1A1-1

ラベル信頼度を用いたブースティング手法のインバランスデータへの応用

a Boosting Method Utilizing Label Reliability in Imbalanced Datasets

中田 康太[*1] 村上 知子[*1]
Kouta Nakata Tomoko Murakami

[*1]株式会社東芝研究開発センター
Corporate Research & Development Center, TOSHIBA

In supervised learning, accurate learning requires accurate labels, and accurate labels requires accurate knowledge and rich experiences. Domain experts are expected to label data. However, having such experts for all data is often impossible because of its high cost and sometimes we have to make use of 'cheaper' labels of amateurs. In such a case, experts' and amateurs' labels are not discriminated even when amateurs' wrong labels may make the resultant classifier poor. We proposed Credit AdaBoost (CAB) that utilizes experts' labels and suggested that CAB achieves better classifiers than AdaBoost on various UCI datasets. But in reality, CAB faces a class imbalance problem; CAB might not perform well on imbalanced datasets where the number of big classes overwhelms that of small ones in experts' data. In this paper, we propose a boosting method coping with the class bias by extending CAB. An empirical study on internet video recommendation suggests effectiveness of our method on real world data.

1. はじめに

教師あり学習におけるラベル信頼度問題について注目する。ラベル信頼度問題とは、エキスパートによりラベル付けされた「質の良い（獲得コストの高い）少量の教師データ」と非エキスパートによりラベル付けされた「ノイズを含む（獲得コストの低い）多量の教師データ」が混在する場合、両者を同一の学習データとして用いたとき、後者に含まれるノイズによって精度の悪い分類器が生成する状況を表している。我々はラベル信頼度問題に対して Boosting 手法を用いた Credit AdaBoost (CAB) 手法を提案した [Nakata 08]。CAB はラベル信頼度問題において質の異なる教師データを別々に扱うことで高い精度の分類モデルを構築する手法で、1) 質の高い教師データを基にして質の悪い教師データのラベルに信頼度を付与し、2) ラベルの信頼度を反映した学習を行い分類モデルを構築することを主な特徴としている。

CAB は質の高い教師データにおいて正例と負例の比率が大きく異なるインバランスなデータセットに対しては精度が向上しないことが予想される。現実的な問題においてはデータはインバランスであることが多いことから、本論文ではインバランスデータに対応した CAB 手法を提案する。ラベル信頼度問題が起こる具体的な状況としてネット動画コンテンツの推薦に注目し、実験により提案手法の性能を検証する。本論文の構成を以下に示す。2 章では CAB のラベル信頼度の重要性を検証し、インバランスデータに対応した CAB 手法を提案する。3 章では、ネット動画コンテンツの推薦について精度実験を行い、提案手法の有用性を検証する。4 章で本論文をまとめる。

2. 提案手法

2.1 Credit AdaBoost (CAB)

我々は AdaBoost([Freund 96]) に基づき、ラベル信頼度問題に対してラベル信頼度付きブースティング手法 Credit AdaBoost (CAB) を提案した [Nakata 08]。ラベル信頼度問題では、前提としてエキスパートによりラベル付けされた「質の良い少量の教師データ」と非エキスパートによりラベル付けされた「ノイズを含み多量の教師データ」が混在する状況を考える。本論文では前者をエキスパートデータ、後者を非エキスパートデータと呼ぶ。ラベル信頼度問題に対処するために、CAB は全ての学習データ (x_i, y_i) に対してラベル信頼度 c_i を付与し、ラベル信頼度付き学習データ (x_i, y_i, c_i) を用いて学習を行う。CAB においては c_i は例えば式 (1) で与えられる。

$$c_i = \begin{cases} 1 & \text{if } x_i \in \mathcal{D}_{\text{ex}} \\ \sum_{j=1, y_i = Y_{ij}}^{k} \frac{1}{d(x_i, X_{ij})} / \sum_{j=1}^{k} \frac{1}{d(x_i, X_{ij})} & \text{if } x_i \in \mathcal{D}_{\text{ne}} \end{cases} \quad (1)$$

$\mathcal{D}_{\text{ex}}$ はエキスパートデータの集合、$\mathcal{D}_{\text{ne}}$ は非エキスパートデータの集合、(X_{ij}, Y_{ij}) は学習データ (x_i, y_i) に j 番目に近いエキスパートデータ、$d(x_i, X_{ij})$ は (x_i, y_i) と (X_{ij}, Y_{ij}) の距離を表している。$d(x_i, X_{ij})$ にはユークリッド距離を用いる。

