文書分類におけるサポートメジャーマシンの有効性

澤井 裕一郎　　吉川 友也　　松本 裕治
奈良先端科学技術大学院大学　情報科学研究科
{sawai.yuichiro.sn0, yuya-y, matsu}@is.naist.jp

1 はじめに

文書分類は，入力された文書を予め定められたクラスのいずれかに分類するタスクである．文書分類では，サポートベクターマシン (SVM)[4] が高い分類性能を示すことが知られており，広く用いられている．

サポートベクターマシンの一般化として，Muandet らはサポートメジャーマシン (SMM) を提案した [6]. SMM は，分布として表現されたデータ点に対してカーネルを定め，マージン最大化に基づき求められる分離平面でデータ点を分類する．Muandet らは SMM を画像分類に適用し，従来手法に比べ高い分類精度が達成できることを示した．しかし，自然言語処理における SMM の適用例はまだ少ない．

文書分類においては，文書を単語ベクトルの分布と見なすことにより，SMM を適用できる．Yoshikawa らは，単語ベクトルを潜在変数として識別学習を行うことで SMM を拡張し，従来手法に比べて高い分類精度を達成した [8]．しかし，この手法は，単語ベクトルの学習の計算量が大きいという欠点がある．

一方で，近年 word2vec[5] などの大規模コーパスから教師なしで単語ベクトルを学習する手法が注目されている．学習済みの単語ベクトルが簡単に使える状態でインターネット上で公開されており，これらを教師あり学習アルゴリズムの素性として用いることができる．例えば，含まれる単語のベクトルの和や平均を文書の素性として利用する方法がある [1]．しかし，この方法では，単語ベクトルの 2 次以上のモーメントの情報が失われてしまう．SMM は，単語ベクトルを高次モーメントの情報を保持しつつ文書分類で利用するための自然な手法である．

本研究では，自然言語処理における基本的な問題である文書分類に SMM を適用し，その有効性を分析する．具体的には，分類において有効なカーネルの種類や単語ベクトルの種類を標準的な文書分類データセットによる実験によって分析する．

2 サポートメジャーマシン

サポートメジャーマシンは，入力として与えられる分布間に定められるカーネルによって規定される高次元空間で，マージンが最大になるような分離平面を求め，分布を分類する．

カーネル k により規定される再生核ヒルベルト空間 (RKHS)H_k への分布 $\mathbb{P}$ のカーネル埋め込みは，

$$\mu_{\mathbb{P}} := \mathbb{E}_{x\sim\mathbb{P}}[k(\cdot,x)] = \int k(\cdot,x)d\mathbb{P} \in H_k \qquad (1)$$

で与えられる．以後 k を埋め込みカーネルと呼称する．

実際の応用では，分布 $\mathbb{P}$ は未知であり，代わりに分布 $\mathbb{P}$ からのサンプル集合 X が与えられる．このサンプル集合を経験分布 $\hat{\mathbb{P}}$ と見なし，次式で与えられる経験カーネル埋め込みを考える．

$$\hat{\mu}_{\mathbb{P}} := \frac{1}{|X|}\sum_{x\in X} k(\cdot,x) \in H_k \qquad (2)$$

SMM は経験カーネル埋め込みに対する SVM である．経験カーネル埋め込み間に定義されるカーネルを，以後レベル 2 カーネルと呼称する．経験分布 $\hat{\mathbb{P}}_i$，$\hat{\mathbb{P}}_j$ の RKHS での内積は，次式で与えられる．

$$\langle\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j}\rangle_{H_k} = \frac{1}{|X_i||X_j|}\sum_{x\in X_i}\sum_{x\prime\in X_j} k(x,x\prime) \qquad (3)$$

ただし，X_i，X_j はそれぞれ分布 $\mathbb{P}_i$，$\mathbb{P}_j$ からのサンプル集合である．レベル 2 カーネルに線形カーネルを用いる場合，最終的な分布間のカーネルは，$K(\mathbb{P}_i,\mathbb{P}_j) = \langle\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j}\rangle_{H_k}$ である．また，レベル 2 カーネルに RBF カーネルを用いる場合，$K(\mathbb{P}_i,\mathbb{P}_j) = \exp\Big(-\frac{\lambda}{2}\|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j}\|^2_{H_k}\Big) = \exp\Big(-\frac{\lambda}{2}(\langle\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_i}\rangle_{H_k} - 2\langle\hat{\mu}_{\mathbb{P}_i},\hat{\mu}_{\mathbb{P}_j}\rangle_{H_k} + \langle\hat{\mu}_{\mathbb{P}_j},\hat{\mu}_{\mathbb{P}_j}\rangle_{H_k})\Big)$ である．ただし，λ はバンド幅パラメータである．

以上より，埋め込みカーネルとレベル 2 カーネルの組み合わせで，SMM のカーネルが定まる．

2.1 SMM の文書分類への適用

図 1 に SMM の文書分類への適用の概念図を示す．分類対象の各文書は，文書が含む単語のベクトルの分布として表現される．文書を表す各分布を SMM のカーネルで埋め込み，RKHS 上で分離することにより，文書分類が行われる．

