期待再現率における期待適合率最大化モデルの学習方法

佐々木 健太郎　　土田 正明　　田村 晃裕
NEC 情報・ナレッジ研究所
{k-sasaki@ez, m-tsuchida@cq, a-tamura@ah}.jp.nec.com

1 はじめに

テキストを効率よく自動分析するため，数多くのテキスト分類手法が提案されている．それらの手法は，例えば，ニュース分類，スパムフィルタリング，意見抽出などに適用され，効果をあげている [1].

本稿では，テキスト分類の中で，特に，正例の事例を負例に誤って分類した場合に致命的な影響を与えてしまうタスクに着目する．以降，正例の事例を負例に誤って分類することを「分類漏れ」と呼ぶ．例えば，企業等において，大量のテキストストリームから不正行為や違法行為が疑われるテキストを分類する場合，分類漏れが生じると漏れた行為に対する対策が取れないため，企業に大きなダメージを与える可能性がある．このような分類漏れを極力少なくしなければならないタスクを想定した機械学習手法を提案する．

単純には，再現率を最大化する学習を行うことで解決できるように見えるかもしれない．しかし，適合率を完全に無視すると，結果として負例の事例を誤って正例と判断してしまうことが多くなり，実用に耐えない分類器になってしまう．したがって，我々が着目するタスクにおいても，再現率だけでなく適合率も考慮することが必要不可欠である．

本稿では，高再現率制約のもとで適合率を最大化する新たな定式化を行う．この定式化により，高再現率，つまり，分類漏れが少ないもとで可能な限り高い適合率でテキストを分類するモデルを学習することができる．提案方式は，F 値最大化問題を F 値の期待値で表すことにより定式化した Jansche の手法 [6] を参考に，再現率と適合率を期待値で表して定式化する．

我々の知る限り，本タスク設定を直接扱った定式化は存在しない．したがって，提案手法は，我々のタスクの必要条件である高再現率のもとで，従来手法に比べて適合率が良くなると期待できる．以降，2.1 で述べる期待 F 値最大化ロジスティック回帰と比較した実験により提案手法の有効性を確認する．

表 1: TP,FP,FN,TN の定義

		真のラベル	
		正例	負例
予測ラベル	正例	TP	FP
	負例	FN	TN

2 定式化

分類結果は，真のラベルと分類器の予測ラベルに応じて 4 種類に分けることができる (表 1)．分類器の評価指標には，適合率 (Precision) と再現率（Recall）が存在する．

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

ここで，TP，FP，FN の個数もそれぞれ TP，FP，FN と表記した．本研究の興味は分類漏れの軽減であるため，高再現率の状態を対象とする．ただし，再現率と適合率にはトレードオフの関係があるため，提案手法では，高再現率における適合率を極力高めるための機械学習を考える．ある値 τ 以上の再現率が必要な時に適合率を最大化する問題は以下の通りとなる．

$$\begin{cases} \text{maximize} & \textbf{Precision} \\ \text{subject to} & \textbf{Recall} \geq \tau \end{cases}$$

上記を満たすようにモデルのパラメタを学習することが提案手法の目的であるが，**Precision** ,**Recall** の計算のための TP, FP, FN が離散値となり，モデルのパラメタに対して不連続な関数となる．結果，微分不可能となるため数値最適化問題としては扱いが難しい．そこで，提案手法では，Jansche の方法 [6] を参考に TP, FP, FN を各々の期待値で置換え，微分可能な連続関数に変換して最適化を行う．

2.1 期待 F 値 (Jansche, 2005)

本節以降，ラベル付きデータ $\{(x_n, y_n)\}_{n\in N}$ のラベル $\{y_n\}_{n\in N}$ は $\{1, -1\}$ の二値とし，値が $+1$ のときを正例とする．データの添え字集合を正例と負例に対応して二分しておく．$N = N^+ \coprod N^-$, $N^+ = \{n \in N | y_n = +1\}$ $N^- = \{n \in N | y_n = -1\}$. また，$N^+$ の濃度なども N^+ で表記する．

ロジスティック回帰　ロジスティック回帰とは，n 番目のデータ点が正例である確率を

$$p(+1|x_n, \theta) = \sigma(x_n \cdot \theta),$$

として計算するモデルである．σ はシグモイド関数 $\sigma(z) = \frac{1}{1+\exp(-z)}$ で，θ がモデルのパラメタである．学習時には対数尤度が最大になるように学習し，判別時にはこの確率がある閾値 τ 以上の時に x_n が正例であると予測する．通常は $\tau = 0.5$ である．

$$y_{pred} = \begin{cases} +1 & p(+1|x_n, \theta) \quad \geq \tau \\ -1 & p(-1|x_n, \theta) \quad \leq \tau. \end{cases}$$