エキスパートデータについてはラベル付けが信頼できると考えられるため、全ての $x_i \in \mathcal{D}_{\text{ex}}$ で $c_i = 1$ とする。非エキスパートデータについてはラベル付けが必ずしも信頼できるとは限らないため、近傍のエキスパートデータを用いてラベル信頼度を算出する。式 (1) では、$x_i \in \mathcal{D}_{\text{ex}}$ の近傍 k 個のエキスパートデータを参照し、ラベルが y_i と等しいエキスパートデータから距離に反比例したラベル信頼度をもらう。非エキスパートデータ (x_i, y_i) のラベル信頼度は近傍 k 個のエキスパートデータが全て y_i と等しいときに最大値 1 となる。

表 1 に CAB のアルゴリズムを示す。表 1 では、従来の AdaBoost と比較して過程 1 と過程 3(c) が拡張されている。過程 1 においては、式 (1) により各学習データのラベル信頼度を算出する。過程 3(c) においては、データ重み $D_t(i)$ を更新する際にラベル信頼度 c_i をかけることでラベル信頼度を学習に反映する。過程 3(c) のラベル信頼度の反映方法により、ラベル信頼度が低いデータは仮に誤分類された場合でもデータ重みが大きくならず、ノイズの過学習を防ぐことが期待できる。[Nakata 08] では、公開データを用いた実験において CAB による分類精度向上が確認されている。

連絡先: 中田康太，kouta.nakata@toshiba.co.jp

Algorithm: Credit AdaBoost

1. ラベル信頼度 $c_i, i = 1, ..., N$ を計算する
2. 初期データ重みを与える $D_1(i) = 1/N, i = 1, ..., N$.
3. $t = 1, ..., T$ について以下を (a)-(c) を繰り返す:
   (a) 弱仮説 $h_t(x)$ をデータ重み $D_t(i)$ を用いて学習する
   (b) 弱仮説重み $\alpha_t = \frac{1}{2} \log \frac{\epsilon_t}{1-\epsilon_t}$ を計算する。
   ϵ_t はデータ重み付きエラー $\epsilon_t = \sum_{y_i \neq h_t(x_i)} D_t(i)$
   (c) データ重みを更新する:
   $D_{t+1}(i) = D_t(i) c_i \exp[-y_i \alpha_t h_t(x_i)]$, $i = 1, ..., N$
   $D_{t+1}(i) \leftarrow D_{t+1}(i) / \Sigma_{i=0}^{N} D_{t+1}(i)$.
4. 最終仮説 $H(x)$ を出力: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

表 1: Credit AdaBoost

2.2 ラベル信頼度の重要性について

AdaBoost については過去に非常に多くの研究がなされており、非常に高い性能を示すことが知られている。Schapire et al. は AdaBoost の性能を Margin と呼ばれる量を導入して説明している [Schapire 97]。Margin は各学習データ (x_i, y_i) に対して定義され、T 回目まで弱仮説 h_t が生成されたときの Margin は式 (2) で表される。

$$\mathrm{margin}(x_i, y_i) = \frac{y_i \sum_{t=0}^{T} \alpha_t h_t(x_i)}{\sum_{t=0}^{T} \alpha_t} \tag{2}$$

ここで h_t、α_t は表 1 と同様に弱仮説と弱仮説重みを表す。式 (2) は、T 個の弱仮説が学習データ (x_i, y_i) に対してどの程度正しい分類を行なったかを表しており、最終仮説による境界面とデータとの距離と解釈することができる。AdaBoost は、データ重みを大きくすることで境界面の近くの学習データについても Margin を大きくするように仮説を生成し、結果として高い汎用性を実現する Large Margin Classifier であることが知られている [Schapire 97]。

ここでは Schapire et al. の導出に従って、Margin の観点から CAB の性質を考察する。今、CAB において T 個の弱仮説が生成している状況を考えると、手順 2(b) の重みの更新の式から以下の関係が得られる。

$$\begin{aligned} D'_{t+1}(i) &= \frac{c_i D'_t(i) \exp[-y_i \alpha_t h_t(x_i)]}{Z_t} \\ &= \cdots \\ &= \frac{c_i^T \exp[-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)]}{\Pi_{t=1}^{T} Z_t} \end{aligned}$$

