Vol.2021-MPS-133 No.13
Vol.2021-BIO-66 No.13
2021/6/29

大規模不均衡データに対する 2段階無作為抽出法の提案

田川 聖治[1,a)]

概要：教師あり学習に基づく分類器の性能は，教師データ（特徴量とラベル）の内容に大きく左右される．本稿では，2 クラス問題を対象とした分類器の学習において，大規模不均衡データから教師データを生成するために 2 段階無作為抽出法を提案する．まず，大規模不均衡データでは，正常なデータに対する異常なデータの比率が極端に低いものとする．そこで，提案する 2 段階無作為抽出法では，特徴量が含まれる領域の選択と，特徴量（標本）の選択を分けることで，データの密度が疎な領域から教師データが選ばれる確率を高くする．また，分類器にサポートベクトルマシン（SVM）を用いた数値実験により，SVM の性能は一般的な無作為抽出法と比較し，提案法による教師データの方が高くなることを確認する．

Two-Stage Random Sampling Technique For Large Imbalanced Datasets

Tagawa Kiyoharu[1,a)]

1. はじめに

人工知能の主要な技法の 1 つに機械学習がある．さらに，機械学習の技法は，教師あり学習，教師なし学習，強化学習に大別される．ここで，教師あり学習に基づく分類器は，特徴量（実数ベクトル）を入力とし，それが属するクラスのラベルを出力とする．代表的な分類器として，深層学習を含むニューラルネットやサポートベクトルマシン（SVM：Support Vector Machine）[1] が挙げられる．

教師あり学習に基づく分類器の構築には，教師データ（特徴量とラベル）を必要とする．また，その教師データの質と量が，分類器の性能を大きく左右する．このため，様々な教師データの生成法が報告されている [2]．ここで，教師データの生成法は，分類器を使用するものと使用しないものに大別できる．前者としては，学習結果に基づき教師データを追加する増分学習 [3] や能動学習 [4] がある．また，後者としては，教師なし学習であるクラスタリングや教師データの構造解析に基づく技法 [5] がある．

少数派と多数派のクラスに属する特徴量の比率が大きく異なる不均衡データに対しては，多数派の教師データを減らすアンダーサンプリングと，少数派の教師データを増やすオーバーサンプリングがある [6, 7]．例えば，基本的なオーバーサンプリング手法である SMOTE [6] では，少数派の教師データの近傍に人工的なデータを合成する．

近年, 無線センサネットワークや IoT（Internet of Things）など情報通信技術の発達により，様々な分野でビッグデータと呼ばれる膨大なデータが蓄積されている [8]．そこで，著者は特徴量の大規模不均衡データから分類器の教師データを抽出する技法の開発に取り組んでいる．ただし，大規模不均衡データに含まれる特徴量のクラスは未知であり，その診断には時間や費用が掛かるものとする．先に提案した空間層化抽出法 [9] では，大規模データの主成分分析を必要とし，適切な層数の設定方法が課題であった．

本稿では，大規模不均衡データから教師データを生成する 2 段階無作為抽出法を提案する．まず，大規模データに対しては，計算負荷の観点から，巧妙な教師データの生成法よりも単純な無作為抽出法が適している．さらに，提案

[1] 近畿大学 理工学部
School of Science and Engineering,
Kindai University, Higashi-Osaka 577-8502 Japan
[a)] tagawa@info.kindai.ac.jp

する 2 段階無作為抽出法では，特徴量が含まれる領域の選択と，特徴量 (標本) の選択を分けることで，データの密度が疎な領域から教師データが選ばれる確率を高くする．

分類器に SVM を用いた数値実験の結果から，一般的な無作為抽出法と比較して，提案法による教師データで学習した方が SVM の性能は高くなることが確認できた．

2. 2 クラス問題

特徴量 $\boldsymbol{x}_i \in \Re^D$ の大規模なデータ $\boldsymbol{B} = \{\boldsymbol{x}_i\}$ が与えられ，各特徴量 $\boldsymbol{x}_i \in \boldsymbol{B}$ は異常（Positive）か正常（Negative）の何れか一方のクラスに属するものとする．また，異常なデータを $\boldsymbol{B}^+ \subseteq \boldsymbol{B}$ とし，正常なデータを $\boldsymbol{B}^- \subseteq \boldsymbol{B}$ とする．特徴量が $\boldsymbol{x}_i \in \boldsymbol{B}^+$ であるとき，そのラベルを $y_i = 1$ とし，$\boldsymbol{x}_i \in \boldsymbol{B}^-$ であるとき $y_i = -1$ とすれば，

$$\left[\begin{array}{rcl} \boldsymbol{B} & = & \boldsymbol{B}^+ \cup \boldsymbol{B}^- \\ \boldsymbol{B}^+ & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, y_i = 1\} \\ \boldsymbol{B}^- & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, y_i = -1\} \end{array}\right. \tag{1}$$

