Vol.2014-DBS-160 No.28
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.28
2014/11/18

単語の共起関係に基づく機械学習による文書分類

福元 伸也[1,a)]　渕田 孝康[1,b)]

概要：近年，ビッグデータと呼ばれる大規模データから有益な情報を抽出しようとする試みが広く行われており，テキストデータの解析に関する多くの研究がなされている．本研究では，シソーラスの分類語彙表を用いて，単語の特徴ベクトルである共起行列を生成する手法を提案する．出現単語のみによる共起行列を，単語の意味を考慮した分類語に変換することにより，共起行列の次元数が増大するのを抑えることができ，単語の特徴ベクトルをより的確なベクトルとして表現できる．また，得られた共起行列から分類を行うための学習器には，アンサンブル学習の 1 つであるランダムフォレストと大規模データに対して高度な分析が可能な機械学習フレームワークである Jubatus を用いた．実験では，ニュース記事のカテゴリ分類を行い，複数の学習アルゴリズムについて検証した．

キーワード：文書分類，共起行列，シソーラス，機械学習

1. はじめに

近年，インターネットの普及やクラウド環境の充実により，膨大な量のデータを扱う機会が増大しており，また人々の情報検索に対する要求も高まり，テキストデータの解析に関する情報処理技術の研究が盛んに行われている．情報検索においては，テキストデータを，ある特徴に従いグループ分けするために，テキスト分類に関する研究も行われており，大量の文書データを，効率よく分類する手法も数多く提案されている [1], [2]．テキスト分類では，テキストに含まれる文章を構成する語の重みを特徴ベクトルとして表現し，文書ベクトル間の類似度を定義したのち，文書分類を行う．このため，テキストが文字データとして扱われており，語の意味が考慮されておらず，我々人間の感覚と異なってしまうことがある．

本研究では，特徴ベクトルの生成において，文書中に現れた出現単語とシソーラスの単語の意味属性を用いて，共起頻度による共起行列を生成する．ここでは，シソーラスに分類語彙表を用いる [3]．分類語彙表は，長い年月にわたり語を意味によって分類・整理した類義語集であるため，語の持つ意味をうまく反映させることが期待できる．

通常，文書内には，似たような意味を持つ複数の単語が存在する．共起とは，ある単語に隣接して別の単語が現れることを共起と言い，単語間の共起頻度を利用した共起行列を用いて文書分類を行うと，本来似ている意味の単語が，距離の離れた特徴ベクトルとして表現されてしまい，精度の低下が生じてしまう [4]．

本研究では，語を意味により分類したシソーラスである分類語彙表を用いることで，単語の持つ意味を考慮した共起行列を作成する．その共起行列を学習データとして，分類のための学習器に与える．学習器には，アンサンブル学習の 1 つであるランダムフォレストを利用し [5]，文書分類を行った．また，別の学習器による分類も試みた．学習器には，大規模データに対し，高度な分析が可能な機械学習フレームワークである Jubatus を用いた [6]．

Jubatus は，オンライン処理，分散並列処理などの特徴を持つ機械学習フレームワークで，大規模データのさまざまなデータ分析に優れた性能を示している．実験では，ニュース記事の分類を行い，ランダムフォレストと他のアンサンブル学習法との比較や，Jubatus における複数の学習アルゴリズムでの識別率の比較を行った．

2. 関連研究

テキスト解析は，大量の非構造的テキストデータが蓄積されていることを考えると非常に重要な技術である．これまで文書分類の研究として，単語の係り受け関係を用いて分類を行う研究や文書中に現れる語の共起関係を用いたものなどがある [7]．また，テキスト解析に関する多くの研究が，分類精度の向上にチャレンジしており，分類に関するさまざまな学習法を提案している．Wang らは，語の重

[1] 鹿児島大学大学院理工学研究科
Graduate School of Science and Engineering, Kagoshima University, 1-21-40, Korimoto, Kagoshima 890–0065, Japan
[a)] fukumoto@ibe.kagoshima-u.ac.jp
[b)] fuchida@ibe.kagoshima-u.ac.jp

Vol.2014-DBS-160 No.28
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.28
2014/11/18

要度の決定において，RageRank アルゴリズムを用いることが，分類に有効であることを示した [8]．また，単語の共起行列を作成するために，文書に現れた単語間の共起頻度を利用した手法がある．単に単語間の共起頻度では，意味的に近い単語であっても，別の共起頻度としてカウントされ，その特徴ベクトルが離れてしまう問題があった [9]．別所らは，単語間の共起頻度ではなく，単語とコーパスにおける単語に付随する意味属性との共起頻度を取る手法を提案し，共起ベクトルの質が向上することを示した [10].