ここで Z_t は規格化因子であり、式 (3) で表される。

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)} \tag{3}$$

$\exp[-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)]$ について解いて、両辺の $-\ln$ をとることで、以下の関係を得る。

$$y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) = \sum_{t=1}^{T} \ln \frac{c_i}{Z_t} - \ln D'_{t+1}(i)$$

この結果を学習データ (x_i, y_i) の Margin の式 (2) に代入することで、式 (4) を得る。

$$\mathrm{margin}(x_i, y_i) \geq \frac{\ln c_i - \frac{1}{T} \sum_{t=1}^{T} \ln \sqrt{4\epsilon_t(1-\epsilon_t)}}{\frac{1}{T} \sum_t^T \sqrt{\ln \frac{1-\epsilon_t}{\epsilon_t}}} \tag{4}$$

式 (4) は、ラベル信頼度が c_i である学習データ (x_i, y_i) の Margin の下限が $\ln c_i$ によることを示している。この下限から、高い信頼度を与えられた学習データは CAB の学習過程で Margin は大きくなるが、低い信頼度を与えられた学習データは Margin が小さいまま学習に反映されないことが分かる。

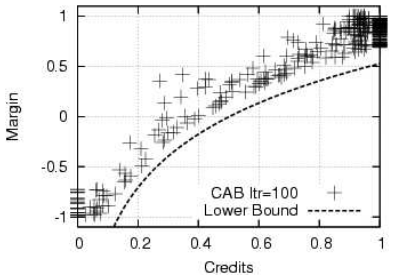

図 1: CAB のラベル信頼度と Margin の関係の例

図 1 は、公開データ breast_cancer [Asuncion 07] を用いた実験において観測されるラベル信頼度と学習 Margin の分布を表している。ここで縦軸は Margin の値、横軸はラベル信頼度、各点は学習データを表している。図 1 では、信頼度と最小 Margin の関係が明確に得られている。これらの関係は式 (4) で見積もられる下限を満たしており、$D'_t(i) = c_i D_t(i)$ の方法で信頼度を反映した場合には、$\log c_i$ に依った形で各データの Margin の下限が決定されることが明確になっている。そのためラベル信頼度 c_i の値が正確に与えられれば、CAB により有効な学習が行えることが分かる。

2.3 インバランスデータへの対応

CAB はラベル信頼度を式 (1) により付与している。このようなラベル信頼度付けは、エキスパートデータの正例と負例の比率が大きく異なるインバランスの場合には必ずしも適切でないと考えられる。本論文ではエキスパートデータがインバランスである状況において、ラベル信頼度付与にクラス重みと適応近傍数を導入することで精度の向上を図る。

2.3.1 クラス重み

エキスパートデータが負例に偏っている場合、非エキスパートデータの近傍には負例が存在する確率が高く、特に非エキスパートデータの正例に対して適切なラベル信頼度が付与できない可能性が考えられる。そのためラベル信頼度の算出の際に、各クラスの事例数の比率に応じたクラス重みを導入することで、より適切なラベル信頼度を付与する。式 (5) はクラス重み付きラベル信頼度の例である。

$$c_i = \sum_{j=1, y_i = Y_{ij}}^{k} \frac{W_{Y_{ij}}}{d(x_i, X_{ij})} \Big/ \sum_{j=1}^{k} \frac{W_{Y_j}}{d(x_i, X_{ij})} \quad \text{if } x_i \in \mathcal{D}_{\mathrm{ne}} \tag{5}$$

ここで $W_{Y_{ij}}$ はクラス Y_{ij} のクラス重みである。正例と負例のクラス重みを W_+, W_- とすると、W_+, W_- は例えば以下のように表すことが可能である。

$$\begin{aligned} W_+ &= |C_+^{\mathrm{ex}}| / (|C_+^{\mathrm{ex}}| + |C_-^{\mathrm{ex}}|) \\ W_- &= |C_-^{\mathrm{ex}}| / (|C_+^{\mathrm{ex}}| + |C_-^{\mathrm{ex}}|) \end{aligned} \tag{6}$$

$|C_{+,-}^{\mathrm{ex}}|$ はエキスパートデータ中の正例、負例の数を表している。事例の少ないクラスには高い重みが与えられるため、非エキスパートデータの正例の近傍にエキスパートデータ正例が 1

