イベント関連表現に基づいた事例の偏りの補正法

櫻井　茂明†　　折原　良平†

† 株式会社 東芝 研究開発センター　〒 212–8582 神奈川県川崎市幸区小向東芝町 1
E-mail: †{shigeaki.sakurai,ryohei.orihara}@toshiba.co.jp

あらまし　テキストデータを特定のイベントの記述の有無によって分類する分類モデル学習において、イベントを含んでいるテキストデータ (正例) はイベントを含んでいないテキストデータ (負例) に比べて、その数が非常に少なくなっている。このため、分類器によって学習される分類モデルは、負例に過度に依存する傾向にある。本論文では、テキストデータの特徴を利用することにより、事例の偏りを解消する方法を提案する。提案法においては、テキストデータを特徴付けるイベントに関連する表現を利用することにより、2 種類の重要な負例を抽出する。また、残りの負例を冗長な負例とみなすことにより、事例の偏りを補正する。提案法を掲示板サイトから収集したテキストデータに対して適用し、従来法と比べた提案法の効果を示すとともに、利用する関連表現の影響を評価する。
キーワード　テキスト分類、事例の偏り、関連表現

Adjustment Method of Unbalanced Examples based on Expressions Related to an Event

Shigeaki SAKURAI† and Ryohei ORIHARA†

† Corporate Research & Development Center, Toshiba Corporation,　1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, Kanagawa, 212–8582, Japan
E-mail: †{shigeaki.sakurai,ryohei.orihara}@toshiba.co.jp

Abstract　A classification model can identify whether an item of textual data describes the contents related to a specific event or not. In the learning of this model, the number of items related to the event (positive examples) is much smaller than the number of the items unrelated to the event. Therefore, the model excessively tends to depend on the negative examples. The paper proposes a method that decreases imbalance of examples based on features of textual data. The method identifies two kinds of important negative examples by using expressions related to the event, and regards remaining examples as redundant examples. The paper also applies the method to items of textual data collected from bulletin board sites. In addition, the paper verifies its effect through the comparison of previous methods and evaluates such influence that relevant expression sets give to the classification efficiency.
Key words　Text classification, Imbalance of training examples, Relevant expressions

1. はじめに

コンピュータ環境の普及に伴って、多数のテキストデータが簡便に収集、蓄積されるようになっている。これらのテキストデータの中には人間の意思決定にとって重要な情報が含まれており、テキストデータを分析することの必要性が高まっている。研究レベルにおいては、テキストマイニング [1] [6] [7] として、1990 年代の後半から研究が活発化しており、我々のグループにおいても、ほぼ同時期から研究を開始し、成果をあげてきている [3] [11] [12] [13] [14]。

テキストマイニングといっても発見される知識には、様々なタイプのものが存在しており、現在、我々のグループでは、テキストデータからテキストの内容を代表するイベントを自動的に抽出する技術の研究開発に注力している。また、この際のベースとなる技術として、学習データに基づいて知識を発見する帰納学習法を採用している。帰納学習法においては、学習に利用するデータの良し悪しが学習される知識の良し悪しに大きく依存するため、多数の質の高い学習データを収集することが重要になっている。しかしながら、質の高い学習データを収集するには、通常多くの労力が必要となる。特に、全体としてはあまり起こらないイベントの出現を学習する場合には、そのイベントを含むデータ (正例) を多数収集することは困難であ

る。これに対して、そのイベントを含まないデータ (負例) は比較的簡便に収集することができる。このため、多数の負例と少数の正例からなる学習データが生成される傾向にある。このような学習データに基づいて帰納学習を行った場合、帰納学習法は負例に偏った知識を発見する傾向にある。本問題は、帰納学習分野において、imbalanced 問題 [4] として知られた問題であり、与えられている正例の重みを高くしたり、学習データ間の距離に基づいて冗長な負例を削減 [4] したりする方法等が提案されている。しかしながら、これらの手法は、必ずしもテキストデータを想定した手法にはなっていないため、テキストデータの性質を有効に活用することはできない。このため、テキストデータの性質を利用することにより、テキストデータにおける事例の偏りを精度よく補正することが期待できる。

そこで、本論文では、テキストデータを対象とした imbalanced 問題の新たな解決策を提案し、英語掲示板サイトから収集した記事に適用し、その効果を検証する。以下においては、2 節で、テキストデータからのイベント抽出における問題点を指摘するとともに、機械学習研究の分野で提案されている従来の事例の偏りの補正法を紹介する。また、3 節では、テキストデータの特徴を加味した事例の偏りの補正法を提案し、4 節で、提案法を英語掲示板サイトから収集した記事データに適用した実験について説明する。最後に、5 節でまとめと今後の課題について述べる。

2. イベントの抽出

テキストデータに記述されている主体、行動、感情をイベントとして、テキストデータから抽出することを考える。このうち、主体に対応するイベントは、ある程度対象とするテキストを限定することにより、その主体の種類を制限することができる。また、このような主体は、表現のバリエーションが比較的少ないため、予めそのバリエーションを辞書に登録しておくことができる。これに対して、行動、感情に対応するイベントは、対象とするテキストを限定することにより、そのイベントの種類を限定することはできるものの、表現のバリエーションが非常に多いばかりか、複雑な構造を持って記述されることもある。このため、すべてのバリエーションを矛盾無く辞書に登録しておくことはできない。従って、行動、感情といったイベントに対しては、不完全な辞書を想定したイベント抽出を行う必要がある。

