F-041

# SVM 学習のためのデータサンプリング法の提案
# A Proposal of Data Sampling Method for SVM learning

高橋　洸 † 松本 一則 ‡ 橋本 和夫 † 徳山 豪 †
Takeru Takahashi Kazunori Matsumoto Kazuo Hashimoto Takeshi Tokuyama

## 1. はじめに

サポートベクターマシン (Support Vector Machine, SVM)[2] は，2 クラスの分類問題を取扱う学習機械の 1 つであり，高い性能と扱いの容易さからパターン認識の分野で広く用いられている．一般に SVM では学習に用いるデータの数が多いほど分類精度は向上するが，大規模データに手作業でクラスを与える作業は膨大なコストを要し，また計算時間も爆発的に増加するといった問題点がある．従って，効率のよい SVM 学習を行うため，クラスを付与するデータをサンプリングし能動的に学習データを増加させながら学習を進める枠組みが必要とされている．SVM 学習のためのデータサンプリング法で最も一般的なものは Tong らが 2000 年に提案した SIMPLE 法と呼ばれる手法であり [7]，判別面の形成に大きく影響を与えるデータをサンプリングしてくため判別面を効率よく安定させることができるが，学習の初期での性能が悪いことが知られる．Baram らによる KFF 法は，学習済みのデータから最も離れたデータをサンプリングする手法であり，学習初期の性能に特化しているが，有効となるデータセットが限られている [1]．この 2 つの手法を組み合わせた手法として Baram らによる COMB 法 [1] や Osugi らによる EXPLORATION 法 [4] は，SIMPLE 法や KFF 法に比べ多くのデータセットで高い効果を示しているが，アルゴリズムが非常に複雑になってしまうといった問題点がある．本研究では，SIMPLE 法を大域的に拡張することにより，アルゴリズムを複雑にすることなく，効率良くサンプリングを行う手法を提案する．

## 2. 既存のデータサンプリング手法

### 2.1 SIMPLE 法

SIMPLE 法 [7] はカーネル空間中で判別面に近いラベルなしデータをサンプリングしていくという手法であり，SVM を用いた能動学習では最も良く用いられている．判別面に近いデータはラベルが不確実であり，同時に判別面の形成に大きく影響を与えるため，効率よく判別面を安定させていくことが出来る．カーネル空間中のデータ $\boldsymbol{x}$ から判別面までの距離 $d_{sim}(\boldsymbol{x})$ は式 (1) のように表されるため容易に計算可能である．なおこの距離はデータ $\boldsymbol{x}$ がソフトマージンの境界上にある場合 $|f(\boldsymbol{x})| = 1$ となるように正規化されている．

$$d_{sim}(\boldsymbol{x}) = |f(\boldsymbol{x})| = \left| \sum_{i=1}^{L} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right| \quad (1)$$

SIMPLE 法のサンプリングのアルゴリズムを Algorithm 1 に示す．LABEL() はデータにラベルを付与する関数であり，7 行目でサンプリングされたラベルなしデータにラベルを与え，8 行目でこのデータをラベルありデータ集合 $L$ に加えている．

**Algorithm 1** SIMPLE 法

**Require:** $L$:ラベルありデータ集合, $U$:ラベルなしデータ集合

1: **loop**
2: Compute $\boldsymbol{\alpha}$ and $b$ using SVM($L$).
3: **for all** $\boldsymbol{x} \in U$ **do**
4: $d_{sim}(\boldsymbol{x}) \leftarrow \left| \sum_{\boldsymbol{s} \in L} \alpha_i K(\boldsymbol{s}, \boldsymbol{x}) + b \right|$
5: **end for**
6: $\boldsymbol{x} \leftarrow \arg\min d_{sim}(\boldsymbol{x})$
7: $y \leftarrow \mathrm{LABEL}(\boldsymbol{x})$
8: $L \leftarrow L \cup (\boldsymbol{x}, y)$
9: $U \leftarrow U \setminus \boldsymbol{x}$
10: **end loop**

この手法は判別面が理想的な判別面にある程度近い形をしていれば効率よく判別面を詳細化することが可能であるが，データセット全体のうち判別面の周辺のみをを局所的にサンプリングしていくため，学習初期で判別面から離れた座標に重要なデータがあるような場合，これをサンプリングすることができない．また，本来の判別面に近いデータはラベルが付け難いデータである場合も考えられ，そうしたデータのみを学習させてしまうと過学習に繋がり効率が悪い．