件でも含まれている場合にはラベル信頼度は大きくなり、事例の比率の違いを軽減できると考えられる。

2.3.2 適応近傍数

非エキスパートデータのラベル信頼度は近傍のエキスパートデータを参照して決定するため k 近傍法 (kNN) と関連が深い。kNN による分類においては、正例と負例の比率が大きく異なる状況において、クラス毎に k を変化させることで精度を向上する手法が提案されている (e.g. [Baoli 04])。CAB においても、エキスパートデータの正例と負例の比率が大きく異なる場合、非エキスパートデータが参照する適切な k の値は正例と負例で同一であるとは限らない。エキスパートデータがインバランスである状況では、正例と負例で非エキスパートデータが参照するエキスパートデータの数を変えることで適切なラベル信頼度の算出が期待される。

本論文では非エキスパートデータが参照するエキスパートデータ数を可変にしたラベル信頼度を式 (7) により導入する。

$$c_i = \sum_{j=1,y_i=Y_{ij}}^{K_{y_j}} \frac{1}{d(x_i,X_{ij})} / \sum_{j=1}^{K_{y_j}} \frac{1}{d(x_i,X_{ij})} \quad \text{if } x_i \in \mathcal{D}_{\text{ne}} \quad (7)$$

ここで K_{y_j} は非エキスパートデータ (x_j, y_j) の参照するエキスパートデータの数を表している。(x_j, y_j) が正例の場合に参照するエキスパートデータの数を K_+、負例の場合に参照するエキスパートデータの数を K_- とすると、K_+, K_- は式 8 のように表すことができる。

$$\begin{aligned} K_+ &= a_0 + k_0 \frac{|C_+^{\text{ex}}|}{|C_-^{\text{ex}}|} \\ K_- &= a_0 + k_0 \end{aligned} \quad (8)$$

$|C_{+,-}^{\text{ex}}|$ はエキスパートデータ中の正例、負例の数を表しており、a_0, k_0 はあらかじめ設定された正の整数である。エキスパートデータが負例に偏っている場合には $|C_+^{\text{ex}}|/|C_-^{\text{ex}}| \sim 0$ となるため、例えば $a_0 = 3$、$k_0 = 7$ とすることで、正例の非エキスパートデータは $K_+ = 3$ の近傍エキスパートデータを参照し、負例の非エキスパートデータは $K_- = 10$ の近傍エキスパートデータを参照することになる。式 (6) と式 (8) から、本論文で用いるラベル信頼度は式 (9) で表される。

$$c_i = \sum_{j=1,y_i=Y_{ij}}^{K_{y_j}} \frac{W_{Y_{ij}}}{d(x_i,X_{ij})} / \sum_{j=1}^{K_{y_j}} \frac{W_{Y_{ij}}}{d(x_i,X_{ij})} \quad \text{if } x_i \in \mathcal{D}_{\text{ne}} \quad (9)$$

式 (9) によるラベル信頼度を用いた CAB を adaptive-K based Weighted CAB (KWC) と呼ぶ。

3. ネット動画コンテンツ推薦への応用

3.1 ネット動画コンテンツ推薦とラベル信頼度問題

Youtube(http://jp.youtube.com/) に代表されるネット動画配信サイトでは、膨大な数の動画が配信されている。このようなサイトで配信されている動画は、ユーザが自ら制作・投稿を行っていることから User Generated Contents (UGC) と呼ばれる。膨大な数の UGC の中からユーザが所望するコンテンツを視聴することは困難であり、コンテンツを有効に推薦する方法が必要であると考えられる。

一般的な UGC 配信サイトでは検索サービスを提供しており、ユーザは自分の興味のある出演者や話題の動画を積極的に探すことが多い。またユーザは嗜好に合う動画をお気に入りとして何度も視聴する傾向にある。そのため視聴履歴をユーザの好みを表す教師データとしてユーザの嗜好を学習することで、有効な UGC 推薦が期待できる。一方で、膨大な数の UGC の中でユーザは限られた量のコンテンツしか視聴できないため、ユーザ個人が視聴する UGC だけでは嗜好を学習するための十分な視聴履歴が得られない可能性がある。