一方、テキストデータに対して、イベントの有無を指定した教師データが与えられるとするならば、テキストデータと教師データを組み (学習データ) にすることにより、イベントの有無を予測する分類モデルを学習することができる。このとき、学習データをいかにして集めるかが問題となる。Web などの普及に伴って、テキストデータそのものを収集することは容易になっているものの、特定のイベントを含むテキストデータを収集することはそれ程簡単ではない。このため、イベントを含むテキストデータ (正例) は、イベントを含まないテキストデータ (負例) よりも、通常、その数が非常に少なくなっている。このような負例に偏った学習データの場合、学習される分類モデルは負例に偏ったクラスを判別する傾向にあり、本問題を解決する手法が従来から研究されている。

そのひとつである正例への重み付け法は、負例に対する重みを 1 とする一方、正例に対して 1 を超える重みを与える方法である。各正例の重みを各負例の重みよりも大きくすることにより、見かけ上、正例が多くなるため、負例に偏っていた分類モデルの偏りをある程度解消することができる。しかしながら、本手法の場合、その重みを決定する明示的な指標が必ずしも与えられていないため、試行錯誤を通じて、その値を調整する必要がある。特に、テキストデータの場合には、学習データを特徴付ける次元は非常に高くなるため、分類モデルを学習するには多くの時間が必要であり、試行錯誤を通したパラメータ調整は困難であった。また、学習法によっては同一の学習データを事前に集約する機能を持っているため、その重みが必ずしも学習結果に反映されないという危険性があった。

一方、別の手法として、距離に基づいた負例の削減法 [4] も提案されている。本手法は、図 1 のアルゴリズムに従って、正例と負例が Tomek links になるかどうかを判定することにより、Tomek links となる事例を冗長な負例として学習データから削除する。ただし、異なるクラスを割り当てられたふたつの学習データ x、y において、距離関数 d が定義されており、$d(x,z) < d(x,y)$ あるいは $d(y,z) < d(y,x)$ となるような z が存在しない場合に、x、y に Tomek links の関係があると定義する。このような Tomek links となる事例間の場合、その事例間よりも近い位置に他の事例をひとつも含んでいないため、当該事例間を分割するような境界を引く必要がない。このため、Tomek links となる負例を冗長な事例として削除することができる。なお、本論文では、ふたつの学習データ x、y 間の距離を式 (1) によって定義する。

$$d(x,y) = \sqrt{\sum_i dist(x_i, y_i)} \tag{1}$$

ここで、x_i、y_i は x、y を構成する i 番目の要素であるとし、$dist()$ は、x_i 及び y_i が離散値の場合に式 (2)、x_i 及び y_i が数値の場合に式 (3)、によって定義されるとする。

$$dist(x_i, y_i) = \begin{cases} 1, & x_i = y_i \\ 0, & x_i \neq y_i \end{cases} \tag{2}$$

$$dist(x_i, y_i) = (x_i - y_i)^2 \tag{3}$$

1. S を学習に用いる学習データ全体とする。
2. 全正例とランダムに選択したひとつの負例からなる学習データの集合を C とする。
3. C を利用した 1-Nearest Neighbor 法で S を分類し、誤分類した負例を C に追加する。
4. C から Tomek links となる負例を削除する。

図 1　距離に基づいた負例削減アルゴリズム

このような負例の削減を行うことにより、正例の割合が相対

的に大きくなるため、imbalanced 問題をある程度解消することができる。しかしながら、テキストデータを対象とした場合には、学習データを特徴付ける次元は非常に高くなるため、学習データ間の距離を計算するのに多大なる時間が必要である。また、テキストデータの場合には、スパースでノイズの多い特徴付けが通常なされるため、学習データ間の距離をそれ程正確に計測することはできない。このため、学習データ間の距離を利用して冗長な負例を判定したとしても、必ずしも妥当な負例を削減できないという問題があった。

このように、テキストデータの場合、従来法では、必ずしも imbalanced 問題を適切に扱うことができなかった。そこで、テキストデータの特徴を反映した事例の偏りの補正法を次節で提案する。

3. イベント関連表現辞書に基づいた事例の偏りの補正

属性ベクトルによって特徴付けられたテキストデータからイベントを抽出する簡便な方法として、イベントに関連する表現を辞書 (イベント関連表現辞書) に登録して、本表現に合致する表現がテキストに含まれている場合に、テキストにイベントが含まれていると判定する方法が考えられる。このようなイベント関連表現辞書をすべてのイベントを正しく抽出できるように作成することは難しいものの、どのような表現がイベントと関連しているかは、ある程度想定することはできる。このため、ある程度妥当なイベント関連表現辞書を作成できると考えられる。しかしながら、イベント関連表現辞書を用いた方法では、複雑な構造を持った表現の場合に、本来イベントを含んでいないテキストデータをイベントを含んでいると誤って判定する危険性がある。例えば、図 2 に示すように、「不満」といった表現がイベント関連表現辞書に登録されているとし、テキストデータに不満が記述されているかどうかをイベントとして取り出す場合を考えてみることにする。このとき、「不満がある」といった表現がテキストデータに含まれているとすれば、イベント関連表現辞書は、当該テキストデータがイベントを含んでいると正しく判定することができる。一方、「不満がない」といった表現がテキストデータに含まれているとすれば、イベント関連表現辞書は、当該テキストデータに対してもイベントを含んでいると判定し、誤った判定をすることになる。このような判定がなされるのは、「不満」という表現だけで、イベントの有無を判定しているためであり、「ある」、「ない」といった表現も判定の際に評価することにすれば、両者の表現を正しく判定することができる。このように、本来はイベントを含んでいないテキストデータをイベントを含んでいると判定する場合、本負例には、イベント関連表現辞書では想定していなかった複雑な構造を持った表現が含まれていると考えられる。そこで、このような負例を、正例とともに学習データとして利用することにより、複雑な構造を持った表現を判定する分類モデルを学習することが期待できる。

