値と一致するベクトル空間 $\mathcal{H}_F$ が存在する [6]，という三段階で証明される．

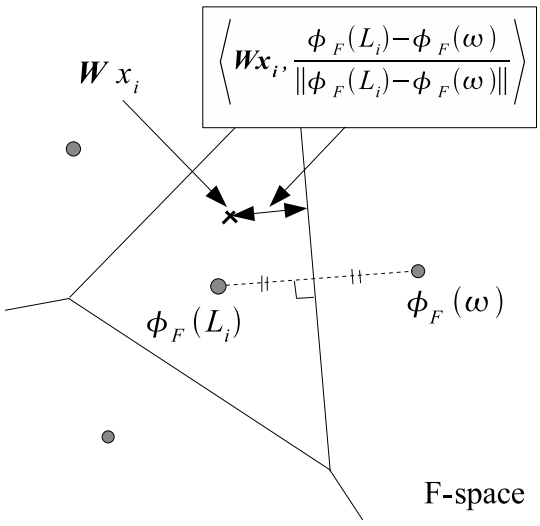

図 1: 最大マージンラベリング法

最適化問題の解となるものを求めよ．

$$\min_W \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{m}\sum_{\omega\in\Omega,\omega\neq L_i} \xi_i^\omega$$
$$\text{s.t.} \left\langle W\mathbf{x}_i, \frac{\phi_F(L_i)-\phi_F(\omega)}{\|\phi_F(L_i)-\phi_F(\omega)\|} \right\rangle_F \geq 1-\xi_i^\omega$$
$$\xi_i^\omega \geq 0 \quad \text{for} \quad 1\leq i\leq m, \forall\omega\in\Omega, \omega\neq L_i. \qquad (4)$$

ただし，必ずしもサンプルに対して正しいラベリングが可能とは限らないので，式（4）においては，条件が破られた度合い ξ_i^ω に応じて，ペナルティ C を課している．

実際に式（4）を解くにあたっては，式（4）と等価な以下の双対問題の形が便利である．

$$\max_{\alpha_i^\omega} \sum_{i=1}^{m}\sum_{\omega\in\Omega,\omega\neq L_i} \alpha_i^\omega - \frac{1}{2}\sum_{i,j=1}^{m}\sum_{\omega\neq L_i}\sum_{\omega'\neq L_j} \alpha_i^\omega\alpha_j^{\omega'}(\mathbf{x}_i\cdot\mathbf{x}_j)$$
$$\times\frac{F(L_i,L_j)-F(L_i,\omega')-F(L_j,\omega)+F(\omega,\omega')}{2\sqrt{(1-F(L_i,\omega))(1-F(L_j,\omega'))}}$$
$$\text{s.t.} \quad 0\leq\alpha_i^\omega\leq C \text{ for } 1\leq i\leq m, \forall\omega\in\Omega, \omega\neq L_i. \quad (5)$$

式（5）は式（4）と異なり，ϕ_F を含んでおらず，直接計算可能な量だけで構成されている．

MMLによるテキスト $\mathbf{x}$ のラベリングは，W により写像された点に最も近いラベリング L を選びだすことで行われる．すなわち，式（4）の解 W，もしくは式（5）の解 α_i^ω を用いて

$$\max_{L\in\Omega} \langle W\mathbf{x}, \phi_F(L)\rangle_F$$
$$= \max_{L\in\Omega} \sum_{i=1}^{m}\sum_{\omega\in\Omega,\omega\neq L_i} \alpha_i^\omega(\mathbf{x}\cdot\mathbf{x}_i)$$

$$\times \frac{F(L,L_i)-F(L,\omega)}{\sqrt{2(1-F(L_i,\omega))}} \qquad (6)$$

の解として求めることができる.

4 実験

提案手法 MML を，医学生物学文献 [4] と Web ページ [3] のラベリングに適用し，性能評価をおこなった．また，同一のデータを用いて，SVM および最近傍法により各ラベルの有無を判別する方法も適用し，MML と比較した．各学習手法の実験時に使用したパラメータは次の通りである．

MML ペナルティCは１とした.