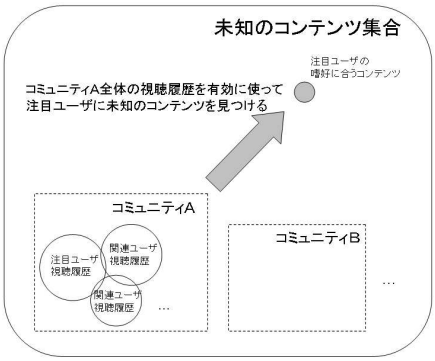

図 2: コミュニティ集合の例

このような不足を補う方法として、興味の類似した他のユーザの視聴履歴を用いる方法が考えられる。インターネット上では多くのユーザが UGC を視聴しており、興味の共通するユーザも多く存在する。興味の共通するユーザの視聴履歴を補足的な教師データとして用いることで、教師データの不足を補うことが期待される。しかし、興味の共通する他のユーザの視聴履歴を教師データとして用いる場合、その教師信号が必ずしも正確とは限らない。例えば、一部において興味が共通しているユーザ同士であっても、他の部分においては異なる興味を持っていると考えられる。このとき一方の興味によって視聴されたコンテンツは、必ずしも他方にとって興味のあるコンテンツであるとは限らない。そのため注目するユーザの視聴履歴と興味の共通する他のユーザの視聴履歴を組合わせて嗜好モデルを学習した場合、他のユーザの視聴履歴が大量のノイズとなり、推薦精度が低下すると予想される。

ここで注目するユーザの視聴履歴をエキスパートデータ、興味の共通する他のユーザの視聴履歴を非エキスパートデータとみなすと、両者を用いて嗜好モデルを学習することはラベル信頼度問題に帰着することができる。よって本論文では UGC 推薦に CAB/KWC 手法を応用し、注目したユーザの視聴履歴と興味の共通する他のユーザの視聴履歴を有効に組合わせることで精度の高い推薦を行う。ここでユーザの視聴履歴は全データと比較して非常に少数であるため、特にインバランス性を考慮した KWC の効果が期待される。今後は簡単のため、推薦を行うユーザを注目ユーザ、注目ユーザと興味の類似したユーザを関連ユーザと呼ぶこととする。

3.2 コミュニティの作成

本論文では、ある話題 K についての動画を一定以上持っているユーザを話題 K に興味のあるユーザとみなし、その集合を話題 K のコミュニティと呼ぶ。同一のコミュニティに所属しているユーザは興味が共通しているため、お互いを関連ユーザとみなすことが可能である。本実験ではあるコミュニティに属する注目ユーザに対して、コミュニティ内の関連ユーザの視聴履歴を利用した推薦を行う。

図 2 は、本論文で注目する UGC 推薦の状況を模式的に表した例である。外枠は全コンテンツ集合を表している。注目

ユーザが視聴できるUGCは限られているため、全体のコンテンツ集合のごく一部となっている。注目ユーザの視聴履歴のみを用いて嗜好モデルを学習した際には、学習データが不足し精度の高い推薦が行えない可能性が考えられる。学習データの不足に対し、関連ユーザからなるコミュニティA 全体の視聴履歴を用いることで、より多くの視聴履歴を学習に利用することが可能である。しかしコミュニティ全体の視聴履歴を全て用いた場合、ノイズにより推薦の精度が低下すると考えられる。本論文では注目ユーザの視聴履歴をエキスパートデータ、コミュニティ全員の視聴履歴を非エキスパートデータとして扱い、CAB/KWC を応用することで推薦精度の向上を図る。

3.3 実験

本論文では、日本オペレーションズ・リサーチ学会の実践的データマイニング研究部会が主催したリコメンデーションコンテスト 2009 ([RC2009 09]) において提供されたデータを用いる。提供データはチームラボ社のサグールテレビ (http://sagool.tv/) で貯められた一般ユーザの動画の視聴履歴データとお気に入りデータおよび全動画のメタ情報データで構成されている。視聴履歴データとお気に入りデータは 448 名の一般ユーザのデータであり、視聴履歴は 3ヵ月、お気に入りは 1 年 3ヵ月の期間記録されている。メタ情報データにはタイトル、投稿者などのメタ情報が約 180 万件含まれている。

本論文では、RC2009 のデータを用いて擬似的にコミュニティを作成し実験を行う。擬似的なコミュニティは以下の手順で生成する。448 名のユーザのうち、お気に入り N_f 件に特定の出演者名や単語 K を含むコンテンツが N_c 件以上存在するユーザを抽出する。抽出されたユーザのうち、K を含むコンテンツの割合 N_c/N_f の値が P_c 以上であるユーザをコミュニティ K に属するユーザとする。本実験では $N_c = 10$、$P_c = 0.15$ とする。この抽出基準により 5 人以上のメンバーを持つコミュニティを 5 件作成した。表 2 に作成したコミュニティの特徴を示す。人数はコミュニティに含まれるユーザ数、属性はコミュニティの話題の属性を示している。5 件のコミュニティのうち 4 件が出演者、1 件が単語に関するコミュニティである。