ここで、イベント関連表現辞書には、イベントに関連するすべての表現が予め登録されていないことに注意する必要がある。このため、本来はイベントを含んでいるテキストデータであったとしても、イベントを含んでいないと誤って判定する危険性がある。例えば、テキストデータに「対応が悪い」といった表現が含まれているとし、イベント関連表現辞書に「対応が悪い」といった表現が登録されていないとする。このとき、イベント関連表現辞書では、「対応が悪い」といった表現を不満に関連した表現として抽出することができないため、当該テキストデータには、イベントが含まれていないと判定されることになる。ここで、当該テキストデータが正例として与えられているとすれば、「対応が悪い」といった表現を含むテキストデータを、イベントを含むテキストデータとして抽出する分類モデルを学習器は学習することができる。しかしながら、当該テキストデータを正例として与えただけでは、分類モデルにおいてどこまでを正例の範囲として良いかが不明確である。例えば、図 3 に示すように、「悪い」、「対応」、「ない」といった表現によって構成される空間に事例が配置されているとする。このとき、「対応が悪い」といった正例だけでは、「ない」の軸方向に関しては、情報が全く与えられていないため、「ない」の軸方向に関しては、任意の位置を通る分類モデルを学習することができる。このため、「対応」、「悪い」といった表現が一致しており、空間上において近接していると考えられる「対応が悪くはない」といった表現を含むテキストデータに関して、分類器は、その下を通るような分類モデルを学習することもできるし、その上を通るような分類モデルを学習することもできる。従って、当該テキストデータを、正例として判定するべきか負例として判定するべきかは、「対応が悪い」を含むテキストデータだけでは判定することができない。このとき、「対応が悪くはない」といった表現を含むテキストデータが負例として与えられているとすれば、分類器は「対応が悪い」と「対応が悪くはない」とを分割する分類モデルを学習することができる。このため、イベント関連表現辞書によって誤って負例と判定された正例が、影響を与える範囲を規定する負例を導入することが必要である。従って、このような当該正例の近くにある負例も学習データとして利用することにより、イベント関連表現辞書に含まれていない表現にも対応した分類モデルを学習することが期待できる。

学習データには多くの負例が存在するため、以上に検討した 2 種類の負例以外にも、多数の負例が存在すると考えられる。しかしながら、残りの負例は、正例と負例の境界の決定には、あまり寄与していないと考えられる。一方、このような負例を学習データとして残した場合、負例の数の力によって、本来は

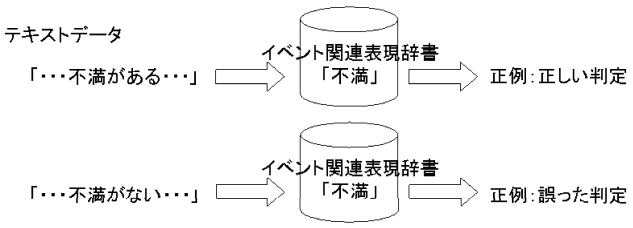

図 2　イベント関連表現辞書による誤った評価

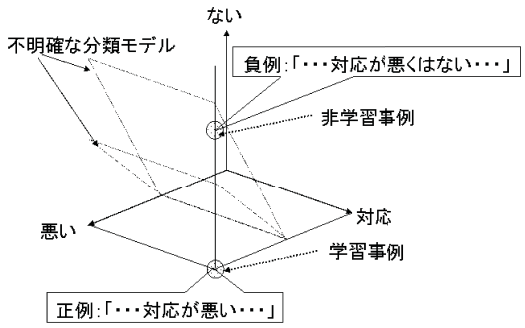

図 3　誤った分類モデルの学習可能性

重視すべきではない表現が誤って重視される危険性が増大する。このため、負例に過度に偏った分類モデルを学習する危険性が高くなる。そこで、上記の 2 種類以外の負例を冗長な負例として削除することにより、事例の偏りを補正し、2 種類の負例と正例に基づいて分類モデルを学習することにする。すなわち、提案法では、図 4 に概略を示すように、イベント関連表現に基づいた評価を行うことにより、学習データを 4 種類の事例集合に分割する。また、教師が負例と判定し、評価も負例と判定した事例の中から、教師が正例と判定し、評価が負例と判定した事例からの距離が遠い事例を、冗長な事例として削除する。これにより、事例の偏りを解消する。