### 2.2 KFF(Karnel Farthest First) 法

KFF 法 [1] は Baram らにより 2004 年に提案された手法であり，カーネル空間中で学習済みのデータセットから最も遠いデータ，則ちラベルありデータ集合に最も似ていないデータのラベル付けを行う．学習済みデータ $\boldsymbol{s}$ からラベルなしデータ $\boldsymbol{x}$ までのカーネル空間中での距離を $d_{kff}(\boldsymbol{s}, \boldsymbol{x})$ とすると $d^2_{kff}(\boldsymbol{s}, \boldsymbol{x})$ は式 (3) で表される．

$$\begin{aligned} d^2_{kff}(\boldsymbol{s}, \boldsymbol{x}) &= \|\phi(\boldsymbol{s}) - \phi(\boldsymbol{x})\|^2 & (2) \\ &= K(\boldsymbol{s}, \boldsymbol{s}) + K(\boldsymbol{x}, \boldsymbol{x}) - 2K(\boldsymbol{s}, \boldsymbol{x}) & (3) \end{aligned}$$

サンプリングのアルゴリズムを Algorithm 2 に示す．3 行目ではラベルなしデータ $\boldsymbol{x}$ と全ラベルありデータとの距離の 2 乗の和 $D_{kff}(\boldsymbol{x})$ を求め，5〜7 行目で $D_{kff}(\boldsymbol{x})$ が最大となるデータ $\boldsymbol{x}$ にラベルを与え，ラベルありデータ集合に加えている．

SIMPLE 法とは異なりデータセット全体を大局的に見るため，ラベルありデータの少ない学習初期でも判別面の大まかな形状を知ることができる．データセッ

†東北大学大学院情報科学研究科，Graduate School of Information Science, TOHOKU University
‡KDDI 研究所，KDDI R&D Laboratories, Inc.

**Algorithm 2** KFF 法

**Require:** $L$:ラベルありデータ集合, $U$:ラベルなしデータ集合

**loop**
  **for all** $\boldsymbol{x} \in U$ **do**
    $D_{kff}(\boldsymbol{x}) \leftarrow \sum_{\boldsymbol{s} \in L} (K(\boldsymbol{s},\boldsymbol{s}) + K(\boldsymbol{x},\boldsymbol{x}) - 2K(\boldsymbol{s},\boldsymbol{x}))$
  **end for**
  $\boldsymbol{x} \leftarrow \arg\max D_{kff}(\boldsymbol{x})$
  $y \leftarrow$LABEL($\boldsymbol{x}$)
  $L \leftarrow L \cup (\boldsymbol{x}, y)$
  $U \leftarrow U \setminus \boldsymbol{x}$
**end loop**

トによってはSIMPLE 法を大きく上回る効果があるが，その他のデータではSIMPLE 法やランダムサンプリングに劣ることが実験的に示されている [1].

### 2.3 SIMPLE 法と KFF 法の組み合わせ

Baram らは KFF 法と同時に SIMPLE 法と KFF 法などの手法を組み合わせた COMB 法を提案している [1]．COMB 法では多腕バンディット問題の概念を用いて，各手法の有効性の評価値と各手法から求めたサンプリング対象データの優先度から，総合的にサンプリングを行う．比較的多くのデータセットで高い性能を示すが，総データ数が少ないようなデータセットでは性能が悪く，また毎回複雑なデータ構造を更新していく必要があるという難点がある．Osugi らによる EXPLORATION 法は COMB 法よりも手法の組み合わせアルゴリズムを平易にしたもので [4]，KFF 法から SIMPLE 法へ確率的に切り替える手法であり，KFF 法が有効となるようなデータセットでは高い効果があるが，そうでなければ SIMPLE 法と同等かそれ以下となる．COMB 法と EXPLORATION 法それぞれのアルゴリズムを Algorithm 3，4 に示す.

### 2.4 既存手法のまとめ

本章では，SVM の能動学習で用いられる既存のデータサンプリング手法について述べた．それぞれの特徴をまとめたものを表 1 に示す．最も一般的なサンプリング手法である SIMPLE 法はデータセット全体の中で判別面のすぐ近くのみを局所的にサンプリングする手法であり，判別面を効率よく安定させていくことが出来るが，同時に学習初期の性能が悪く，過学習に陥りやすいという問題点がある．KFF 法は，データセット全体から大域的にデータサンプリングすることが出来るが，学習がある程度進んだ後は効率が悪く，また有効なデータセットも限定的である．これらの手法を学習状況に応じて組み合わせる COMB 法や EXPLORATION 法などの手法では，サンプリングの効率は向上するがアルゴリズムが複雑すぎるなどの課題がある．これらの問題点を解決するため，本研究では判別面の近傍を局所的にサンプリングする SIMPLE 法を大域的に拡張した手法を提案する.

