不均衡データにおけるアンダーサンプリングによる分類精度向上法

情報数理応用研究　　5212C006-1　井上大樹
指導教員　後藤正幸

Classification Accuracy Improvement Method by Undersampling in Imbalanced Data

INOUE Taiki

1　研究背景・目的

近年の高度情報化に伴い，膨大な量のテキストデータが蓄積され，その有効活用が望まれている．しかし，その膨大さにより，人手での分析は物理的に不可能となりつつある．このため，コンピュータを用いたテキストデータの自動分析手法の確立が望まれる [1],[2]．テキストデータに対する分析には主に文書分類，クラスタリング，要約などがあげられるが，本研究ではこのうち文書分類に焦点をあてる．テキストの自動分類に関する研究は既に数多く手法が提案され [3]，その中でも高い分類性能を示す 2 値分類器として Support Vector Machine(以下 SVM) が提案されている [5],[6],[7]．しかし，SVM は 2 値分類器であるため，カテゴリ数が 3 以上の多値分類問題を対象とした場合，複数の 2 値分類器を組み合わせて多値分類を行うことが多い．2 値判別器の組み合わせにより多値判別を行う手法は既に様々なものが提案されているが，本研究では最も一般的な one-vs-the-rest 法に焦点をあてる．しかしながら，one-vs-the-rest 法では，正例を一つのカテゴリ，負例をその他全てのカテゴリとして学習を行うため，各カテゴリの学習データ数が等しい場合であったとしても，正例と負例の学習データ間に偏りが生じてしまう．これは，不均衡データの問題 [4] と言われ，各分類器の分類精度低下を招く [8].

本研究では，不均衡データに対するアプローチとして，one-vs-the-rest 法 [8] における負例の学習データを多数派データとして定義し，多数派データのデータ数を減少させるアンダーサンプリングの中でも，亀井らが提案した One-sided selection（以下 ONESS） [10] を対象とし，正例と負例の学習データ数の偏りを補正した分類に着目する．しかし，ONESS によるアンダーサンプリング法では，多数派データの削除数はランダムに選択されたデータに強く依存してしまう．このため，適切な削除データ数になるまで複数回の試行を必要とする点や，少数派データとの境界付近にある多数派データが選択された場合，ほとんどが分類に寄与しないデータと判断され，本来であれば分類に寄与するデータまで削除されてしまうという問題点がある．

そこで，本研究では ONESS におけるランダム性に起因する問題を解決するため，データをランダムに選択するのではなく，多数派データの重心とする方法を提案する．これにより，分類に寄与しやすい正例と負例の境界付近のデータを削除することがないため，精度の向上が期待できる．また，重心を基準とすることで数回の試行を行う必要がないため，計算量の削減も見込める．しかしながら，重心を基準に多数派データを削減しただけでは，少数派データと多数派データのバランスが十分に取れたデータセットが得られるとは限らない．そのため，重心を基準とした ONESS を行った後，さらに各少数派データと残っている多数派データのすべての組み合わせについてユークリッド距離を求め，小さいものから少数派データ数分選択する手法を提案する．これにより，分類に寄与する境界付近の学習データを削除することなく，正例と負例の学習データ数の偏りがない学習データセットを用意することができるため，分類精度を向上させることができる．提案手法の有効性を新聞記事データを用いた分類実験を行うことで示す.

2　準備

2.1　多値分類問題

分類問題とは，あらかじめカテゴリが付与された文書から分類ルールを学習し，カテゴリが未知の新規文書を与えられたカテゴリに分類する問題である．いま，学習データの文書集合 $\mathcal{X}_{learn}$ を $\mathcal{X}_{learn} = \{\boldsymbol{x}_i | i = 1, 2, \cdots, I\}$，テストデータの文書集合 $\mathcal{X}_{test}$ を $\mathcal{X}_{test} = \{\boldsymbol{x}_j^t | j = 1, 2, \cdots, J\}$ とし，カテゴリ集合を $\mathcal{C} = \{c_k | k = 1, 2, \cdots, K\}$ とする．$K \geq 3$ の場合が本研究の対象とする多値分類問題となる．本研究では 2 値分類器として SVM を用いた.