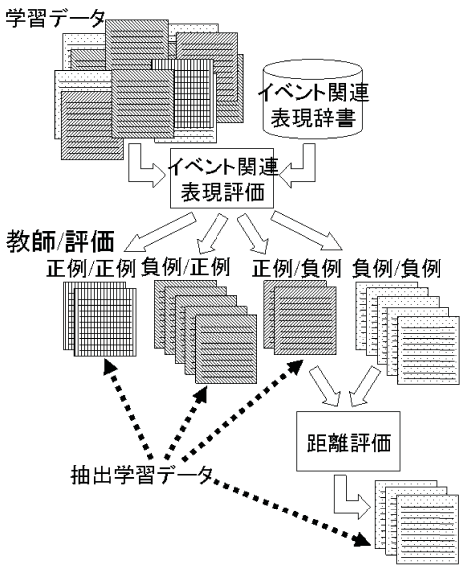

図 4　イベント関連表現辞書に基づいた事例の偏りの補正法の概略

本事例の偏りの補正法のアルゴリズムは、図 5 に示す C 言語ライクな擬似コードによって記述することができる。図 5 においては、学習データ S、イベント関連表現辞書 D、負例と評価された正例の集合との近さを示す値の最小値 Th を入力として与えることにより、事例の偏りを補正した学習データ C が出力される。また、本図においては、$flag[]$ がイベント関連表現辞書によって学習データが評価された場合に与えられる正例 ($Positive$) あるいは負例 ($Negative$) を識別する値を格納する領域、$nearness[]$ が負例と負例と評価された正例の集合との近さを示す値 ($\in [0,1]$) を格納する領域を表しているとする。加えて、$getTrain()$ が指定された学習データからひとつの事例を取り出す関数、$evalDic()$ が指定された事例をイベント関連表現辞書に適用することにより、正例あるいは負例かを評価する関数、$getClass()$ が指定された事例の教師によって与えられた正例あるいは負例の値を取り出す関数、$calcSim()$ が指定された事例間の近さを計算する関数、$\oplus$ が指定されたふたつの事例間の近さに対して実施する和演算を表しているとする。この他、P が教師が正例と判定した学習データ、N が教師が負例と判定した学習データ、N_n が教師が負例と判定し、評価も負例と判定した学習データ、S_t が S の一時変数、P_t が P の一時変数、N_t が N の一時変数、N_{nt} が N_n の一時変数、x 及び y が事例を格納する一時変数、cl が事例の教師による評価を格納する一時変数、sim が事例間の近さを示す値を格納する一時変数を表しているとする。

4. 数 値 実 験

本節では、提案法の効果を検証するために実施した、数値実験の概要を説明し、実験結果を考察する。

4.1 実験データ

6 つの英語掲示板サイトから 4 度の試みによって 14,860 件の記事を収集する。このうち、第 1 の試みによって収集した 2,240 件の記事 (D_1) 及び第 2 の試みによって収集した 4,067 件の記事 (D_2) は、コンピュータ関連の内容を対象とした記事である。また、各記事においては、不満イベント、会社イベントの集合である会社イベントクラスに含まれるイベント、機器イベントの集合である機器イベントクラスに含まれるイベントの有無が判定されている。一方、第 3 及び第 4 の試みによって収集した 8,553 件の記事 (D_3) は、コンピュータ関連の製品を扱う会社に関連した内容を対象とした記事であり、不満イベントの有無だけが判定されている。D_3 においては、会社を特定した記事収集となっているため、コンピュータ関連以外の記事も含まれていることに注意する必要がある。

不満記事を含む学習データを正例とした場合、D_1 においては、201 件が正例となっており、正例の割合は 8.97%である。一方、D_3 においては、不満イベントを含む記事をできるだけ収集することを目的として記事収集を行ったため、3,814 件が正例となっており、正例の割合が 44.8%と高くなっている。我々の最終目的としては、事例の偏りの有無に関わらず、分類性能の高い分類モデルを学習することであるため、学習データを最大限利用した、すべての記事 ($D_{all} = D_1 + D_2 + D_3$) を用いた評価実験 (不満率は 29.0%) を行う。一方、提案法の効果を検証する上では、事例の偏りの大きな学習データを利用した方がよいため、D_1 を利用した評価実験も行うことにする。

4.2 評 価 基 準

本論文では、イベントの抽出性能及び計算量を基準として、提案法の効果を検証する。イベントの抽出性能を示す評価基準としては、式 (4)～ 式 (6) に定義される、再現率、適合率、g-測度を採用する。特に、g-測度は、正例と負例のイベント抽出性能のバランスを取ったイベント抽出性能に関する総合的な指標であり、imbalanced 問題の解消の効果を検証する上では、重要

```
//辞書評価
St = S;
while((x = getTrain(St))! = NULL){
  //辞書評価結果の設定
  flag[x] = evalDic(x, D);
  St = St − x;
}
//正例・負例設定
St = S; P = φ; N = φ;
while((x = getTrain(St))! = NULL){
  //事例のクラスの判定
  if((cl = getClass(x)) == Positive) P = P + x;
  else N = N + x;
  St = St − x;
}
C = P;
//誤分類負例設定
Nt = N; Nn = φ;
while((y = getTrain(Nt))! = NULL){
  //辞書評価結果による事例の分類
  if(flag[y] == Positive) C = C + y;
  else Nn = Nn + y;
  Nt = Nt − y;
}
//誤分類正例対応負例距離計算
Nnt = Nn;
while((y = getTrain(Nnt))! = NULL){
  //近さの初期設定
  nearness[y] = 0;
  Nnt = Nnt − y;
}
Pt = P;
while((x = getTrain(Pt))! = NULL){
  if(flag[x] == Negative){
    Nnt = Nn;
    while((y = getTrain(Nnt))! = NULL){
      sim = calcSim(x, y);
      //近さの再計算
      nearness[y] = nearness[y] ⊕ sim;
      Nnt = Nnt − y;
    }
  }
  Pt = Pt − x;
}
//誤分類正例対応負例設定
Nnt = Nn;
while((y = getTrain(Nnt))! = NULL){
  //近さによる負例の判別
  if(nearness[y] ≧ Th) C = C + y;
  Nnt = Nnt − y;
}
```