3. 共起行列の作成

3.1 問題点

単語と単語の関係において，例えば，1 つの文中に現れた単語は，その単語と別の単語の位置が近く，意味的にも近い関係であろうという予測のもと，これらの単語は共起関係にあると言う．その共起関係に基づき，ある単語と別の単語の共起関係の頻度を成分にした行列が共起行列である．単語間の共起頻度を利用した共起行列では，対象となったすべての単語が含まれてしまうため，行列の大きさが巨大になってしまう．行列の次元数が大きくなると次のような問題がある．

- 次元数の増大に伴い，計算コストも増大する
- スパースな行列になる
- 本来，近い関係にあるべき特徴ベクトルが離れた状態になってしまう

そこで，単語行列を属性行列に変換する手法が提案されており，笠原らは，国語辞典を用いる手法を提案している [11].

3.2 単語による特徴ベクトル

全テキストデータに含まれる単語を w_i とし，N 個の単語が含まれているとすると，単語 w_i の特徴ベクトルは次のように表される．

$$w_i = (v_{i1}, v_{i2}, \cdots, v_{iN}) \tag{1}$$

ただし，v_{ij} は w_i における重みである．単語特徴ベクトル w_i を要素とした列ベクトルは，次のような行列で表される．

$$F_w = \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} = \begin{pmatrix} v_{11} & \cdots & v_{1N} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{NN} \end{pmatrix} \tag{2}$$

この単語特徴行列から属性行列を生成する．属性数（行列の列数）を m とすると，単語の属性ベクトル $\acute{w}_1, \cdots, \acute{w}_N$ および，その列ベクトルは次の行列で表される．

分類語
意味属性　図書

出現語＼共起語	---	a: 小説	b: 雑誌	---
---	---	---	---	---
A: 本棚	---	3	80	---
B: 書棚	---	73	5	---
---	---	---	---	---

図 1 共起行列

$$F_p = \begin{pmatrix} \acute{w}_1 \\ \vdots \\ \acute{w}_N \end{pmatrix} = \begin{pmatrix} \acute{v}_{11} & \cdots & \acute{v}_{1m} \\ \vdots & \ddots & \vdots \\ \acute{v}_{N1} & \cdots & \acute{v}_{Nm} \end{pmatrix} \tag{3}$$

ただし，$\acute{v}_{ij}$ は $\acute{w}_i$ における重みである．

3.3 分類語による共起行列

共起行列の作成において，単語間の共起頻度を利用した共起行列の生成では，似た意味の単語であるのに，単語間の共起ベクトルの距離が離れてしまう問題が指摘されている [10].

図 1 の表は，左端の列が記事中に現れた単語の例を表しており，上段の共起語は，同一文章中に現れた共起語を表している．また，表中の数字は，その出現頻度を表している．単語間の共起に基づいた共起行列の作成手法では，出現語 A と B が，意味の近い単語であったとして，その共起語 a と b が別々にカウントされる．（a と b も意味の近い単語）そうなると，A と B のベクトルは，意味の近い単語であるにもかかわらず離れてしまう．そこで，共起語に現れた a と b を同じ意味を表す 1 つの単語にまとめることが出来れば，A と B のベクトルは離れない．

具体例で見てみると，本棚と書棚は，意味の近い単語である．その共起行列を見たときに，本棚は雑誌の出現頻度が高く，書棚は小説の出現頻度が高いと，それぞれの特徴ベクトルは，離れた状態となる．これを小説と雑誌の分類語である「図書」にまとめることができれば，それぞれの特徴ベクトルの向きは離れずに済む．

本研究では，意味の似ている語をまとめると共起ベクトルの距離は近くなるという点を踏まえ，単語間の共起頻度を用いるのではなく，単語に付随する意味属性を利用する．

レコード ID 番号／見出し番号／レコード種別／
類／部門／中項目／分類項目／分類番号／
段落番号／小段落番号／語番号／見出し／
見出し本体／読み／逆読み

図 2 分類語彙表の項目

Vol.2014-DBS-160 No.28
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.28
2014/11/18

単語の意味属性には，単語を意味によって分類整理したシソーラスである分類語彙表を利用し分類語に適用する．分類語彙表を構成する項目は，**図 2** のようになっており，共起行列に用いる意味属性には，その中の「分類項目」を用いた．共起行列の 1 列目には，形態素解析の結果得られた単語のうち，名詞のみを取り出し入力し，数字の部分は，1 文中に共起する頻度をカウントした数が入った行列となっている．また，1 行目には，意味属性として分類語彙表の分類項目の語が入る [12].