TP, FP, FN の期待値　TP の個数は

$$\mathrm{TP} = \sum_{n\in N} 1_{N^+(n)} 1_{\{n|y_{pred}=+1\}} = \sum_{n\in N^+} 1_{\{n|y_{pred}=+1\}}$$

と定義される．1 は指示関数である．分類器としてロジスティック回帰を用いる場合には，正例と予測されたラベル $1_{\{n|y_{pred}=+1\}}$ の予測確率は $p(+1|x_n, \theta) = \sigma(x_n \cdot \theta)$ であるため，TP の期待値は以下の通りとなる．

$$\mathbb{E}(\mathrm{TP}) = \sum_{n\in N^+} p(+1|x_n, \theta) = \sum_{n\in N^+} \sigma(x_n \cdot \theta).$$

FP, FN の期待値も同様である．

$$\begin{aligned} \mathbb{E}(\mathrm{FP}) &= \sum_{n\in N^-} p(+1|x_n, \theta) = \sum_{n\in N^-} \sigma(x_n \cdot \theta), \\ \mathbb{E}(\mathrm{FN}) &= \sum_{n\in N^+} p(-1|x_n, \theta) = \sum_{n\in N^+} (1 - \sigma(x_n \cdot \theta)). \end{aligned}$$

これらは，シグモイド関数の和の形であるため，パラメタ θ に対して微分可能な関数となる．

期待 F 値　$\mathbb{E}$(TP), $\mathbb{E}$(FP), $\mathbb{E}$(FN) を用いることで，F 値の定義より，期待 F 値も次のように表すことができる．これもパラメタ θ で微分可能な関数となる．

$$\mathrm{F}_\alpha = \frac{\mathbb{E}(\mathrm{TP})}{\mathbb{E}(\mathrm{TP}) + \alpha\mathbb{E}(\mathrm{FN}) + (1-\alpha)\mathbb{E}(\mathrm{FP})}.$$

2.2 期待 Precision , 期待 Recall

期待再現率 ($\mathbb{R}$) と期待適合率 ($\mathbb{P}$) を同様に定義する．

$$\begin{aligned} \mathbb{P} &= \frac{\mathbb{E}(\mathrm{TP})}{\mathbb{E}(\mathrm{TP}) + \mathbb{E}(\mathrm{FP})} \\ &= \frac{\sum_{n\in N^+} \sigma(x_n \cdot \theta)}{\sum_{n\in N^+} \sigma(x_n \cdot \theta) + \sum_{n\in N^-} \sigma(x_n \cdot \theta)} \\ &= \frac{\sum_{n\in N^+} \sigma(x_n \cdot \theta)}{\sum_{n\in N} \sigma(x_n \cdot \theta)}, \\ \mathbb{R} &= \frac{\mathbb{E}(\mathrm{TP})}{\mathbb{E}(\mathrm{TP}) + \mathbb{E}(\mathrm{FN})} \\ &= \frac{\sum_{n\in N^+} \sigma(x_n \cdot \theta)}{\sum_{n\in N^+} \sigma(x_n \cdot \theta) + \sum_{n\in N^-} (1 - \sigma(x_n \cdot \theta))} \\ &= \frac{\sum_{n\in N^+} \sigma(x_n \cdot \theta)}{N^+}. \end{aligned}$$

上記を用いることで，提案手法の定式化も θ で微分可能な関数で表現できる．

$$\begin{cases} \text{maximize} & \frac{\sum_{n\in N^+} \sigma(x_n \cdot \theta)}{\sum_{n\in N} \sigma(x_n \cdot \theta)} \\ \text{subject to} & \sum_{n\in N^+} \sigma(x_n \cdot \theta) \geq \tau N^+. \end{cases}$$

パラメタの学習には，制約付き非線形最適化問題を扱う方法を用いることができる．内点法や逐次二次計画法 (SQP) などが有名であるが，次節で述べる実験では内点法を用いた．提案手法では，狙いとする再現率のレベルにあわせて学習できる．例えば，90% 以上の再現率を狙う場合は，ハイパーパラメタ τ を 0.9 と設定すれば良い．ただし，τ もクロスバリデーションなどでチューニング可能である．