図 5　イベント関連表現辞書に基づいた事例の偏りの補正アルゴリズム

な指標になっている。このため、イベント抽出性能の評価においては g-測度を中心とした評価を行う。

$$\text{再現率} = p_{rec} = \frac{a}{a+b} \tag{4}$$

$$\text{適合率} = p_{pre} = \frac{a}{a+c} \tag{5}$$

$$\text{g-測度} = p_{g\text{-}per} = \sqrt{\frac{a}{a+b} \times \frac{d}{c+d}} \tag{6}$$

ただし、a、b、c、d は表 1 の各セルの値に対応する記事の件数とする。

表 1　学習データ数間の関係

		評価結果	
		イベントあり	イベントなし
教師による判定	イベントあり	a	b
	イベントなし	c	d

一方、計算量に関しては、事例の偏りを補正することによって削減される負例の割合を中心として評価するとともに、事例間の距離計算を行う回数に基づいて評価する。

4.3　属性ベクトルの生成

記事とその正例あるいは負例を判定した教師データを組にした学習データは、属性の構造化が行われていないため、このままでは、分類器を用いて分類モデルを学習することができない。このため、テキストデータからその特徴を取り出して属性の構造化を行う必要がある。今回の実験では、記事を構成するテキストデータに対して、形態素解析 [5] [9] を適用することにより、テキストデータを語幹の列に変換する。また、全記事から抽出される各語幹に対して、tf-idf 値 [15] を式 (7) によって計算する。

$$\text{tf-idf}_i = \frac{1}{D} \cdot \log_2 \left(\frac{D}{d_i}\right) \cdot \sum_j \frac{\log_2 (t_{ij} + 1)}{\log_2 w_j} \tag{7}$$

ここで、D は記事の総数、d_i は i 番目の語幹をもつ記事の数、w_j が j 番目の記事に含まれる語幹の数、t_{ij} が j 番目の記事に含まれる i 番目の語幹の数を示しており、各記事は少なくともふたつの語幹によって構成されているとする。本式を用いることにより、多くの文章に頻繁に現れる語幹の影響を取り除いた、記事を特徴付ける語幹を抽出することができる。

計算した tf-idf 値を基準とすることにより、指定したしきい値以上となる語幹を記事を特徴付ける語幹として抽出する。各記事においては、抽出された語幹を記事が含んでいれば 1、含んでいなければ 0 を割り当てることにより、属性ベクトルを生成する。例えば、記事を特徴付ける語幹として、$a_1 \sim a_3$ が抽出されており、記事に語幹 a_1 及び a_2 が含まれている場合、当該テキストに対しては、(1,1,0) といった属性ベクトルが生成される。ただし、本属性ベクトルの表現においては、i 番目の語幹の有無が i 番目の属性値の値に対応している。

なお、今回の実験においては、D_1 の場合に、4,244 個の語幹によって属性ベクトルが構成されており、D_{all} の場合に、1,913 個の語幹によって属性ベクトルが構成されている。

4.4 実験方法

記事の中に不満イベントが存在するかどうかを識別する分類モデルは、記事を不満イベントの有無によって分類することと等価である。また、記事の分類はテキスト分類問題のひとつと考えることができる。テキスト分類問題においては、学習器として、SVM(Support Vector Machine) [16] を採用した場合に、分類性能が高くなることがいくつかの論文によって指摘されている [8] [10]。そこで、本実験では、学習器としては、SVM を利用することにする。具体的には、参考文献 [2] からダウンロードできる SVM ソフトウェア libsvm を利用する。本ソフトウェアにおいては、特徴空間上でのふたつのパターンの内積を表す関数であるカーネル関数として、線形カーネル、多項式カーネル、回転対象基底関数カーネル (RBF カーネル)、シグモイドカーネルを提供している。参考文献 [2] では、RBF カーネルの利用を推奨しており、経験則に基づいた RBF カーネルのパラメータの調整法を紹介している。しかしながら、RBF カーネルのパラメータ調整を実施するには、後に示すように多くの組み合わせを調査する必要があり、提案法を複数の手法と比較する場合には、実験に時間がかかるといった問題があった。ここで、手法の比較に主眼を置くことを考えてみると、同一の条件で実験を実施するかぎり、選択するカーネルやパラメータの比較結果に対する影響はそれ程大きくないと考えられる。このため、パラメータ調整を実施することなしに、比較実験を行うことにする。ただし、予備実験の結果、RBF カーネルのデフォルトパラメータに基づいた分類モデル学習では、イベント抽出性能が著しく低くなる一方、線形カーネルのデフォルトパラメータに基づいた分類モデル学習では、比較的高いイベント抽出性能を得ることができた。このため、D_1 に基づいた提案法の従来法との比較実験である第 1 の実験では、線形カーネルのデフォルトパラメータに基づいた分類モデル学習を実施することにする。このとき、第 1 の実験では、他手法として、事例の偏りの補正を行わない方法 (origin)、正例の重みを 2(weight2) 及び 10(weight10) とする方法に加えて、Tomek links に基づいた方法 (dist)、誤分類正例に対応した負例の追加を実施しない提案法 (non_ev_dist) を利用する。また、イベント関連表現辞書としては、135 個のイベント関連表現を登録した辞書を利用する。