となる．ただし，$\boldsymbol{B}^+ \cap \boldsymbol{B}^- = \emptyset$ である．

上記のデータ $\boldsymbol{B}$ のサイズは非常に大きく，かつ，特徴量 $\boldsymbol{x}_i \in \boldsymbol{B}$ のラベル $y_i \in \{-1, 1\}$ の診断には時間や費用が掛かるものとする．そこで，2 クラス問題として，教師あり学習に基づく分類器を用いて特徴量のラベルを推定する．したがって，分類器の入力は特徴量 $\boldsymbol{x}_i \in \boldsymbol{B}$ であり，出力は推定されたラベル $\hat{y}_i \in \{-1, 1\}$ である．ここで，分類器の学習に使用する教師データ（特徴量とラベル）を

$$\boldsymbol{T} = \{(\boldsymbol{x}_n, y_n) \,|\, \boldsymbol{x}_n \in \boldsymbol{S} \subseteq \boldsymbol{B}, y_n \in \{1, -1\}\} \tag{2}$$

とする．ただし，教師データのサイズ $N = |\boldsymbol{T}|$ は，大規模なデータ $\boldsymbol{B}$ に比べて非常に小さく $N \ll |\boldsymbol{B}|$ である．

教師データ $(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}$，$n = 1, \cdots, N$ に依存する分類器の決定関数を $f : \boldsymbol{B} \to \Re$ として，決定関数 f の値に基づき $\boldsymbol{x}_i \in \boldsymbol{B}$ のラベルの推定値 $\hat{y}_i \in \{-1, 1\}$ を

$$\hat{y}_i = \begin{cases} 1 & \text{if } f(\boldsymbol{x}_i \,|\, \boldsymbol{T}) > 0 \\ -1 & \text{if } f(\boldsymbol{x}_i \,|\, \boldsymbol{T}) < 0 \end{cases} \tag{3}$$

とする．

分類器を用いて大規模なデータ $\boldsymbol{B}$ を分割すると，

$$\left[\begin{array}{rcl} \boldsymbol{B} & = & \widehat{\boldsymbol{B}}^+ \cup \widehat{\boldsymbol{B}}^- \\ \widehat{\boldsymbol{B}}^+ & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \hat{y}_i = 1\} \\ \widehat{\boldsymbol{B}}^- & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \hat{y}_i = -1\} \end{array}\right. \tag{4}$$

となる．ただし，$\widehat{\boldsymbol{B}}^+ \cap \widehat{\boldsymbol{B}}^- = \emptyset$ である．

分類器の推定は常に正しく，$y_i = \hat{y}_i$ であるとは限らない．そこで，大規模なデータ $\boldsymbol{B}$ は，正解のラベルと推定されたラベルの値に基づき，以下のように分割される．

Algorithm 1 Simple Random Sampling (SRS)

```
Input B: data set; N: sample size
Output S: a set of samples
Begin algorithm
S ← ∅
while |S| < N do
    Select a sample x_n ∈ B randomly.
    S ← S ∪ {x_n}
end while
End algorithm
```

$$\left[\begin{array}{rcl} \boldsymbol{B} & = & \boldsymbol{B}_{TP} \cup \boldsymbol{B}_{FN} \cup \boldsymbol{B}_{FP} \cup \boldsymbol{B}_{TN} \\ \boldsymbol{B}_{TP} & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, y_i = 1 \wedge \hat{y}_i = 1\} \\ \boldsymbol{B}_{FN} & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, y_i = 1 \wedge \hat{y}_i = -1\} \\ \boldsymbol{B}_{FP} & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, y_i = -1 \wedge \hat{y}_i = 1\} \\ \boldsymbol{B}_{TN} & = & \{\boldsymbol{x}_i \in \boldsymbol{B} \,|\, y_i = -1 \wedge \hat{y}_i = -1\} \end{array}\right. \tag{5}$$