3 実験

3.1 実験設定

本節では，提案手法の有効性評価として，一般的なロジスティック回帰と期待 F 値最大化 [6] と比較し，実際に高再現率における適合率が向上するかを確認する．

実験データには，ロイターコーパス vol.1 (RCV1) と Twitter から作成した 2 種類のインハウスデータを使用した．RCV1 からは 2 万記事を抽出し，1 万件を学習データ，残りの 1 万件を評価データとし，全 51 分野のうち，正例の割合が 10% 未満のもの中から無作為に選んだ 10 ラベル（E512, C18, GSPO, G15, GDIP, C41, C33, M141, GJOB, C21）を使用した．Twitter データは「災害」「事故」の 2 種類あり，事故は，車両と人，物の接触があり，それによって被害が生じたことを述べた tweet を判別するタスクで，全データ 14969 件中，正例は 4037 件で約 27.0% となっている．災害は，現

表 2: RCV1 の 10 ラベルの適合率の平均値

再現率	80%	85%	90%	95%	100%
LR	0.539	0.464	0.386	0.294	0.0497
$F_{0.5}$	0.513	0.421	0.356	0.306	0.0485
F_TUNED	**0.554**	0.453	0.377	0.317	**0.0509**
提案手法	0.528	0.466	**0.402**	0.321	0.045
提案手法_TUNED	0.534	**0.486**	0.400	**0.328**	0.0508

在および過去の天災（豪雨，地震，雷など）の発生に言及した tweet を判別するタスク，全データ 17714 件中，正例は 1076 件で約 6.1% となっている．いずれも，半分を学習データ，残りの半分を評価データとした．

求める高再現率のレベルとして，再現率 80%, 85%, 90%, 95%, 100% を設定した．

期待 F 値最大化 [6] については，適合率と再現率の調和平均に置いて，どちらを重視するかを調整するためのハイパーパラメタ α があるため，標準的 $\alpha = 0.5$ と，ハイパーパラメタチューニング用のデータを用いて，0.0 から 1.0 の 0.1 刻みで調整した場合の 2 通りの実験を行った．これは，重み付き F 値の重みの調整が，高再現率で高精度な分類に適するかどうかの検証の意味もある．

提案手法も同様に，ハイパーパラメタ τ を狙いとする再現率に固定した場合と，チューニングした場合の適合率を算出した．ハイパーパラメタのチューニング範囲は，評価する際の再現率にあわせて $\tau = 0.8, 0.85, 0.90, 0.95, 0.99$ とした[1]．

ロジスティック回帰，期待 F 値最大化，提案手法のいずれのにもパラメータの L2 ノルム $|\theta|^2$ を正則化項として追加した．その際の係数について，ロジスティック回帰では $a \cdot 10^m$ とし，$F_{0.5}$ および 期待 F 値最大化ロジスティック回帰では，素性数 $\cdot\, a \cdot 10^m$ として，a は 0 から 9 まで 1 刻み，m は −3 から 3 まで 1 刻みで動かして最良の値を選択できるようにした，

モデルのパイパーパラメタ，正則化の係数の調整は，学習データを半分に分け，一方で学習を行い他方で評価を行い最良の結果を出したものを採用した．

3.2 実験結果

まず，RCV 1 コーパスの結果について，各再現率 (80%, 85%, 90%, 95%, 100%) における全てのラベルにわたる適合率の平均値を (表 2) に示す．提案手法は，通常のロジスティック回帰 (LR) および期待 F 値最大化 ($F_{0.5}$) と比較し，概ね適合率が上回っている．さらにパラメータ調整を施した期待 F 値最大化（F_TUNED）と比較しても，再現率 85%, 90%, 95%では適合率が上回っている．

[1]再現率 100% の場合には，$\tau = 1.0$ を制約とすると $\mathbb{E}(\mathrm{FP}) = 0$ となる，これは数理的に絶対満たされないため，$\tau = 0.99$ とした．

表 3: Twitter の「事故」の各再現率における適合率

再現率	80%	85%	90%	95%	100%
LR	**0.613**	**0.571**	0.512	0.439	0.29
$F_{0.5}$	0.596	0.555	0.494	0.426	0.298
F_TUNED	0.583	0.563	0.516	**0.455**	**0.306**
提案手法	0.57	0.57	**0.521**	0.449	0.287
提案手法_TUNED	0.604	0.56	0.516	0.449	0.292

表 4: Twitter の「災害」の各再現率における適合率

再現率	80%	85%	90%	95%	100%
LR	0.252	0.19	0.145	0.099	0.044
$F_{0.5}$	0.154	0.137	0.113	0.09	**0.049**
F_TUNED	0.234	0.215	0.208	0.134	0.044
提案手法	0.216	**0.217**	0.212	**0.176**	0.047
提案手法_TUNED	**0.256**	**0.217**	**0.224**	**0.176**	0.045