次に、分類性能を高くすることを意図しつつ、イベント関連表現辞書に含まれる表現の違いの影響を見るために、D_{all} を用いて、イベント抽出性能を比較する第 2 の実験を実施する。比較対象としては、事例の偏りを補正しない方法 (origin) と誤分類正例に対応した負例の追加を実施しない提案法 (non_ev_dist) を利用する。第 2 の実験では、最終的な分類性能を高くすることも念頭においているため、参考文献 [2] の推奨に従って、RBF カーネルを、パラメータ調整を実施しつつ利用する。すなわち、RBF カーネルにおける 2 種のパラメータ C 及び γ を、$C = 2^{-5}, 2^{-3}, \cdots, 2^{15}$ 及び $\gamma = 2^{-15}, 2^{-13}, \cdots, = 2^{3}$ というように変化させ、g-測度が最大となる実験結果を 110(= 11 × 10) 通りの実験結果の中から選択する。ただし、C は判別に失敗する記事をどの程度許容するかを調整するためのパラメータであり、γ は不満イベントを含む記事と含まない記事とを分離する超平面と記事までの最小距離 (マージン) に関するパラメータである。また、利用するイベント関連表現辞書としては、表 2 に示す個数のイベント関連表現を登録した辞書を利用することにする。ただし、本辞書においては、次の関係 $dic1 \subseteqq dic2 \subseteqq dic3 \subseteqq dic4$ が成立している。

表 2　イベント関連表現の数

イベント関連表現辞書	dic1	dic2	dic3	dic4
イベント関連表現数	467	491	556	578

なお、本論文で実施する実験においては、提案法の基本的な効果を検証するために、$calcSim()$ における事例間の近さを、最もシンプルな式 (8) によって計測する。

$$nearness(x, y) = \begin{cases} 1, & y = y_{min}(x) \\ 0, & y \neq y_{min}(x) \end{cases} \quad (8)$$

ただし、$y_{min}(x) = \{y' | min_{y' \in N}(d(x, y'))\}$ とする。また、$\oplus$ におけるふたつの事例間の近さの和演算として、max 演算を適用することにし、誤分類正例に対応する負例として、学習データに負例を加えるためのしきい値 Th に 1 を設定する。

このような設定で負例を選択することにより、誤分類された正例の最も近くに、イベント関連表現辞書の評価によって負例と判定された負例がある場合に、当該負例を重要な負例として、学習データに加えることができる。

4.5 実験結果

第 1 の実験におけるイベント抽出性能を比較した実験結果を図 6 に示す。図 6 においては、事例の偏りの補正を行わない場合の各評価値を 1 とした場合における改善率を示している。また、x 軸が手法、y 軸が改善率、各折れ線がイベント抽出性能の種類を示している。

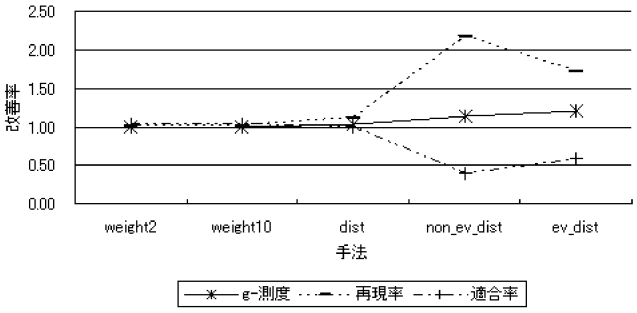

図 6　イベント抽出性能の改善率

次に、事例の偏りの補正を行った場合の負例の削減率を表 3 に示す。ただし、事例の重み付け法の場合には、負例は削減されないため、実験結果は割愛されている。また、各実験結果は、事例の偏りの補正を行わない場合の負例数を 1 とした場合の削減率を示している。

表 3　負例の削減率

手法	dist	non_ev_dist	ev_dist
削減率	0.027	0.861	0.822

加えて、図 7～ 図 9 に、第 2 の実験におけるイベント抽出性能を示す。すなわち、図 7、図 8、図 9 が、それぞれ、イベント関連表現辞書の変化に対する g-測度、再現率、適合率の変化を示している。各図においては、x 軸がイベント関連表現辞書を示しており、y 軸が各イベント抽出性能を示している。また、各折れ線が手法を示している。