2.2　SVM

テキストデータの分類問題は素性の数が非常に多い高次元な単語頻度ベクトルを用いるため，相対的にパラメータ数が多くなり，過学習が生じやすくなる．これにより学習データに分類器が過度にフィッティングしてしまい，結果として分類精度も低下してしまうという問題がある．SVM は高次元特徴空間における 2 値分類問題を目的としており，過学習せず最適な識別関数を構成することで高精度な分類を達成することが可能となる [5].

SVM は 2 つのカテゴリに分かれたデータに対して，線形識別関数から最も近いデータまでの距離 (マージン) を最大化することでデータの識別能力を向上させる手法である．正例と負例の 2 つのカテゴリが付与された学習データを $(\boldsymbol{x}_i, y_i)$ で定義する．y_i は正例であれば $+1$ を，負例であれば -1 をとるものとする．いま，線形識別関数を以下で定義する．

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b \tag{1}$$

$\boldsymbol{w}$ は重みベクトル，b はバイアス項を示しており，$\boldsymbol{w}$ と b は制約付き最適化問題により求めることができる．正例，負例間の最も近いデータ同士のマージンは $2/||\boldsymbol{w}||$ で与えられるため，$||\boldsymbol{w}||$ が小さいほどマージンは大きくなる．また，線形識別関数により誤って分類された学習データに対して分離超平面との距離を ζ_i とする．式 (1) における正例は $\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i \geq 1$，負例は $\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i < -1$ を満たすことか

ら，SVM では以下の制約付き最適化問題を解くことで，線形識別関数を得る．

$$\min \quad \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + M\sum_{i=1}^{I}\zeta_i \tag{2}$$

$$s.t. \quad y_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i \geq 1 \tag{3}$$

$$\zeta_i \geq 0 \tag{4}$$

2.3 one-vs-the-rest 法

複数の 2 値分類器を組み合わせることで多値判別を行う方法は既に様々なものが提案されている．その一つに，あるカテゴリ c_k とそれ以外のカテゴリとを分類するための 2 値分類器をカテゴリ数分用意する one-vs-the-rest 法が存在する．しかし，この手法では，正例をひとつのカテゴリ，負例をその他全てのカテゴリとするため，各カテゴリの学習データ数が等しい場合であったとしても，多値判別問題を対象とした場合，正例のカテゴリが持つ学習データ数と負例のカテゴリが持つ学習データ数に偏りが生じてしまう．例えば，10 カテゴリがすべて等しく 100 サンプルずつの学習データを有している場合，one-vs-the-rest 法では，学習データが 1 カテゴリの正例 100 件とその他のカテゴリの不例 900 件となってしまう．学習せずに全て負例と予測しても正解率 90% を達成してしまうことになる．これは，不均衡データの問題と言われ，少数派のカテゴリに属するデータを分類することが難しくなることで，各分類器を統合したときの多値分類の精度の低下を招く．

2.4 ランダムアンダーサンプリング (RUS)

不均衡データの問題を解消する手法の一つに，多数派データのデータ数を減少させるランダムアンダーサンプリング法（RUS）がある．これは，ランダムに任意の多数派データを選択し削除する操作を繰り返す手法である．従来手法 [9] においては，ランダムアンダーサンプリングの考えを応用し，one-vs-the-rest 法による多値判別を行う．

3 従来手法

本手法では，one-vs-the-rest 法による学習データの分割を行う際に，ランダムアンダーサンプリングの考えを応用した亀井らによる手法 [9] を適用した後，SVM を用いたテキストの自動分類を行う．前述の通り，one-vs-the-rest 法はある 1 つのカテゴリとその他のカテゴリとを分類する識別関数を K 個構築するため，偏りのある学習データ (不均衡データ) を形成してしまう．そこで，K 個の 2 値分類器を作成する前の段階で，各データセットに対してアンダーサンプリングを行う．これにより，各分類器の汎化能力の向上を図る．RUS はランダムに多数派データを削除していくため，分類に寄与するデータも削除してしまう可能性がある [9]．そこで，RUS の多数派データの選択法を工夫した従来手法である ONESS がある．ONESS においては，多数派データをランダムに 1 つ選択しそのデータを代表元として最近傍法を繰り返し行うことで，冗長なデータを削除しつつ，少数派データに対しても最近傍法を適用し，近傍が多数派データであった際にノイズとして削除する．これにより多数派データの特徴を捉えながらデータ数を減らすことが可能となる．ここで，冗長なデータとは分類に寄与しないデータ，ノイズデータとは多数派データの中で統計的特徴が少数派データと類似する分類に悪影響を与えるデータを指す．以下に従来手法の概念図を示す．