次に，Twitter の「事故」「災害」の結果をそれぞれ (表 3) (表 4) に示す.「事故」の結果を見ると，各手法ごとの適合率の傾向の差は見られなかった.「災害」では RCV1 の平均と同様に，再現率 80%のときの LR , 100%のときの $F_{0.5}$ を除き，提案手法の適合率が上回っている．F_TUNED との比較でも再現率 80%の場合を除き，やはり提案手法の適合率が上回った．

総括すると，提案手法は，所望の再現率において高い適合率を達する傾向にあり，特に，再現率 85%, 90%，95%では安定していた．一方，80%や 100%など，比較的低め，もしくは，極端に高い場合には，効果が見られなかった．

また，F_TUNED により，高再現率における適合率を高める効果が見られることもわかった．実際，選ばれたハイパーパラメタ α を見ると，0.8 や 0.9 など再現率重視の場合が多かった．すなわち，F1-measure というバランスの良い分類指標のの最大化は，高再現率における適合率の向上には必ずしも適していないものの，再現率へ重みを付けることで，類似の効果が得られていた．

4 関連研究

提案手法のように，評価指標を直接的に最大化する先行研究は複数存在する．評価指標としては，F 値 [4, 6]，AUC [5, 11, 3], partial AUC [8, 9], 平均適合率 [2, 10, 12], break-even point[7] などが取り組まれている．これらの評価指標の最大化は，本研究の目的とす

る高い再現率での適合率の向上には直接的ではないと考えられる．例えば，AUC, break-even point, 平均適合率は，高い再現率という条件を直接的に表現することはできない．F 値は，適合率と再現率の調和平均の重みによって，再現率を重視することができるが，相対的に適合率は犠牲となり，ある再現率での適合率を最大化するという目的には直接的でない．partial AUC は，AUC の特定の区間の面積を最大化するが，これはランキングした際の FP の割合の区間における TP の数を最大化すること，すなわち，適合率に関係する条件のもとで，再現率の最大化を行っているとみなせる．高い再現率における適合率の向上という目的には直接的でないが関連は深い．

5 おわりに

本稿では，分類漏れを極力防ぎつつも高い適合率を達成するために，再現率そ制約として，適合率を最大化する学習方法について述べた．提案手法は，ロジスティック関数を用いて，適合率，再現率を期待値で表現し，再現率が一定以上という制約で適合率を最大化する最適化問題として定式化した．実験では，ロジスティック回帰，期待 F 値最大化と比較し，複数のテキスト分類タスクにて，高い再現率の適合率を高める効果があることを確認した．

参考文献

[1] Charu C. Aggarwal and ChengXiang Zhai. Chapter. 6: A survey of text clustering algorithms. In *Mining Text Data*, pp. 77–128. Springer, 2012.

[2] Chris Buckley and Ellen M. Voorhees. *Evaluating Evaluation Measure Stability*. ACM Press, 33–40, 2000.

[3] Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. In *Advances in neural information processing system*, pp. 313–320, 2004.

[4] Krzysztof. Dembczynski, Willem. Waegeman, Weiwei. Cheng, and Eyke. Hullermeier. An exact algorithm for f-measure maximization. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, 2014.

[5] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiveroperating characteristic (roc) curve. In *Radiology*, 1982.

[6] Martin Jansche. Maximum expected f-measure training of logistic regression models. In *Proceeding of HLT '05 Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP)*, pp. 692–699, 2005.

[7] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the International Conference on Machine Learning*, 2005.

[8] Harikrishna Narasimhan and Shivani Agarwal. A structural SVM based approach for optimizing partial AUC. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 516–524, 2013.

[9] Harikrishna Narasimhan and Shivani Agarwal. Svm_pAUC̃ight: A new support vector method for optimizing partial auc based on a tight convex upper bound. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 692–699, 2013.

[10] Warren Greiff William Morgan and John Hende. Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. In *Human Language Technology conference /North American Chapter of the Association for Computational Linguistics(HTL-NAACL 2004)*, 2004.

[11] Lian Yan, Robert Dodier, Michael C. Mozer, and Richard Wolniewicz. Optimizing classifier performance via the wilcoxon-mann-whitney statistic. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pp. 848–855, 2003.

[12] Yisong. Yue, Thomas. Finley, Filip Radlinski, and Thorsten. Joachims. A support vector method for optimizing average precision. In *Special Internet Group on Information Retrival (SIGIR)*, 2007.