このようにして得られた共起行列は，式 (3) に相当し，単語間の共起行列である式 (2) から式 (3) を導き出す作業は，次式で表される変換行列 K を求めることに等しい [13].

$$F_p = F_w K \tag{4}$$

ただし，K は，N 行 m 列の行列である．

4. 学習器

4.1 ランダムフォレストによる学習

アンサンブル学習は，弱学習器と呼ばれるあまり精度の高くない学習器を複数用いて，その結果を組み合わせて，精度向上をはかる機械学習法である．アンサンブル学習では，異なるサンプルから単純なモデルを複数生成し，それらを統合することにより，全体としての精度を実現するモデルを構築する．アンサンブル学習の中には，与えられたデータセットから，ブートストラップによって，複数の学習用データセットをサンプルとして生成し回帰・分類を行うバギング（Bagging）[14] やランダムフォレスト（Random Forest）[5] などの学習法がある．

ランダムフォレストは，複数の木（tree）によって構成される機械学習アルゴリズムである [5]．ここでの木は，決定木のことで，それぞれの決定木の性能はあまり高くなく，それらを複数組み合わせることにより，高い予測精度を持つ学習器となる．ランダムフォレストでは，決定木として二分決定木が主に用いられ，個々の決定木がアンサンブル学習における弱学習器となる．ランダムフォレストのアルゴリズムを以下に示す [15]．（**図 3**）

(1) 与えられたデータセットから n 個のブートストラップ・サンプル $B_1, B_2, \cdots, B_n$ を作成する．ただし，構築したモデルを評価するために 1/3 のデータを除いてサンプリングする．除いたデータを OOB (out-of-bag) データと呼ぶ．

(2) $B_k (k = 1, 2, \cdots, n)$ における M 個の変数の中から m 個の変数をランダムサンプリングする．M は，データセットの中の変数の数を表し，m は，$m = \sqrt{M}$ が多く用いられる．

(3) ブートストラップ・サンプル B_k の m 個の変数を用いて，未剪定の最大の決定木 T_k を生成する．

(4) n 個のブートストラップ・サンプル B_k の決定木 T_k について，OOB データを用いてテストを行い，推測誤差を求める．

(5) その結果を統合し，新たに分類器を構築する．分類の問題では多数決をとる．

本研究では，学習器としてランダムフォレストを用い，そのアルゴリズムとして上記を用いて文書分類を行う．

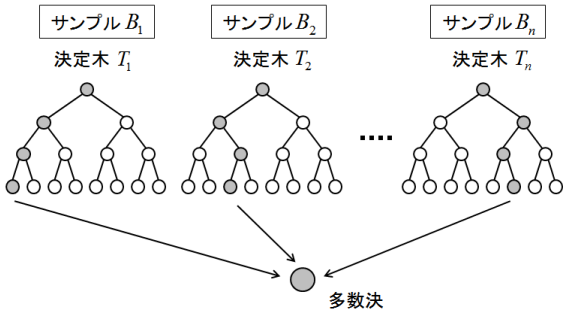

図 3 ランダムフォレスト

4.2 Jubatus による学習

分散並列処理システムがオープンソースで提供され，パソコンなどの安価なハードを用いることで，大規模データの分析が可能になりつつある．大規模データ処理基盤では，Hadoop[16] がオープンソースソフトウェアとして提供されているが，Hadoop は，データを蓄積してから解析するバッチ処理方式である．一方，ストリーム型のデータに対応可能な新たな機械学習フレームワークとして，Jubatus が提供されている [6]．Jubatus は，機械学習やデータマイニングによるデータ分析に特化した大規模データ処理基盤であり，ビッグデータ解析に必要なリアルタイム・ストリーム処理，分散並列処理，機械学習やマイニングなどの深い分析といった特徴を持つ [17].