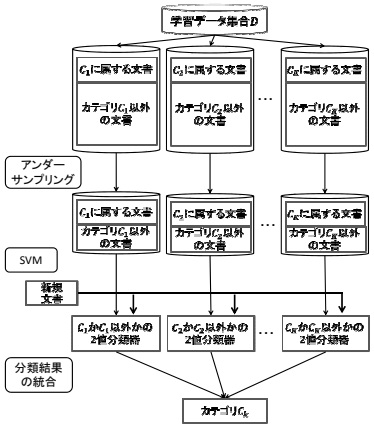

図 3. ONESS の例

$\boldsymbol{x}_{s_b}$ は多数派データのある一つの文書をとし，以下に ONESS のアルゴリズムを示す．

Step.1 データの選択
少数派データの集合と，多数派データからランダムに選んだ 1 個のデータ $\boldsymbol{x}_{s_b}$ を合わせたデータセットを G とする．$\boldsymbol{x}_{s_b}$ は多数派データのある一つの文書を示す．

Step.2 冗長なデータの削除
データセット G の少数派データを正例，$\boldsymbol{x}_{s_b}$ を負例の学習データとして，$\boldsymbol{x}_{s_b}$ 以外の多数派データをテストデータとして最近傍法による分類を行う．このテストデータが最近傍法によって $\boldsymbol{x}_{s_b}$ 最も近く多数派データと分類される場合，冗長なデータであると判断し，その学習データは使わずに取り除く．最近傍法には亀井らの手法と同様にユークリッド距離を適用する．

Step.3 ノイズデータの削除
各少数派データからの最近傍データが多数派データの場合，ノイズデータとして多数派データを学習データから取り除く．

Step.4 繰り返し
データの選択を行う際に削除する学習データ数が最初に選択された $\boldsymbol{x}_{s_b}$ に強く依存することから，指定の数になるまでデータセットを初期化し，Step.1 に戻り最初の $\boldsymbol{x}_{s_b}$ を選びなおして，データセットの生成を繰り返す．残った多数派データ数が少数派データ数より少なくならないようにする．

カテゴリ c_k の属する学習データを少数派データ，カテゴリ c_k 以外に属する学習データを多数派データとして Step.1 〜Step.4 を K 回繰り返し，K 個のデータセットの生成を行う．生成された K 個のデータセットから，識別関数を構成する．テストデータ $\boldsymbol{x}_j^t$ を全カテゴリ数分の識別関数に適用させ $f_k(\boldsymbol{x}_j^t) \geq 0$ となった場合，カテゴリ c_k に所属させる．テストデータ $\boldsymbol{x}$ が複数の識別関数で正になった場合，正になったカテゴリでランダムに割り当てる．