ここで以下の手順でデータセットを作成し評価を行う。まずコミュニティから推薦対象となる注目ユーザを選択し、注目ユーザのお気に入りデータを正例として一時学習データを作成する。注目ユーザが所属していないコミュニティのユーザのお気に入りを負例として一時学習データに追加する。どのコミュニティにも属さないコンテンツを全て評価データとする。一時学習データをランダムに 2 分割し片方を評価データに含め、他方を学習データ (エキスパートデータ) とする。同じコミュニティ内の他ユーザの学習データを注目ユーザの非エキスパートデータとする。一時学習データを分割する際には同じ件数の正例を学習データと評価データに配分し、評価データ中の正例を正解として推薦精度を評価する。以上の分割を 10 回繰り返し、平均の精度を最終的な推薦精度とする。学習と評価にはタイトルから抽出したキーワード、出演者、ジャンル、動画投稿者を特徴量としてベクトル化したデータを用いる。

本実験では、ラベル信頼度を用いた CAB、KWC を従来の AdaBoost 手法 (Traditional AdaBoost, TAB) と比較する。注目ユーザのデータのみを用いて TAB により推薦を行った場合を TAB1、注目ユーザのデータに加えコミュニティ全体のデータを全て用いて TAB により推薦を行った場合を TAB2 で表す。これらの Boosting 手法は分類を行う際には最終仮説 $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$ の符号によりクラスを分類しているが、今回の推薦タスクでは $H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$ の出力の実数値をそのまま用い、出力値の高い順に推薦リストを作成する。各手法は推薦リストの上位 10 件の適合率で評価する。

ID	人数	属性	TAB1	TAB2	CAB	KWC
1	8	出演	7.3	5.0	12.0	**13.8**
2	12	出演	10.8	2.5	13.5	**14.2**
3	10	出演	24.4	7.0	24.6	**26.7**
4	5	出演	27.0	25.0	**31.5**	**31.5**
5	5	単語	15.9	3.8	17.1	**17.2**

表 2: 各コミュニティの特徴と推薦精度 (適合率)

表 2 に各手法の推薦精度を示す。数値はコミュニティ内の対象ユーザの平均の適合率を表している TAB2 では TAB1 と比較して全てのコミュニティで精度が低下しており、関連ユーザのお気に入りがノイズとなり精度の低下を引き起こしていると推測することができる。CAB/KWC の精度は全てのコミュニティで TAB を上回る結果となっており、ラベル信頼度の導入によりコミュニティのお気に入りデータを有効に利用し推薦を行っていることを示唆している。また KWC の推薦精度は CAB を上回っており、本論文で導入したインバランスデータに対する手法が有効であることを示している。

4. まとめ

本論文では、インバランスデータに対応したブースティング手法を提案し、その有用性について検証を行った。学習データがインバランスであることは現実の問題において頻繁に起こるため、本手法によりラベル信頼度付き手法の応用可能性が示されたと言える。今後は、更なる応用に向けて検討を続ける。

謝辞: コンテストを通してデータを提供していただいた日本オペレーションズ・リサーチ学会 実践的データマイニング研究部会の方々に感謝いたします。

参考文献

[Asuncion 07] Asuncion, A. and Newman, D. J.: UCI Machine Learning Repository, http://www.ics.uci.edu/∼mlearn/MLRepository.html

[Baoli 04] Baoli, L. and SHIWEN, Y.: An adaptive k-nearest neighbor text categorization strategy, ACM Transactions on Asian Language Information Processing, vol.3, 215-226 (2004)

[Freund 96] Freund, Y. and Schapire, R. E.: Experiments with a New Boosting Algorithm, In Proc. 13th International Conf. on Machine Learning, pp. 148-156 (1996)

[Nakata 08] Nakata, K., Sakurai, S. and Orihara, R.: Classification Method Utilizing Reliably Labeled Data, Lecture Notes in Computer Science, vol.5177, pp. 114-122 (2008)

[RC2009 09] http://kgmod.jp/contest

[Schapire 97] Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. E.: Boosting the margin: a new explanation for the effectiveness of voting methods, In Proc. 14th International Conf. on Machine Learning, pp. 322-330 (1997)