本報告では，ランダムフォレストとは別の学習器として，Jubatus を用い，Jubatus の持つ複数の学習アルゴリズムによりテキスト分類を行う．文書分類に必要な操作は，多クラス分類であり，Jubatus は，線形識別器を用いて，これを実現している．Jubatus の多クラス分類において利用できる学習アルゴリズムには次のものがある．(i) Perceptron [18], (ii) Passive Aggressive (PA) [19], (iii) Confidence Weighted Learning (CW) [20], (iv) Adaptive Regularization Of Weight vectors (AROW) [21], (v) Normal Herd (NHERD) [22] である．(i) の Perceptron は，学習データが与えられたとき，現在の分類器で正しく分類できるかどうかを確かめ，正しく分類出来なかった場合は，重みを更新する．(ii) の PA は，学習データが正しく分類できたら，重みを変更しない．逆に正しく分類出来なかったら，分離面を更新する．(iii) の CW では，出現頻度を考慮し，重みベクトルにガウス分布を導入して更新する．(iv) の AROW は，CW と同様の手法を実現しつつ，複数の条件を同時に考慮しながら最適化を行う．(v) の NHERD は，特徴ベク

Vol.2014-DBS-160 No.28
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.28
2014/11/18

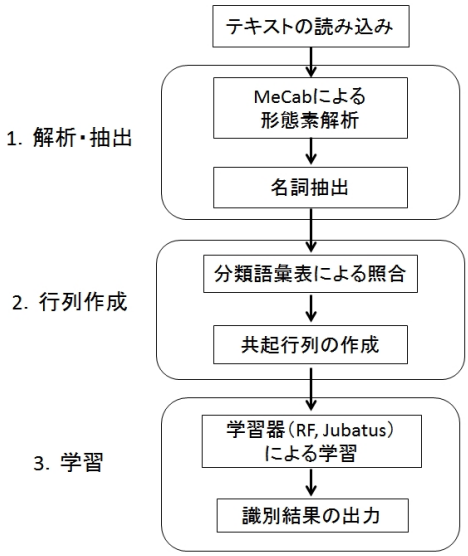

図 4 処理の流れ

トルが正規分布に従って生成されているモデルを利用し学習を行い，ベクトル間の共分散情報も利用することで効率的な学習を実現する．今回我々は，NHERD を除いた 4 つの学習アルゴリズムを用いて文書分類を試みる．

共起行列の生成とランダムフォレストおよび Jubatus の学習器による文書識別までの処理の流れを**図 4** に示す．

5. 実験

5.1 記事分類における識別率

実験では，ニュース記事のカテゴリ分類を行った．毎日新聞社のサイト [23] より記事を収集し，それを，政治，経済，社会，スポーツ，エンターテイメントの 5 つのカテゴリに分類する．分類に用いた記事の数は，1,000 である．収集した記事を，MeCab を用いて形態素解析し，その中から名詞の単語を共起行列作成のための出現単語として用いる．抽出された名詞の数は，重複を除いて 13,275 個であった．従来手法では，この出現語を用いて，共起行列を生成していた．ここでは，シソーラスの分類語彙表を用いて共起行列を生成した．そのため，共起行列の列の数は，507 個と大幅に削減された．ただし，単語の数も，分類語彙表に掲載されている単語となり，その数は，8,320 個となった．生成された共起行列を学習データとして学習器に与える．学習器には，RandomForest (RF) と，比較のための Bagging[14]，Support Vector Machine (SVM)[24] を使用した．また，RF に用いられた木の数は 500 である．実験の結果，得られた正識別率を表 1 に示す．その結果，RF での識別率が最も高かった．

つぎに，生成された共起行列と学習器に Jubatus を用いた実験を行った．Jubatus の学習アルゴリズムとして，Perceptron，Passive Aggressive (PA)，Confidence Weighted Learning (CW)，Adaptive Regularization Of Weight vectors (AROW) の 4 つの学習アルゴリズムを用いて文書識別を行った．各学習アルゴリズムごとの識別結果を表 2 に示す．4 つの学習アルゴリズムの中では，AROW による識別率が最も高かったが，RF と比べると，Jubatus による識別率は全体的に低い結果となった．

表 1 識別結果

	RF	Bagging	SVM
正識別率 (%)	88.7	79.5	86.7

表 2 学習アルゴリズムによる比較

	Perceptron	PA	CW	AROW
regularization weight	-	-	1.0	1.0
正識別率 (%)	68.4	75.6	77.8	78.4