SVM [7, 8] を参考に，正例と負例とで異なるペナルティC_+, C_-を使用し，$C_- = 1, \frac{C_+}{C_-} = \frac{m_+}{m_-}$ （$m_{+(-)}$は正（負）例の数）とした.

最近傍法 近傍事例の数を３および５とした.

表 1 に使用した文書集合に関する基本的な情報を示す.

4.1 医学生物学文献のラベリング

Tingaud は，PubMed[3]に登録されている論文概要のうち，酵母に関する約２０００件に対して，酵母遺伝子の発現形態や機能に関する１１の GO コード [9] をラベリングするという実験をおこなった [4]．以下，この Tingaud が用いたデータを「Bio データ」と呼ぶ.

我々は，Bio データの中からランダムに５０／１００／１５０／２００／２５０文書を取り出し訓練データとし，残った中からランダムに取り出した５００文書をテストデータとし，実験をおこなった．なお，各手法の評価尺度はテストデータにおける平均Ｆ値とし，データの偏りによる影響を減らすために，全ての手順を１０回繰り返した結果をさらに平均した.

文書のベクトル化は，TF×IDF により各単語の重みを決定したのち，長さが１になるように正規化した.

図 2 に全てのテストデータを用いた場合の結果を示す．全ての訓練データ数において，提案手法 MML が他の手法を大きく上回る結果が得られた．また，SVM は比較的データ数の多い場合に精度が高く，逆に，最近傍法はデータ数の少ない場合に高精度という傾向も見られた.

[3]http://www.ncbi.nlm.nih.gov/PubMed

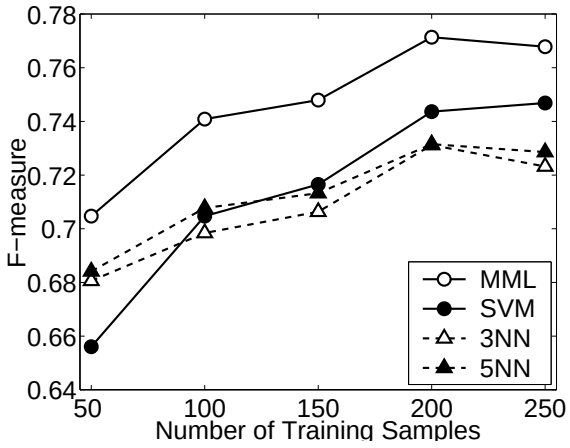

図 2: 全てのテストデータを用いた結果（Bio データ）．横軸：訓練データ数，縦軸：平均Ｆ値．図中，３（５）—ＮＮ は，近傍事例数を３および５とした最近傍法.

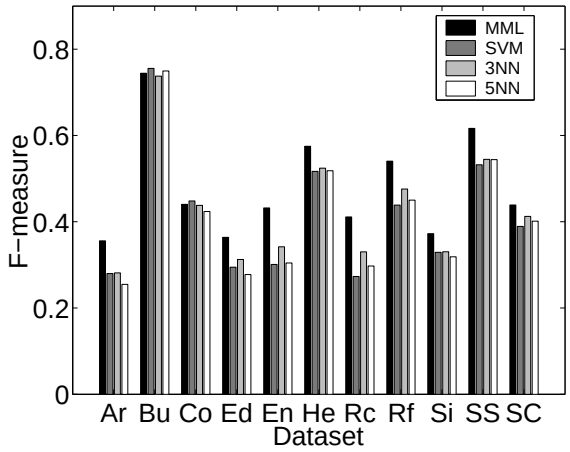

図 3: Ｗｅｂページのラベリング実験結果．データセットの略称は表 1 を参照.