式 (5) の各集合のサイズを以下のように表記する．

$$\left[\begin{array}{rl} \text{True Positive:} & TP = |\boldsymbol{B}_{TP}| \\ \text{False Negative:} & FN = |\boldsymbol{B}_{FN}| \\ \text{False Positive:} & FP = |\boldsymbol{B}_{FP}| \\ \text{True Negative:} & TN = |\boldsymbol{B}_{TN}| \end{array}\right. \tag{6}$$

式 (6) のサイズから，分類器の性能の評価指標として，適合率（Precision），再現率（Recall），正解率（Accuracy）を，式 (7) から式 (9) のように定義する [1].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{9}$$

3. 教師データの抽出法

大規模なデータ $\boldsymbol{B}$ から特徴量の標本 $\boldsymbol{x}_n \in \boldsymbol{S} \subseteq \boldsymbol{B}$，$n = 1, \cdots, N$ を抽出した後，それらのラベル $y_n \in \{-1, 1\}$ を診断して式 (2) の教師データ $\boldsymbol{T}$ とする．標本の抽出法として，最も一般的な無作為抽出法（SRS：Simple Random Sampling）[10,11] を紹介する．次に，2 段階無作為抽出法（2RS：Two-Stage Random Sampling）を提案する．

3.1 無作為抽出法

SRS はデータ $\boldsymbol{B}$ から N 個の標本 $\boldsymbol{x}_n \in \boldsymbol{S}$ をランダムに選択する．SRS の擬似コードを Algorithm 1 に示す．

3.2 2 段階無作為抽出法

大規模なデータ $\boldsymbol{B} \subseteq \Re^D$ を覆う超立方体 $\boldsymbol{\Phi} \subseteq \Re^D$ を設定する．次に，超立方体の各軸を H 等分することで，以下のように超立方体を M 個のセル $\boldsymbol{\Phi}_m \subseteq \Re^D$ に分割する．

Vol.2021-MPS-133 No.13
Vol.2021-BIO-66 No.13
2021/6/29

Algorithm 2 Two-Stage Random Sampling (2RS)

1: **Input** $\boldsymbol{B}$: data set; N: sample size
2: **Output** $\boldsymbol{S}$: a set of samples
3: **Begin algorithm**
4: $\boldsymbol{S} \leftarrow \emptyset$
5: Define cells: $\boldsymbol{\Phi}_m \subseteq \boldsymbol{\Phi}$, $m = 1, \cdots, M$.
6: **while** $|\boldsymbol{S}| < N$ **do**
7: **repeat**
8: Select a cell $\boldsymbol{\Phi}_m \subseteq \boldsymbol{\Phi}$ randomly.
9: **until** $\boldsymbol{\Phi}_m \cap \boldsymbol{B} \neq \emptyset$
10: Select a sample $\boldsymbol{x}_n \in \boldsymbol{\Phi}_m \cap \boldsymbol{B}$ randomly.
11: $\boldsymbol{S} \leftarrow \boldsymbol{S} \cup \{\boldsymbol{x}_n\}$
12: **end while**
13: **End algorithm**

$$\boldsymbol{\Phi} = \boldsymbol{\Phi}_1 \cup \cdots \cup \boldsymbol{\Phi}_m \cup \cdots \boldsymbol{\Phi}_M \tag{10}$$

ただし，次元数 D からセル数は $M = H^D$ である．

式 (10) の超立方体 $\boldsymbol{\Phi}$ のグリッドに基づき，データ $\boldsymbol{B}$ は以下のような部分集合 $\boldsymbol{B}_m \subseteq \boldsymbol{B}$ に分割できる．

$$\boldsymbol{B} = \boldsymbol{B}_1 \cup \cdots \cup \boldsymbol{B}_m \cup \cdots \boldsymbol{B}_M \tag{11}$$

ただし，$\boldsymbol{B}_m = \boldsymbol{\Phi}_m \cap \boldsymbol{B}$，$m = 1, \cdots, M$ とする．

提案する 2RS は，データの部分集合 $\boldsymbol{B}_m \subseteq \boldsymbol{B}$ をランダムに選んだ後，その中から標本 $\boldsymbol{x}_n \in \boldsymbol{S}$ をランダムに選択する．2RS の擬似コードを Algorithm 2 に示す．

各セル $\boldsymbol{\Phi}_m \subseteq \boldsymbol{\Phi}$ に含まれるデータ数の平均は

$$E = \frac{1}{M} \sum_{m=1}^{M} |\boldsymbol{\Phi}_m \cap \boldsymbol{B}| = \frac{|\boldsymbol{B}|}{M} \tag{12}$$