**Algorithm 3** COMB 法

**Require:** $L$:ラベルありデータ集合，$U$:ラベルなしデータ集合，ALG$_j$：サンプリング法

$j = 1, ..., k, w_j = 1$
**for** $t = 1, 2, ...$ **do**
  **for** j=1,...,k **do**
    $\boldsymbol{e}^j(t) \leftarrow (e_1^j(t), ..., e_n^j(t))$ using ALG$_j$
  **end for**
  **for** j=1,...,k **do**
    **for** i=1,...,n **do**
      $b_i^j(t) \leftarrow (\exp\{-\beta(1 - e_i^j(t))\}/Z)$
    **end for**
    $\boldsymbol{b}^j(t) \leftarrow (b_1^j(t), ..., b_n^j(t))$
  **end for**
  **repeat**
    **for all** $\boldsymbol{x} \in U$ **do**
      **if** $\max_j b_i^j > \alpha$ **then**
        $U_e \leftarrow U_e \cup \boldsymbol{x}_i$
      **end if**
    **end for**
    $\alpha \leftarrow \alpha/2$
  **until** $|U_e| \neq 0$
  $\gamma \leftarrow \sqrt{\frac{n_e \ln k}{(e-1) g_m a x}}$
  $W \leftarrow \sum_{j=1}^k w_j$
  **for** $i = 1, ..., n_e$ **do**
    $p_i \leftarrow (1 - \gamma) \sum_{j=1}^k w_j b_j^j(t)/W + \gamma/n_e$
  **end for**
  Randomly draw a point $x_q \in |U_e|$ according to $p_1, ..., p_{n_e}$
  $y_q \leftarrow$LABEL($x_q$)
  $L_t \leftarrow L_{t-1} \cup \{(x_q, y_q)\}$
  $U_{t+1} = U_t \setminus \{x_q\}$
  Compute $C_t$ using SVM($L_t$).
  $H_t \leftarrow H_t(\frac{|C^{+1}(U)|}{|U|})$
  $r(x_q) \leftarrow ((e^{H_t} - e^{H_{t-1}}) - (1 - e)/(2e - 2)$
  **for** $i = 1, ..., n$ **do**
    **if** $i = q$ **then**
      $\hat{r}_i(t) \leftarrow r(x_q)/p_q$
    **else**
      $\hat{r}_i(t) = 0$
    **end if**
  **end for**
  $w_j(t + 1) \leftarrow w_j(t) \exp(\boldsymbol{b}^j(t) \cdot \hat{r}(t)/n_e)$
**end for**

**Algorithm 4** EXPLORATION 法

**Require:** $L$:ラベルありデータ集合，$U$:ラベルなしデータ集合，$\lambda, \epsilon$:パラメータ

$p_0 = 1$
**for** $t = 1, 2, ...$ **do**
  rand $\leftarrow$ random number generated unigormly between 0 and 1
  **if** rand $< p_{t-1}$ **then**
    $\boldsymbol{x} \leftarrow$KFF($U, L$)
  **else**
    $\boldsymbol{x} \leftarrow$SIMPLE($U, h_{t-1}$)
  **end if**
  $y \leftarrow$LABEL($\boldsymbol{x}$)
  $L \leftarrow L \cup (\boldsymbol{x}, y)$
  $h_t \leftarrow$SVM($L$)
  $U \leftarrow U \setminus \boldsymbol{x}$
  $p_t \leftarrow \max(\min(p_{t-1} \exp(d(h_t, h_{t-1})\lambda), 1 - \epsilon)\epsilon)$
**end for**

表 1: SIMPLE 法と KFF 法の特徴

| | SIMPLE 法 |
|---|---|
| 概念 | 局所的 |
| 利点 | 学習後期に有効 |
| 欠点 | 学習初期で得たデータによっては悪化 |
| | KFF 法 |
| 概念 | 大局的 |
| 利点 | 学習初期に有効 |
| 欠点 | 有効なデータセットが限定的 |

表 2: 組み合わせ手法の特徴

| | COMB 法 |
|---|---|
| 概念 | 各手法の有効性を総合 |
| 利点 | 多くのデータセットで高い効果 |
| 欠点 | アルゴリズムが複雑 |
| | EXPLORATION 法 |
| 概念 | KFF 法から SIMPLE 法へ確率的に移行 |
| 利点 | アルゴリズムが平易，X-OR 型問題で高い効果 |
| 欠点 | 有効なデータセットが限定的 |