4.2 Ｗｅｂページのラベリング

上田は，検索エンジンのディレクトリサービスからリンクを辿って収集した Web ページを用いて，各ページが割り当てられたサブカテゴリ[4] をラベリングするという実験をおこなった [3]．上田の用いたデータでは，ラベルとなるサブカテゴリ数が２０〜４０程度あったが，本実験では，計算時間の都合上，各データセットごとに出現頻度の多い１０個のサブカテゴリについてラベリングをおこなった．以下，このデータを「WWW データ」と呼ぶ.

[4]トップカテゴリに関しては多重性が少なかったため，トップカテゴリごとに別のデータセットとし，階層が一つ下のサブカテゴリをラベリングすることがタスクとされている.

データセット名 (略称)	語彙数	総ラベル組合せ数 (訓練データに非出現)	文書あたりのラベル数分布 (%) 0	1	2	3	≥4
医学生物学文献 (Bio)	11,701	82 (6.5–16.8%)	0	59.9	23.7	9.3	7.1
Web ページ (WWW)							
Arts & Humanities (Ar)	21,458	133 (10.3%)	7.9	61.1	23.4	6.1	1.7
Business & Economy (Bu)	19,298	50 (3.1%)	2.6	61.2	27.2	7.9	1.2
Computers & Internet (Co)	28,271	62 (4.1%)	7.6	68.1	19.3	4.4	0.7
Education (Ed)	23,894	116 (8.5%)	6.3	62.1	23.6	6.3	1.9
Entertainment (En)	28,550	88 (6.6%)	2.5	75.3	17.6	3.3	1.4
Health (He)	25,166	83 (5.4%)	2.2	57.1	32.0	6.7	2.1
Recreation (Rc)	26,689	98 (7.5%)	7.5	75.6	14.1	2.2	0.6
Reference (Rf)	34,544	55 (2.8%)	9.4	81.7	8.4	0.5	0.1
Science (Si)	33,931	76 (5.5%)	16.9	66.2	14.1	2.3	0.7
Social Science (SS)	39,744	67 (5.3%)	6.8	78.5	12.7	1.7	0.5
Society & Culture (SC)	28,947	143 (10.5%)	9.4	62.0	20.5	5.5	2.8

表 1: 実験で用いた文書集合に関する情報.

実験は，各データセットごとに，ランダムに取り出した２５０ページを訓練データ，２０００ページをテストデータとし，平均Ｆ値を計算した．また，データによる偏りを減らすために，実験を５回繰り返しその平均を取った．

文書ベクトルは [3] と同様に，単語の出現頻度を重みとし，長さが１になるように正規化したものを用いた．

図 3 に実験結果を示す．Bio データと比べると差は顕著ではないが，全てのテストデータで評価した場合，１１データセット中９セットで MML が最も良い性能を示した．(残り２セットでは，SVM が最も良かった．)

参考文献

[1] 永田昌明 and 平博順. テキスト分類 — 学習理論の「見本市」—、. **情報処理**, 21(1), 2001.

[2] Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[3] Naonori Ueda and Kazumi Saito. Single-shot detection of multiple categories of text using parametric mixture models. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 626–631, 2002.

[4] Frédéric Tingaud, Tomonori Izumitani, Hirotoshi Taira, and Eisaku Maeda. Classifying genes into gene otologoy categories using text-based supervised learning methods. In *Proc. of European Conference on Computational Biology 2003*, pages 81–82, 2003.

[5] 北研二, 津田和彦, and 獅々堀正幹. **情報検索アルゴリズム**, chapter 2. 共立出版, 2002.

[6] Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[7] 平博順, 向内隆文, and 春野雅彦. Support vector machine によるテキスト分類. In **情報処理学会研究報告**, volume 98-NL-128-24, pages 173–180, 1998.

[8] Katharina Morik, Peter Brockhausen, and Thorsten Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proc. of the 16th International Conference on Machine Learning*, pages 268–277, 1999.

[9] The Gene Ontology Consortium. Gene ontlogy: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.