である．

提案する 2RS では，式 (10) のセル数 M を $M = E$ とする．したがって，式 (12) と $M = H^D$ の関係から

$$H = \mathrm{ceil}\left(|\boldsymbol{B}|^{\frac{1}{2D}}\right) \tag{13}$$

となる．H は超立方体の各軸の分割数である．

図 1 に特徴量のデータ $\boldsymbol{B} \subseteq \Re^D$ に対するグリッドの例を示す．データのサイズは $|\boldsymbol{B}| = 600$，次元数は $D = 2$ である．したがって，式 (13) より分割数は $H = 5$ となる．

図 2 に 2RS のイメージを示す．まず，第 1 段でグリッドのセル $\boldsymbol{\Phi}_m$ をランダムに選ぶ．次に，第 2 段で選択したセル内から 1 つの標本 $\boldsymbol{x}_n \in \boldsymbol{B}_m$ をランダムに選ぶ．

従来の無作為抽出法（SRS）により，特徴量 $\boldsymbol{x}_i \in \boldsymbol{B}$ が標本に選ばれる確率は，式 (12) の E と M から

$$\Pr(\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \mathrm{SRS}) = \frac{1}{|\boldsymbol{B}|} = \frac{1}{M} \times \frac{1}{E} \tag{14}$$

である．一方，2RS により $\boldsymbol{x}_i \in \boldsymbol{B}$ が選ばれる確率は，

$$\Pr(\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \mathrm{2RS}) = \frac{1}{M} \times \frac{1}{|\boldsymbol{B}_m|} \tag{15}$$

となる．ただし，$\boldsymbol{x}_i \in \boldsymbol{B}_m = \boldsymbol{\Phi}_m \cap \boldsymbol{B}$ とする．

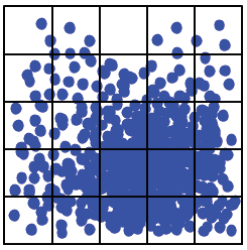

図 1 大規模データ $\boldsymbol{B}$ と超立方体 $\boldsymbol{\Phi}$ のグリッド

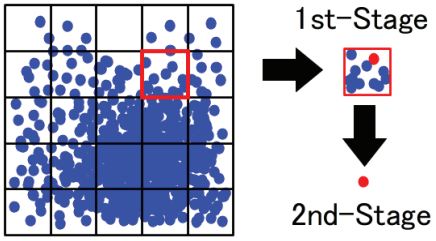

図 2 2 段階無作為抽出法による標本の選択

式 (14) と式 (15) から，$|\boldsymbol{B}_m| \leq E$ のとき

$$\Pr(\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \mathrm{2RS}) \geq \Pr(\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \mathrm{SRS}) \tag{16}$$

となり，セル内のデータ数が $|\boldsymbol{B}_m| \geq E$ のとき

$$\Pr(\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \mathrm{2RS}) \leq \Pr(\boldsymbol{x}_i \in \boldsymbol{B} \,|\, \mathrm{SRS}) \tag{17}$$

となる．

式 (16) と式 (17) から，従来の SRS と比較して，提案する 2RS によれば，データの密度が低い領域では $\boldsymbol{x}_i \in \boldsymbol{B}$ が標本に選ばれる確率が高くなる．逆にデータの密度が高い領域では $\boldsymbol{x}_i \in \boldsymbol{B}$ が標本に選ばれる確率は低くなる．

3.3 教師データによる母比率の推定

標本 $\boldsymbol{x}_n \in \boldsymbol{S}$，$n = 1, \cdots, N$ のラベル $y_n \in \{-1, 1\}$ を診断すると，教師データの標本セット $\boldsymbol{S} \subseteq \boldsymbol{B}$ は

$$\left[\begin{array}{rcl} \boldsymbol{S} & = & \boldsymbol{S}^+ \cup \boldsymbol{S}^- \\ \boldsymbol{S}^+ & = & \{\boldsymbol{x}_n \in \boldsymbol{S} \,|\, y_n = 1\} \\ \boldsymbol{S}^- & = & \{\boldsymbol{x}_n \in \boldsymbol{S} \,|\, y_n = -1\} \end{array}\right. \tag{18}$$