5.2 変数の重要度

学習器に RF を用いたときに，変数の重要度について調べた．これは，分類語彙表を用いて抽出された分類項目が，どの程度重要であったかを意味する．その結果を**図 5** に示す．左の図は，決定木における正解率を基準とした平均値を示しており，また，右の図は，計算の過程で算出された Gini 係数の平均値を大きい順に並べてある．Gini 係数は，分岐の前と後での誤差の改善度合いを表し，判別における変数の重要度として用いられ，図の上位に表された項目ほど，重要度の高い語ということになる．これを見ると，「演劇・映画」，「スポーツ」，「経済・収支」といったニュース記事のカテゴリと一致するような項目が上位に現れていることがわかる．

6. おわりに

本研究では，分類語彙表を用いて単語の意味属性に基づき，共起行列を生成する方法について述べた．これにより，得られた共起行列を学習データとして学習器に与えた．学習器には，ランダムフォレストと Jubatus を使用した．Jubatus では，複数の学習アルゴリズムを用いて，文書分類を行った．その結果，ランダムフォレストに比べると Jubatus は，全体的に低い識別率となった．Jubatus の学習アルゴリズムの中では，AROW を用いた場合，最も高い識別率が得られた．

今後の課題として，Jubatus がランダムフォレストに比べ，低い識別結果となった原因の解明や他の学習アルゴリズムを使用した場合の識別などを行う予定である．

謝辞 本研究の一部は，JSPS 科研費 (24500120) の助成を受けて実施された．

参考文献

[1] R. M. Samer Hassan and C. Banea: “Random-walk term weighting for improved text classification”, Proc. of the First Workshop on Graph Based Methods for Natural