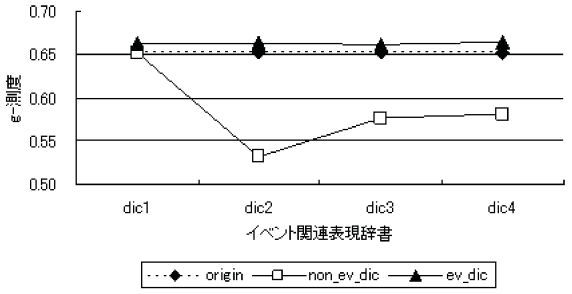

図 7　g-測　　度

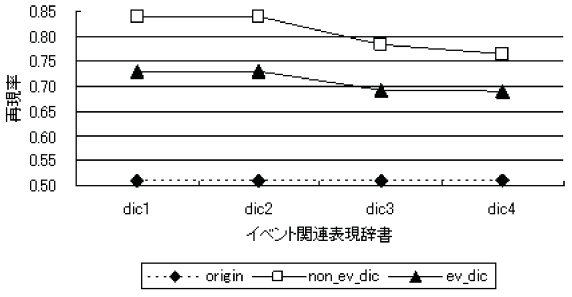

図 8　再　現　率

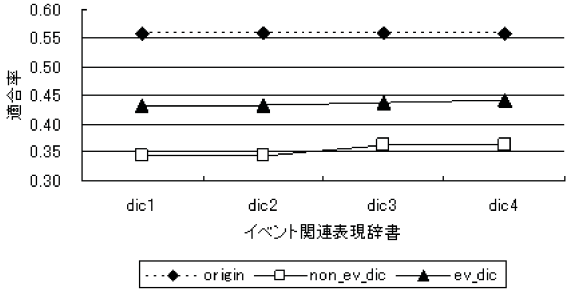

図 9　適　合　率

4.6　考　　察

イベント抽出性能: 図 6 の g-測度の実験結果に着目してみると、事例の調整を行わない方法、正例の重み付け法、Tomek links に基づいた方法、誤分類正例に対応した負例の追加を実施しない提案法、提案法の順に g-測度の値が上昇している。特に、事例の調整を行わない方法との比較においては、19.9% 程 g-測度が改善されており、事例の偏りを補正した効果が得られている。一方、従来の事例の偏りを補正する方法では、Tomek links に基づいた方法でも、その改善率は 3.6% 程度であり、提案法に比べると改善の度合いは小さくなっている。この実験結果は、当初予想したように、高次元の空間に点在するテキスト分類問題においては、必ずしも距離評価を精度よく行うことができず、冗長な負例を適切に判定できないという現象が、発生したことによるものと考えられる。

一方、再現率と適合率に注目してみると、提案法においては、再現率が大幅に改善される反面、適合率がかなり改悪されている。特に、誤分類正例に対応した負例の追加を実施しない提案法の場合には、この傾向が顕著に現れている。提案法の場合、重要な負例以外の負例を冗長な負例と判定しているため、多くの負例を冗長な負例と判定しがちである。このため、この傾向は、正例を重視した分類モデルが学習された結果と考えられる。これに対して、Tomek links に基づいた方法では、冗長な負例に焦点をあてて、事例の偏りを補正しているため、提案法に比べれば正例をそれ程重視してはいない。このため、適合率、再現率の大幅な変動は起こらずに、事例の偏りの若干の補正による小幅な改善がなされたものと考えられる。

今回の実験では、誤分類された正例集合に対応する負例の近さを、当該の各正例に対して最も近い負例の近さを 1 とし、その他の負例の近さを 0 と定義している。このため、重要な負例として負例が残ることに関しては、かなり厳しい定義となっている。従って、この近さの定義を若干緩和して、ある程度近い負例を誤分類する正例に対応する負例として許容することにすれば、正例の影響をある程度緩和することができ、よりバランスのとれた事例の偏りの補正を行うことが期待できる。このため、どのような近さの定義が、より妥当なイベント抽出性能を与えるかについては、今後検討していきたい。

計算速度: 表 3 の実験結果に着目してみると、提案法は、多くの負例を冗長な負例として削除している。一方、Tomek links に基づいた方法は、それ程多くの負例を削除しておらず、若干の負例を削除しているに過ぎない。ここで、Tomek links の判定において、ふたつの学習データと、他の残りの学習データとの間の距離を計算する必要があることに着目してみると、多数の学習データ間の距離を評価する必要があることが分かる。このため、Tomek links に基づいた方法では、学習データ間の距離評価がより重要となる。しかしながら、テキストデータの場合、学習データ間の距離をそれ程正確には計測できないため、冗長な負例の判定が不正確になる傾向にある。このため、本来は冗長な負例と判定されるべき負例と、その負例に対応する正例よりも近くに、何らかの事例が誤って存在する危険性も高くなる。従って、Tomek links に基づいた方法において削除される負例の数が少なくなったものと考えられる。これに対して、提案法の場合には、イベント表現辞書によって負例と判定された正例の近くに存在する負例との間に対してだけ、その距離を考えている。このため、不正確な事例間の距離による影響が少なく、削減できる負例が多くなったものと考えられる。

また、学習データ間の距離計算の回数に着目してみれば、提案法は、距離計算を一部の学習データ間に対してだけ実施すればよく、距離計算回数を大幅に少なくすることができる。Tomek links に基づいた方法及び提案法は、互いに独自の処理を行う部分が存在するため、その計算量を正確に比較することはできないものの、提案法は、Tomek links の場合に比べて、対象とする

負例の数を大幅に少なくしている。加えて、距離計算回数も大幅に少なくしている。このため、提案法は、Tomek links に基づいた方法に比べて、計算速度を大幅に改善することが期待できる。実際、実験のために開発したシステムでは、提案法は数十倍程度も高速に計算を行うことができており、本効果を検証することができる。

一方、先のイベント性能の考察においても指摘したが、表 3 の実験結果に示すように、誤分類された正例集合に対応する負例の追加を行ったとしても、削減される負例の数はそれ程少なくなってはいない。このため、本追加操作では、それ程多くの負例が追加されないことが分かる。この点からも、本操作にともなう近さの定義を若干緩和して、追加される負例を若干拡充したとしても、計算速度にそれ程大きな問題は発生しないと考えられる。