となる．ただし，$\boldsymbol{S}^+ \cap \boldsymbol{S}^- = \emptyset$ である．

式 (19) の母比率（Population Ratio）の推定値である標本比率（Sample Ratio）を式 (20) のように定義する．

$$\text{Population Ratio:} \quad PR = \frac{|\boldsymbol{B}^+|}{|\boldsymbol{B}^+| + |\boldsymbol{B}^-|} \tag{19}$$

$$\text{Sample Ratio:} \quad SR = \frac{|\boldsymbol{S}^+|}{|\boldsymbol{S}^+| + |\boldsymbol{S}^-|} \tag{20}$$

式 (19) の母比率 PR と式 (20) の標本比率 SR から，教師データによる母比率の推定誤差（Ratio Error）を

$$\text{Ratio Error} = |PR - SR| \tag{21}$$

とする．

Vol.2021-MPS-133 No.13
Vol.2021-BIO-66 No.13
2021/6/29

4. サポートベクトルマシン

本稿では，2 クラス問題に対する分類器に SVM を使用する．式 (2) の教師データ $(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}$，$n = 1, \cdots, N$ を用いた SVM の学習は，以下の 2 次計画問題となる．

$$\left[\begin{array}{ll} \max & q(\boldsymbol{\alpha}) = \displaystyle\sum_{n=1}^{N} \alpha_n \\ & \displaystyle -\frac{1}{2} \sum_{n=1}^{N} \sum_{j=1}^{N} \alpha_n \, \alpha_j \, y_n \, y_j \, K(\boldsymbol{x}_n, \boldsymbol{x}_j) \\ \text{s. t.} & \displaystyle\sum_{n=1}^{N} y_n \, \alpha_n = 0 \\ & 0 \le \alpha_n \le C, \; n = 1, \cdots, N \end{array}\right. \tag{22}$$

ただし，$\alpha_n \in \Re$，$n = 1, \cdots, N$ は 2 次計画問題の解（ラグランジュ乗数）である．C は正則化パラメータで SVM の汎化能力を左右する．また，K は基底関数である [1]．

本稿では，以下のラジアル基底関数を使用する．

$$K(\boldsymbol{x}_n, \boldsymbol{x}_j) = \exp\left(-\gamma \, ||\boldsymbol{x}_n - \boldsymbol{x}_j||^2\right) \tag{23}$$

ただし，$||\boldsymbol{x}||$ は $\boldsymbol{x}$ のユークリッドノルムである．

通常，式 (22) の 2 次計画問題の最適解 $\boldsymbol{\alpha} \in \Re^N$ では，多くの要素が $\alpha_n = 0$ となる．ここで，$\alpha_n > 0$ に対応する特徴量 $\boldsymbol{x}_n \in \boldsymbol{S} \subseteq \Re^D$ をサポートベクトルと呼ぶ [1]．

最適解 $\boldsymbol{\alpha} \in \Re^N$ から，式 (3) の決定関数を

$$f(\boldsymbol{x}_i \,|\, \boldsymbol{T}) = \sum_{(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}} \alpha_n \, y_n \, K(\boldsymbol{x}_n, \boldsymbol{x}_i) + b \tag{24}$$

とする．ただし，$b \in \Re$ はバイアス項である．

式 (24) のバイアス項 b は，上限に達していない任意のサポートベクトル $(\boldsymbol{x}_j, y_j) \in \boldsymbol{T}$，$0 < \alpha_j < C$ について

$$b = y_j - \sum_{(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}} \alpha_n \, y_n \, K(\boldsymbol{x}_n, \boldsymbol{x}_j) \tag{25}$$

となる．

式 (24) と式 (25) から，教師データ $(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}$ のうちサポートベクトルのみが SVM の決定関数に寄与する．

5. 数値実験

数値実験により，従来の SRS と提案した 2RS による教師データの特性を比較する．また，それぞれの教師データを用いて構築した SVM の性能を比較する．プログラムの実装と数値実験では，MATLAB [12] を使用した．

5.1 大規模不均衡データと教師データ

図 3 に 2 変量の切断正規分布に基づき人工的に生成した大規模不均衡データ $\boldsymbol{B} \subseteq \Re^2$ を示す．データのサイズは $|\boldsymbol{B}| = 10^4$ であり，中心部ほどデータの密度が高くなる．図 3 の各特徴量 $\boldsymbol{x}_i \in \boldsymbol{B}$ にはラベル $y_i \in \{-1, 1\}$ も付けて