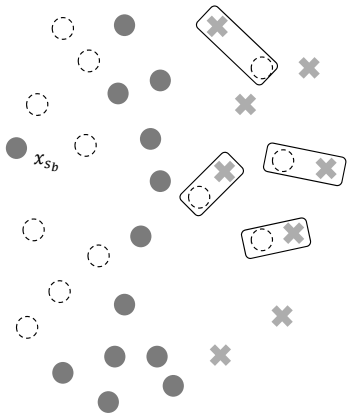

図 3. ONESS の例

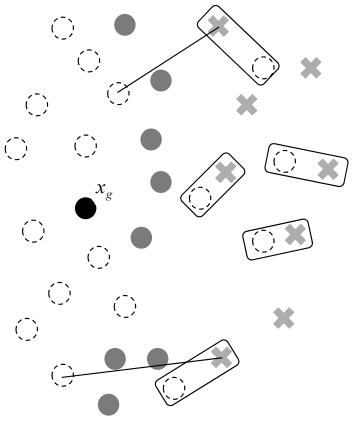

図 4. 提案手法の例

4 提案手法

4.1 重心を基準としたアンダーサンプリング

ランダム性の影響を除くため，ランダムに選択されるデータを ONESS では多数派データからランダムに選んだ 1 個のデータ $\boldsymbol{x}_{s_b}$ がたまたま境界付近にあるデータであった場合，境界付近のデータが最近傍となり，削除されてしまう．そこで，ランダム性に依存しない多数派データの重心を $\boldsymbol{x}_{s_b}$ とする方法を提案する．これにより，数回の試行を必要することなく，一定量の多数派データ数を削減することが出来る．いま，多数派データの重心を $\boldsymbol{x}_g$ としたとき，これをその算術平均で定義する．カテゴリ c_k の属する学習データを少数派データ，カテゴリ c_k 以外に属する学習データ以外を多数派データとし，従来手法の Step.1〜Step.3 を全カテゴリ数分繰り返すことで，K 個の 2 値データセットを作成する．

4.2 バランスを考慮したアンダーサンプリング

4.1 節では重心を基準としたサンプリング法を提案したが，それでも多数派データが少数派データの学習データ数より多くなる．そこで，4.1 節を考慮しつつ少数派データとの学習データ数のバランスも考慮したアンダーサンプリングを提案する．以下にアルゴリズムを示す．

Step.1-Step.3　従来手法の Step.1〜Step.3 と同様.

Step.4　データ数のバランスを考慮
残った多数派データと最も近くなる少数派データとのユークリッド距離を算出し，距離が短い多数派データの上位 A 件 (全少数派データ数 A) 以外を削除する (図 4 の実線で結ばれた上位 A 件以外の多数派データが削除される).

カテゴリ c_k の属する学習データを少数派データ，カテゴリ c_k 以外に属する学習データ以外を多数派データとして Step.1〜Step.4 を K 回繰り返す.

5 実験

提案手法の有効性を検討するため，新聞記事のデータを用いた分類実験を行い，分類精度の評価を行う.

5.1 実験条件

実験では，実験 1 として読売新聞 2005 年の 4 カテゴリの記事のデータセットと，実験 2 として読売新聞 2000 年の 8 カテゴリの記事のデータセットを用いる．記事は唯一のカテゴリに属しており記事が他のカテゴリに重複することはない．また実験 1 では各カテゴリで 600 記事ずつの合計 2400 記事をランダムに選び，学習データとして各カテゴリ 500 個，テストデータとして各カテゴリ 100 個にランダムに分けた.

比較手法には，one-vs-one 法を適用した SVM(1vs1)，one-vs-the-rest 法を適用した SVM(1vsRest)，ONESS を適用した one-vs-the-rest 法 (ONESS1)，4.1 節の方法のみを適用した one-vs-the-rest 法 (提案手法 1)，ONESS を適用後に 4.2 節の Step.4 を適用することでバランスを考慮した one-vs-the-rest 法 (提案手法 2)，4.1 節と 4.2 節の方法を両方に適用した one-vs-the-rest 法 (提案手法 2) について評価を行う.