## 3. 提案手法

本研究で新たに提案する手法は，SIMPLE 法を基にしつつサンプリングするデータの判別面からの距離の閾値を設定し，その閾値よりも判別面からの距離が大きいデータの中から最も近いデータをサンプリングするものである．データ $\boldsymbol{x}$ から判別面までの距離は SIMPLE 法と同様に $|f(\boldsymbol{x})|$ としてこの値の閾値を設定する．このように，現在の判別面に密接したデータのみをサンプリングすることを避け，従来の SIMPLE 法を大局的に拡張することにより，複数の手法を用いてアルゴリズムを複雑にすることなく，サンプリング効率の改善が期待できる．サンプリングのアルゴリズムを Algorithm 5 に示す．第 4 行目で判別面からの距離を判定し，第 5〜9 行目でこれが閾値 $t$ よりも大きいデータ $\boldsymbol{x}$ のうち最小となるもののラベル付けを行い，学習データに加えている．

**Algorithm 5** 提案手法

**Require:** $L$:ラベルありデータ集合, $U$:ラベルなしデータ集合, $t$:閾値

1: **loop**
2: compute $\boldsymbol{\alpha}$ and $b$ using SVM($L$)
3: **for all** $\boldsymbol{x} \in U$ **do**
4: **if** $\left|\sum_{s\in L} \alpha_i K(s,x) + b\right| > t$ **then**
5: $d(\boldsymbol{x}) \leftarrow \left|\sum_{s\in L} \alpha_i K(s,x) + b\right|$
6: **end if**
7: **end for**
8: $y \leftarrow \mathrm{LABEL}(\arg\min_{\boldsymbol{x}} d(\boldsymbol{x}))$
9: $L \leftarrow L \cup (\boldsymbol{x}, y)$
10: $U \leftarrow U \setminus \boldsymbol{x}$
11: **end loop**

## 4. 実験

### 4.1 実験手順

実験には RCV1(Reuters Corpus Volume 1)[3] と呼ばれるデータセットを用いた．このデータセットは単語抽出アルゴリズムである TF-IDF を用いて Reuters によるニュース記事を重み付けしたものである．ここでは CCAT(Corporate/Industrial) と GCAT(Government/Social) の 2 つのクラスの文書分類を考え，学習用データ数は 20,242，テスト用データ数は 677,399 としており互いのデータに重複は無い．実験手順を以下に示す．

1. 閾値 $t$ を設定する．
2. 全学習用データをラベルなしデータとみなす．
3. 学習用データから 10 データをランダムにサンプリングし，それぞれのラベルを求める．
4. 3. で得た 10 データから SVM を用いて判別面を生成する．
5. 生成された判別面を用いて提案手法により 10 データをサンプリングし，それぞれのラベルを求める．
6. 5. で得た 10 データを学習用データに加え SVM を用いて判別面を生成する．
7. 5. および 6. を繰り返す．
8. テスト用データセットを各判別面に適用し，正解率を求める．

### 4.2 実験結果

図 1 は学習データ数とテスト時の正解率の関係である．正解率は 10 セットのランダムな 10 データからデータを増加させていったときの平均をとっている．simple が通常の SIMPLE 法，$t = 0.1$，$t = 0.5$ がそれぞれ閾値を 0.1，0.5 と設定したときの提案手法，randam がランダムに学習データを増加させた場合である．なお SVM の実装は UniverSVM[6] を用いた．