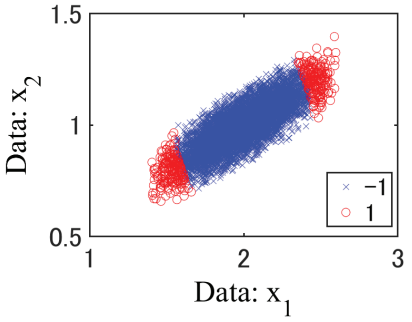

図 3 大規模不均衡データ $\boldsymbol{B} \subseteq \Re^2$ とラベル

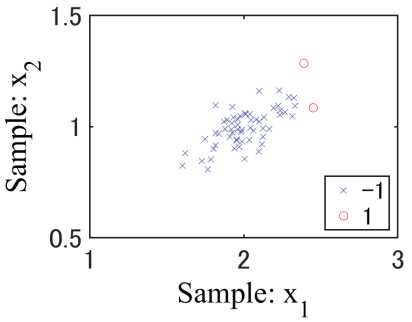

図 4 SRS による教師データ $(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}$

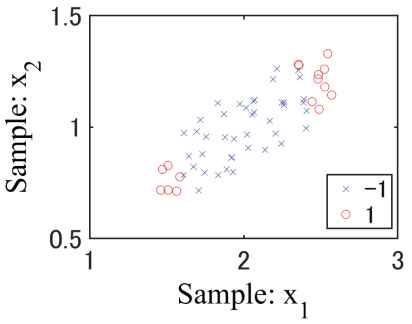

図 5 2RS による教師データ $(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}$

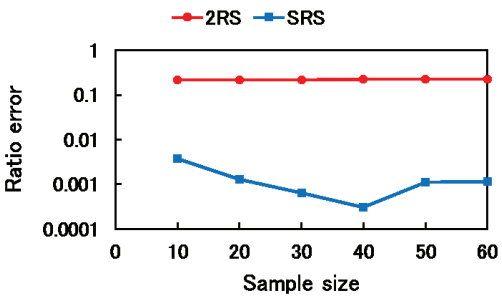

図 6 教師データによる母比率の推定誤差

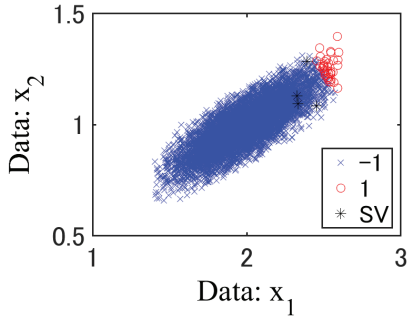

図 7 SRS で学習した SVM によるラベルの推定値

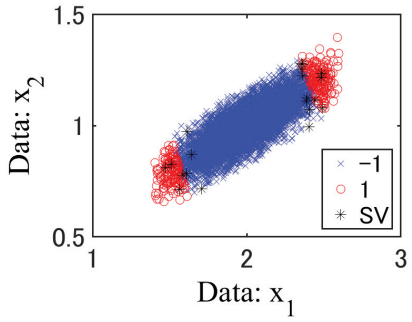

図 8 2RS で学習した SVM によるラベルの推定値

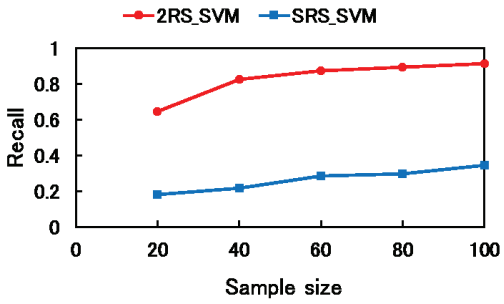

図 9 SRS と 2RS による SVM の再現率（$D = 2$）

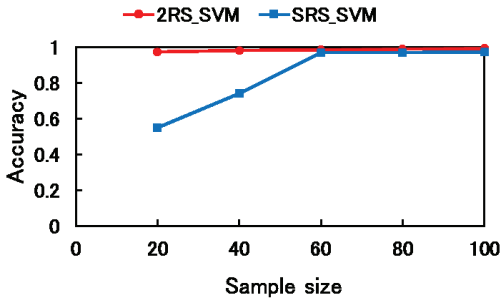

図 10 SRS と 2RS による SVM の正解率（$D = 2$）

いる．ここで，ラベルの値は $|\boldsymbol{B}^+| \ll |\boldsymbol{B}^-|$ と不均衡であり，式 (19) で定義した母比率は $PR = 0.039$ となる．