イベント関連表現辞書の影響: 図 7 の実験結果に示すように、イベント関連表現が充実するに従って、g-測度の値は改善されている。この実験結果は、より多くのイベント関連表現を利用することによって、正例と負例の境界を決定するのに利用される負例を、より多く残す効果が得られたため、正例と負例の境界を精度よく学習できた結果と考えられる。ただし、dic1、dic2 の場合における g-測度の値に着目してみると、その値は事例の偏りを補正しない方法と、同定度の値になっていることに注意する必要がある。第 2 の実験に用いた D_{all} は事例の偏りがあまり大きくないため、提案法の効果をあまり得ることができない。このため、g-測度の値としては同定度の値になったものと考えられる。従って、このような状況下において、イベント関連表現の個数がより少なくなった場合には、事例の偏りを補正しない場合よりもイベント抽出性能が劣化する危険性がある。イベント関連表現辞書によるイベント抽出性能が低いと予想される場合には、提案法を利用しない方がよいものと考えられる。

一方、再現率と適合率に着目してみると、図 8、図 9 の実験結果に示すように、イベント関連表現が充実するのにともなって、再現率が低下する一方、適合率は上昇している。この現象は、イベント関連表現が充実するのにともなって、重要な負例と判定される負例が多くなり、事例の偏りが次第に負例方向に変化することによって、生じたものと考えられる。イベント抽出の目的によっては、これら指標のいずれか一方、または両方を重視する必要もあるので、本現象をイベント関連表現辞書によってどの程度まで偏りを補正すべきかの参考とする必要がある。

以上の議論に基づいて、提案するイベント関連表現に基づいた事例の偏りの補正法は、事例の偏りがあるテキストデータに対して、従来法よりも妥当な分類モデルを高速に学習できると考えられる。特に、イベント関連表現辞書に格納されるイベント関連表現が充実している場合に、イベント抽出性能をより高めることができると考えられる。

5. まとめと今後の課題

今回の論文では、イベント関連表現辞書に基づいた事例の偏りの補正法を提案し、提案法の効果を、6 つの英語掲示板サイトから収集した記事に適用し、その効果を検証した。これにより、事例の偏りが大きい場合に、提案法が従来の事例の偏りの補正法よりも効果的に事例の偏りを補正することを検証することができた。また、利用するイベント関連表現のイベント抽出性能が高い場合に、提案法がより効果的であることを検証することができた。

今後の課題としては、現在のところ、特定の記事を用いて提案法の効果を検証しているに過ぎないので、より多くの他のテキストデータに適用して、提案法の効果を検証していく必要がある。一方、テキストデータからのイベント抽出性能の向上という観点では、形態素解析よりもより多くの情報を与えることができる構文解析を利用することにより、イベント抽出性能が高くなることを期待できる。このため、構文解析を利用したイベント抽出法も検討していきたい。

文　献

[1] R. Feldman, I. Dagan and H. Hirsh: “Mining Text Using Keyword Distributions”, J. of Intelligent Information Systems, 10, 3, 281-300 (1998).

[2] C. -W. Hsu, C. -C. Chang, and C. -J. Lin: “A Practical Guide to Support Vector Classification”, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[3] 市村 由美, 鈴木 優, 酢山 明弘, 折原 良平, 中山 康子: 「日報分析システムと分析用知識記述支援ツールの開発」, 信学論, J86-D-II, 2, 310-323 (2003).

[4] M. Kubat and S. Matwin: “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection”, Proc. of the 14th Int. Conf. on Machine Learning, 179-186 (1997).

[5] J. Kupiec: “Robust Part-of-speech Tagging using a Hidden Markov Model”, Computer Speech and Language, 6, 3, 225-242 (1992).

[6] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen: “Websom for Textual Data Mining”, J. of Artificial Intelligence Review, 13, 5/6, 335-364 (1999).

[7] B. Lent, R. Agrawal, and R. Srikant: “Discovering Trends in Text Databases”, Proc. of the 3rd Int. Conf. on Knowledge Discovery and Data Mining, 227-230 (1997).

[8] L. M. Manevitz and M. Yousef: “One-Class SVMs for Document Classification”, J. of Machine Learning Research, 2, 139-154 (2001).

[9] M. F. Porter: “An Algorithm for Suffix Stripping”, Program, 14, 3, 130-137 (1980).

[10] B. Raskutti, H. Ferrá, and A. Kowalczyk: “Combining Clustering and Co-training to Enhance Text Classification using Unlabelled Data”, Proc. of 8th Int. Conf. on Knowledge Discovery and Data Mining, 620-625 (2002).

[11] 櫻井 茂明, 市村 由美, 酢山 明弘, 折原 良平: 「テキストマイニングシステム向けの構造抽出ルールの自動学習」, 電学論 C, 122, 6, 1009-1015 (2002).

[12] 櫻井 茂明, 酢山 明弘: 「ファジィ帰納学習におけるキー概念集合を含む属性値の扱い」, 日本ファジィ学会誌, 14, 6, 640-647 (2002).

[13] 櫻井 茂明, 酢山 明弘: 「キーフレーズに基づいたテキストの分析」, 日本知能情報ファジィ学会誌, 17, 1, 52-59 (2005).

[14] S. Sakurai and A. Suyama: “An E-mail Analysis Method based on Text Mining Techniques”, Applied Soft Computing, 6, 1, 62-71 (2005).

[15] G. Salton and M. J. McGill: “Intoroduction to Modern Information Retrieval”, McGraw Hill Computer Science Series, (1983).

[16] V. N. Vapnik: “The Nature of Statistical Learning Theory”, Springer, (1995).