5.2 結果と考察

5.3 結果と考察

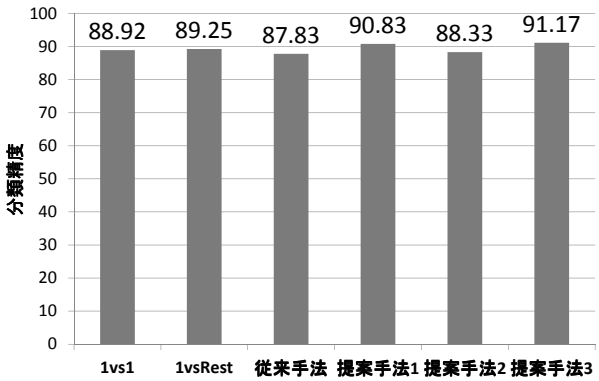

図 5. 実験 1 の分類精度

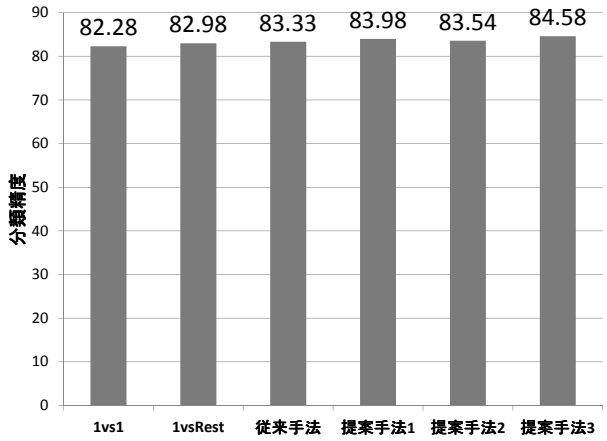

図 6. 実験 2 の分類精度

表 2. 実験 1 における手法別の多数派データ削減数の平均

手法	1vs1	1vsRest	ONESS	提案 1	提案 2	提案 3
削減数	0	0	12.58	396.75	999	1000

表 3. 実験 2 における手法別の多数派データ削減数の平均

手法	1vs1	1vsR	ONESS	提案 1	提案 2	提案 3
削減数	0	0	6.80	3.82	2999	3000

分類精度は従来手法と比較し，提案手法 1 の方が良くなったことから重心をとることで，従来手法では分類に寄与するデータが削減されてしまう可能性を軽減できた結果，有用性が示されたと考えられる．また分類精度は従来手法と比較し，提案手法 2 の方が良くなったことから正例と負例の学習データ数のバランスを考慮することによって不均衡データの問題が解決され，分類精度向上につながった．

また提案 1 から提案 3 において，提案 3 が提案 1，2 より分類精度が高く多数派データの削減数も大きい．このことから，冗長なデータとノイズデータを削除した効果だけではなく，少数派データと多数派データの学習データ数のバランスをとった両方の効果により，汎化能力を高めるモデルが構築できたためであると考えられる．

6 まとめと今後の課題

本研究では，多値分類問題における one-vs-the-rest 法の不均衡データに対して，OENSS に重心を適用した手法とバランスを考慮したアンダーサンプリングによるサンプリング法がテキストの自動分類に有効な手法であることを示した．今後の課題として，SVM 以外の分類器 (ナイーブベイズ分類器や決定木等) の適用や，少数派データを増やすオーバーサンプリングの適用，オーバーサンプリングと提案手法のアンダーサンプリングの組み合わせをした手法が考えられる．

参考文献

[1] 永田昌明，平博順，“ テキスト自動分類-学習理論の見本市-，”情報処理学会誌,Vol.42,No.1,pp. 32-37, 2000.

[2] 鈴木誠，“ カテゴリ間の単語頻度の差分を用いたテキストの自動分類 ”日本経営工学会論文誌, Vol.59,No.4,pp. 2008.

[3] 花井拓也，山村毅，“ 単語間の依存性を考慮したナイーブベイズ法によるテキスト分類，”情報処理学会研究報告，2005-NL-166(14),pp. 101-106, 2005.

[4] N.Japkowicz，“ Learning from Imbalanced Data Sets: A Comparison of Various Strategies，” *AAAI2000 Workhop, Thechnical Report WS-00-05*,pp. 10-15, 2000.

[5] 平博順，向内隆文，春野雅彦，“ Support Vector Machine によるテキスト分類，”自然言語処理研究会報告,98(99),pp. 173-180, 1998.

[6] 栗田哲平，近山隆，“ 多クラス Support Vector Machine を用いた一般物体認識での複数候補提示下における分類性能の傾向，”電子情報通信学会技術研究報告,108(328),pp. 251-258, 2008.

[7] 山田寛康，松本裕治，“ Support Vector Machine の多値分類問題への適用法について，”自然言語処理研究会報告,2001(112),pp. 33-38, 2001.

[8] 小田井良輔，雲居玄道，三川健太，後藤正幸，“ 二値判別器の組み合わせによる RVM 多値文書分類手法に関する一考察，”第 10 回情報科学技術フォーラム,pp. 425-428, 2011.

[9] 亀井靖高，“ オーバーおよびアンダーサンプリング法を用いた Fault-prone モジュール判別モデルの精度評価，”奈良先端科学技術大学院大学情報科学研究科修士論文, 2007.

[10] Nitesh V. Chawla，Kevin W. Bowyer，Lawrence O. Hall，W. Philip Kegelmeyer，“ SMOTE: Synthetic Minority Over-sampling Technique，”*Journal of Artificial Intelligence Research*,Vol.16,pp. 321-357, 2002.

[11] 竹内純一，山西健司，“ データマイニングにおける統計的外れ値検出，” 日本応用数理学会，Vol.11, No.2,pp. 71-75, 2001.