図 3 の大規模不均衡データ $\boldsymbol{B} \subseteq \Re^2$ から従来の SRS によって抽出した教師データ $(\boldsymbol{x}_n, y_n) \in \boldsymbol{T}$，$n = 1, \cdots, N$ を**図 4** に示す．標本サイズは $N = 60$ である．同様に，提案した 2RS により抽出した教師データを**図 5** に示す．

図 4 と図 5 から，従来の SRS に比べて提案した 2RS による標本 $\boldsymbol{x}_n \in \boldsymbol{S}$ は広範囲に分布している．また，2RS は大規模不均衡データの両端から均等に異常な特徴量 $\boldsymbol{x}_i \in \boldsymbol{B}$ を抽出している．一方，図 4 から SRS では標本の抽出が密度の高い大規模不均衡データの中心部に集中し，左端に存在する異常な特徴量 $\boldsymbol{x}_i \in \boldsymbol{B}$ を取得できていない．

図 6 では 2RS と SRS で生成した教師データについて，式 (21) で定義した母比率の推定誤差を比較する．図 6 の横軸は標本サイズ N であり，図 6 の結果は乱数シードが異なる 100 回の試行の平均値である．図 6 から，2RS に比べて SRS による推定誤差は非常に小さい．一方，2RS では標本サイズによらず推定誤差は概ね一定値である．

実験結果から，母比率の推定では SRS の方が提案した 2RS に勝っている．すなわち，SRS の標本比率は母比率に近く，大規模不均衡データ $\boldsymbol{B}$ が $|\boldsymbol{B}^+| \ll |\boldsymbol{B}^-|$ であるため，SRS の標本セット $\boldsymbol{S}$ でも $|\boldsymbol{S}^+| \ll |\boldsymbol{S}^-|$ となる．

5.2 教師データの異なる SVM の性能比較

図 4 の SRS による教師データで学習した SVM により，図 3 の大規模不均衡データの特性量 $\boldsymbol{x}_i \in \boldsymbol{B}$ に対して推定したラベル $\hat{y}_i \in \{-1, 1\}$ を**図 7** に示す．SV はサポートベクトルである．同様に，図 5 の 2RS による教師データで学習した SVM によるラベルの推定値を**図 8** に示す．

図 3 と図 7 から，SRS の SVM では異常な特性量 $\boldsymbol{x}_i \in \boldsymbol{B}$ の多くを誤って正常と判定している．図 4 の SRS による教師データに含まれない右端の異常な特性量を，SVM が学習していないことが原因と思われる．一方，図 3 と図 8 から，両者のラベルの値は酷似しており，2RS の SVM は異常な特性量のラベルの大半を正しく推定している．

図 9 では SRS と 2RS による教師データで学習した 2 種類の SVM について，式 (8) の再現率を比較する．また，**図 10** では上記の 2 種類の SVM について，式 (9) の正解率を比較する．図 9 と図 10 の横軸は標本サイズ N であり，図 9 と図 10 の結果は乱数シードが異なる 30 回の試行の平均値である．大規模不均衡データにおける 2 クラス問題では，分類器が特性量 $\boldsymbol{x}_i \in \boldsymbol{B}$ のラベルをすべて正常と推定し，式 (7) の適合率が算出できない場合がある．このため，適合率に基づく 2 種類の SVM の比較は省略する．

図 9 から，2RS の教師データで学習した SVM の再現率は高く，標本サイズと伴に上昇して $N = 100$ では 90% を超えている．一方，SRS による SVM の再現率は非常に低く，標本サイズが $N = 100$ でも 50% に満たない．

図 10 から，SRS と比較して，2RS による SVM の方が正解率は高く 100% に近い．ただし，標本サイズを増やすと SRS による SVM の正解率も 100% に近くなる．その理由は，大規模不均衡データでは大多数の正常な特性量を正常と判定することで，正解率が上昇するためである．