Vol.2014-DBS-160 No.28
Vol.2014-OS-131 No.2
Vol.2014-EMB-35 No.28
2014/11/18

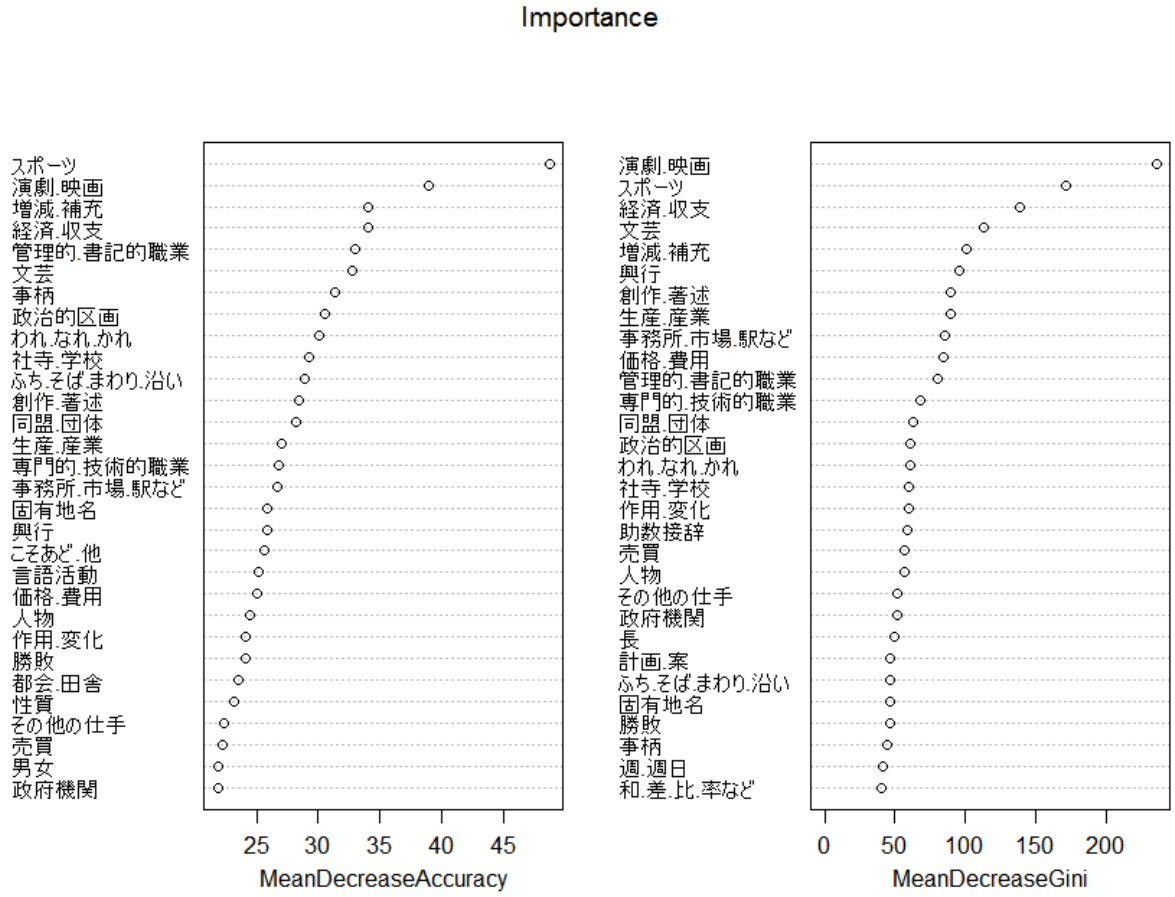

図 5 変数の重要度

Language Processing (2006).
[2] F. Sebastiani: “Machine learning in automated text categorization”, Proc. ACM Computing Surveys, **34**, 1, pp. 1–47 (2002).
[3] 国立国語研究所：“分類語彙表－増補改訂版”, 大日本図書刊 (2004).
[4] 有村博紀：“テキストマイニング基盤技術”, 人工知能誌, **16**, 2, pp. 201–211 (2001).
[5] L. Breiman: “Random forests”, Machine Learning, **45**, pp. 5–32 (2001).
[6] Jubatus, http://jubat.us/ja/.
[7] 渡部広一, 奥村紀之, 河岡司：“概念の意味属性と共起情報を用いた関連度計算方式”, 自然言語処理, **13**, 1, pp. 53–74 (2006).
[8] D. B. D. Wei Wang and X. Lin: “Term graph model for text classification”, Springer-Verlag Berlin Heidelberg 2005, pp. 19–30 (2005).
[9] 片岡 良治：“単語と意味属性との共起に基づく概念ベクトル生成手法”, 人工知能学会第 20 会全国大会論文集, **3C3-1**, pp. 1–3 (2006).
[10] 別所克人, 内山俊郎, 内山匡, 片岡良治, 奥雅博：“単語・意味属性間共起に基づくコーパス概念ベースの生成方式”, 情報処理学会論文誌, **49**, 12, pp. 3997–4006 (2008).
[11] 笠原要, 松澤和光, 石川勉：“国語辞書を利用した日常語の類似性判別”, 情報処理学会論文誌, **38**, 7, pp. 1272–1283 (1997).
[12] 尾脇拓朗, 福元伸也：“単語の意味を考慮した共起ベクトルによるテキスト分類”, DEIM Forum 2014, **C6-2**, (2014).
[13] 笠原要, 稲子希望, 加藤恒昭：“単語の属性空間の表現方法”, 人工知能学会論文誌, **17**, pp. 539–547 (2002).
[14] L. Breiman: “Bagging predictors”, Machine learning, **24**, 2, pp. 123–140 (1996).
[15] 金明哲：“統計的テキスト解析”, ESTRELA, 182 (2009).
[16] Hadoop, http://hadoop.apache.org/.
[17] 岡野原大輔：“大規模データ分析基盤 jubatus によるリアルタイム機械学習”, 人工知能学会誌, **28**, 1, pp. 98–103 (2013).
[18] F. Rosenblatt: “The perception: a probabilistic model for information storage and organization in the brain”, Neurocomputing: foundations of research MIT Press, pp. 89–114 (1988).
[19] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz and Y. Singer: “Online passive-aggressive algorithms”, The Journal of Machine Learning Research, **7**, pp. 551–585 (2006).
[20] M. Dredze, K. Crammer and F. Pereira: “Confidence-weighted linear classification”, Proceedings of the 25th international conference on Machine learningACM, pp. 264–271 (2008).
[21] K. Crammer, A. Kulesza and M. Dredze: “Adaptive regularization of weight vectors”, Advances in Neural Information Processing Systems, pp. 414–422 (2009).
[22] K. Crammer and D. D. Lee: “Learning via gaussian herding”, Advances in neural information processing systems, pp. 451–459 (2010).
[23] 毎日新聞, http://mainichi.jp/.
[24] C. Cortes and V. Vapnik: “Support-vector networks”, Machine learning, **20**, 3, pp. 273–297 (1995).