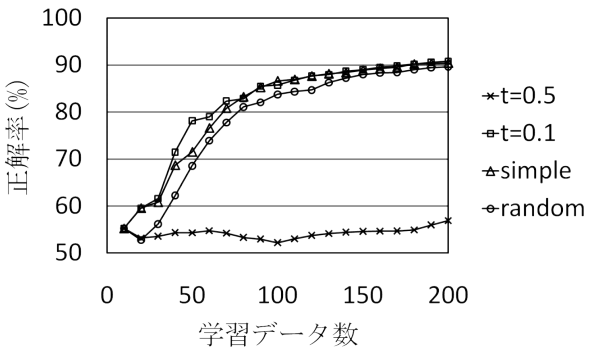


図 1: RCV1 を用いた分類実験の正解率

また，判別面から各ラベルなしデータまでの距離の平均を計測すると図 2 のように結果が得られた．

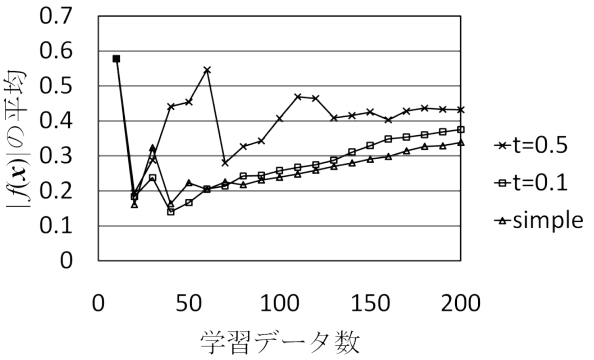


図 2: RCV1 を用いた分類実験の判別面から各ラベルなしデータまでの平均距離

### 4.3 考察

図 1 では，通常の SIMPLE 法と提案手法（$t = 0.1$）はいずれもランダムサンプリングにより学習データを増加させた場合よりも高い正解率が示されている．ただし，$t = 0.5$ では非常に効率が悪くなっている．また，$t = 0.1$ では通常の SIMPLE 法よりも学習初期で速く正解率が上昇しているのが分かる．これは，通常の SIMPLE 法が学習初期で見逃していた重要なデータが，提案手法を用いた場合は効率よくサンプリングされていることを表している．なお閾値は $t = 0.1, 0.5$ の他，$0 < t < 0.1$ の間で幾つかの値で実験を行ったが概ね通常の SIMPLE 法より良い結果を示し，今回用いた閾値の中では全体的には $t = 0.1$ が最も高い結果を示した．

図 2 に関しては，ソフトマージンの幅が 1 ということを考慮すると全体的に小さい値となっている．これは学習データ数に比べデータセット全体が非常に大規模であるためであると考えられる．また，ソフトマージンの幅は SVM に与えるパラメータにも依存するため，閾値 $t$ を決定する際にはこれらの点を考慮する必要があると考えられる．また，学習初期は安定していないが，$t = 0.5$ を除きデータ数が 50 を超えると単調に増加するグラフとなっている．これは，毎回のサンプリングでソフトマージンの実際の幅が小さくなっており，相対的に判別面からデータ $\boldsymbol{x}$ までの距離 $|f(\boldsymbol{x})|$ が大きくなっているためと考えられる．$t = 0.5$ での平均距離は安定せず，閾値の取りようによっては判別面の形成に悪影響を与え得ると言える．$0 < t < 0.1$ の間の複数の値で同様に測定を行った場合では，ほぼ $t = 0, 0.1$ と同様の結果が得られた．

## 5. 結論

本研究では，SVM 学習におけるデータサンプリング法の既存手法の問題点の検証を行い，最も一般的な手法である SIMPLE 法を大域的に拡張した手法を提案し，ベンチマークデータセットである RCV1 を用いた実験を行った．その結果，複数の手法を組み合わせるなどの複雑なアルゴリズムを構築することを避けつつ，特に学習初期のサンプリング効率の向上に成功した．

## 6. 今後の課題

今回は RCV1 を用いた実験を行ったが，他のデータセットについても実験を行い，手法の一般性を検めたい．併せて既存手法との比較も行いたい．閾値 $t$ については，今回検証を行った中では $t = 0.1$ が最も高い性能を示したが，$t = 0.5$ では著しく性能が低下した．このように，提案手法では閾値の選び方が最大の課題である．さらに，閾値の取り方によっては，密集しているデータの周囲から集中してデータを取ってしまい効率が低下してしまっていることも考えられる．こうした事態を避けるため，新たにクラスタリングなどの手法を用いた学習手法について検討を進めたい．初めにランダムサンプリングした 10 データに関しても，これによって結果が大きく変わっていたため，この初めのデータを如何に選択するかも大きな問題である．

## 参考文献

[1] Y. Baram, R. El-Yaniv, K. Luz, “Online Choice of Active Learning Algorithms”, Journal of Machine Learning Research, vol.5, pp.255-291, 2004.

[2] C. Cortes and V. Vapnik, “Support-Vector Networks”, Machine Learning, vol.20, no.3, pp.273-297, 1995.

[3] D. Lewis, et al., “RCV1: A New Benchmark Collection for Text Categorization Research”, Journal of Machine Learning Research, Vol.5, pp.361-397, 2004.

[4] T. Osugi, D. Kun, S. Scott, “Balancing Exploration and Exploitation : A New Algorithm for Active Machine Learning”, 5th IEEE International Conference on Data Mining, pp.330-337, 2005

[5] F. Sinz, R. Collbert, J. Weston, L. Bottou, “UniverSVM, Support Vector Mahine with Large Scale CCCP Functionality”, `http://3t.kyb.tuebingen.mpg.de/bs/people/fabee/universvm.html`

[6] S. Tong and D. Koller, “Support Vector Machine Active Learning with Applications to Text Classification”, Journal of Machine Learning Research, pp.999-1006, 2000.