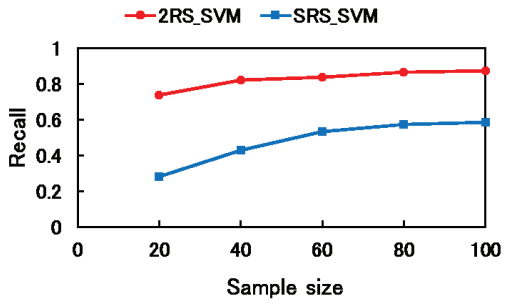

図 11 SRS と 2RS による SVM の再現率（$D = 4$）

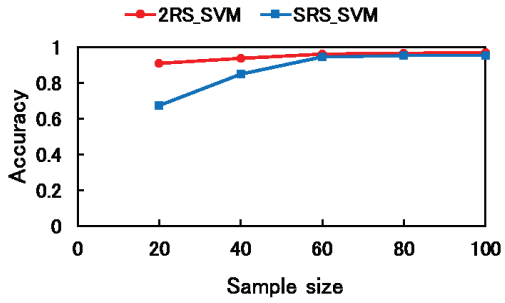

図 12 SRS と 2RS による SVM の正解率（$D = 4$）

5.3 高次元の教師データによる SVM の性能評価

4 変量の切断正規分布に基づき大規模不均衡データ $\boldsymbol{B} \subseteq \Re^4$ を人工的に生成した．データのサイズは $|\boldsymbol{B}| = 10^6$ である．ラベルの値 $y_i \in \{-1, 1\}$ は $|\boldsymbol{B}^+| \ll |\boldsymbol{B}^-|$ と不均衡であり，式 (19) の母比率は $PR = 0.081$ となる．

図 11 では SRS と 2RS による教師データで学習した 2 種類の SVM について，式 (8) の再現率を比較する．また，**図 12** では上記の 2 種類の SVM について，式 (9) の正解率を比較する．図 11 と図 12 の横軸は標本サイズ N であり，それらの結果は 30 回の試行の平均値である．

図 11 の再現率と図 12 の正解率の傾向は，図 9 と図 10 の結果と同様であり，提案した 2RS で学習した SVM の方が，従来の SRS による SVM よりも優れいる．また，図 11 の結果から，高次元の大規模不均衡データ $\boldsymbol{B} \subseteq \Re^4$ に対しても，2RS による教師データで学習した SVM は，比較的に小さな標本サイズで高い再現率を達成している．

6. おわりに

本稿では，大規模不均衡データから分類器の教師データを生成する技法として，2 段階無作為抽出法（2RS）を提案した．また，人工的に生成した大規模不均衡データと分類器に SVM を使用した数値実験において，2RS による教師データで学習した SVM の再現率は，従来の無作為抽出法（SRS）による SVM よりも高いことを確認した．

今後の課題は，現実の世界における大規模不均衡データを対象とした様々な分類器において，提案した 2RS による教師データの効果と実用性を検証することである．

参考文献

[1] Abe, S.: *Support Vector Machines for Pattern Classification (2nd ed.)*, Springer-Verlag (2010).
[2] Nalepa, J. and Kawulok, M.: Selecting training sets for support vector machines: a review, *Artificial Intelligence Review*, Vol. 52, No. 2, pp. 857–900 (2019).
[3] An, J. L., Wang, Z. O. and Ma, Z. P.: An incremental learning algorithm for support vector machine, *Proc. of the 2nd International Conference on Machine Learning and Cybernetics*, pp. 1153–1156 (2003).
[4] Schohn, G. and Cohn, D.: Less is more: active learning with support vector machines, *Proc. of the International Conference on Machine Learning*, pp. 839–846 (2017).
[5] Lopez-Chau, A., Li, X. and Yu, W.: Convex-concave hull for classification with support vector machine, *Proc. of 12th International Conference on Data Mining Workshops*, pp. 431–438 (2012).
[6] Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P.: SMOTE: synthetic minority oversampling technique, *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357 (2002).
[7] Chen, S., He, H. and Garcia, E. A.: RAMOBoost: ranked minority oversampling in boosting, *IEEE Trans. on Neural Networks*, Vol. 21, No. 10, pp. 1624–1642 (2010).
[8] Xu, L. D., He, W. and Li, S.: Internet of things in industries: a survey, *IEEE Trans. on Industrial Informatics*, Vol. 10, No. 4, pp. 2233–2243 (2004).
[9] Tagawa, K.: A support vector machine-based approach to chance constrained problems using huge data sets, *Proc. of the 52nd ISCIE International Symposium on Stochastic Systems Theory and Its Applications*, pp. 46–53 (2020).
[10] Han, J., Kamber, M. and Pei, J.: *Data Mining - Concepts and Techniques*, Morgan Kaufmann (2012).
[11] Tempo, R., Calafiore, G. and Dabbene, F.: *Randomized Algorithms for Analysis and Control of Uncertain Systems: With Applications*, Springer (2012).
[12] Martinez, A. R. and Martinez, W. L.: *Computational Statistics Handbook with MATLAB*, Chapman & Hall/CRC (2008).