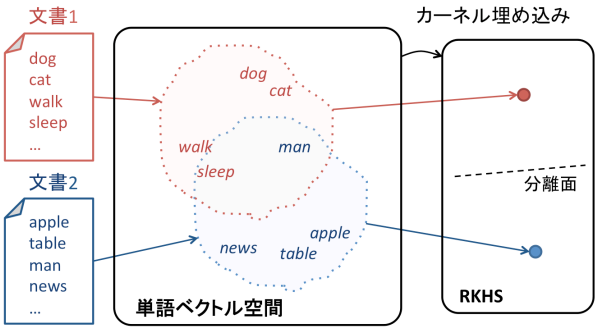

図 1: SMM の文書分類への適用

与えられた N 個の文書 $\{\mathbb{D}_i\}_{i=1}^N$ について，i 番目の文書 $\mathbb{D}_i$ が M_i 個の単語からなる集合 $\{w_{im}\}_{m=1}^{M_i}$ であるとする．また，各単語から d 次元実数値ベクトルへの写像を v とする．

レベル 2 カーネルを線形カーネルとすると，埋め込みカーネルが線形カーネルである場合，文書 $\mathbb{D}_i$，$\mathbb{D}_j$ 間のカーネルは次式で計算される．

$$K_{\mathrm{LIN}}^{\mathrm{LIN}}(\mathbb{D}_i,\mathbb{D}_j)=\frac{1}{M_iM_j}\sum_{g=1}^{M_i}\sum_{h=1}^{M_j}v(w_{ig})^Tv(w_{jh}) \quad (4)$$

埋め込みカーネルが RBF カーネルである場合，バンド幅パラメータを γ とすると，文書間のカーネルは次式で計算される．

$$K_{\mathrm{RBF}}^{\mathrm{LIN}}(\mathbb{D}_i,\mathbb{D}_j)=\frac{1}{M_iM_j}\sum_{g=1}^{M_i}\sum_{h=1}^{M_j}e^{-\frac{\gamma}{2}\|v(w_{ig})-v(w_{jh})\|_2^2} \quad (5)$$

レベル 2 カーネルが RBF カーネルである場合，バンド幅パラメータを λ，埋め込みカーネルを $emb\in\{\mathrm{LIN},\mathrm{RBF}\}$ とすると，文書間のカーネルは次式で計算される．

$$K_{emb}^{\mathrm{RBF}}(\mathbb{D}_i,\mathbb{D}_j)=e^{-\frac{\lambda}{2}(K_{emb}^{\mathrm{LIN}}(\mathbb{D}_i,\mathbb{D}_i)-2K_{emb}^{\mathrm{LIN}}(\mathbb{D}_i,\mathbb{D}_j)+K_{emb}^{\mathrm{LIN}}(\mathbb{D}_j,\mathbb{D}_j))} \quad (6)$$

これらのカーネルを用いて SVM を学習することで，SMM が学習できる．

3 実験

文書分類における SMM のカーネルと単語ベクトルの影響を調査するために，以下の実験を行った．

3.1 実験設定

実験で使用した文書分類データセットの詳細を，表 1 に示す．これらは [3] の著者のウェブサイト [1] からダウンロードでき，前処理としてストップワードの除去と単語のステミングが既に施されている．さらに，今回の実験では低頻度語を除去するために，全訓練データ中 0.1%以上の文書に出現する単語のみを使用した．

表 1: 実験で使用したデータセット

データセット	訓練数	テスト数	クラス数
WebKB	2784	1381	4
Reuters 21578	5485	2189	8
20 Newsgroups	11293	7524	20

SMM に関しては，文書間のカーネル行列を式 (4), (5), (6) を使い計算した後，LIBSVM[2] で学習と分類を行った．多クラス分類は one-vs-one 方式で行った．各データセットについて，訓練データから抽出した 1000 文書を開発データとして使用し，グリッドサーチによりハイパーパラメータの最適化を行った．

各実験では，ベースラインとして，bag-of-words 素性を SVM に使用する場合 (SVM(BoW)) の分類精度を示す．また，bag-of-words 素性に対して SVD で次元圧縮を施した素性を SVM に使用する手法 (SVM(SVD)) との比較も行う．SVD による次元圧縮後の次元数は，100 次元から 1000 次元までの範囲で，最も分類精度が高いものを使用した．

3.2 カーネルの比較実験

埋め込みカーネル，レベル 2 カーネルそれぞれについて，線形カーネル (LIN) と RBF カーネル (RBF) の有効な組み合わせを調べた．

単語ベクトルとしては，Google News corpus（約 1000 億単語）から word2vec[5] で学習された，著者のウェブサイト [3] で配布されている 300 次元の単語ベクトルを用いた．今回使用した前処理済みの文書分類データセットでは，各単語が既にステミングされた状態で公開されている．SMM を適用するためには，これらの各語幹を，配布されている word2vec の単語ベクトルデータの各単語と対応付ける必要がある．そのために，各語幹が接頭辞である単語のうち，最も頻度が高い単語に対応づけるという方法をとった．

図 2 に，訓練に使用する文書数を 200 文書から 1000 文書まで変えたときの分類精度を示す．各結果は，(埋め込みカーネル)+(レベル 2 カーネル) という順で表記している．

埋め込みカーネルとしては，RBF カーネルを用いる場合に精度が高くなる傾向にある．また，レベル 2 カーネルは，RBF カーネルを用いても線形カーネルを用いても精度に大きな差が見られなかった．したがって，埋め込みカーネルとしては RBF カーネルを使用し，レベル 2 カーネルとしてはパラメータ選択の必要

[1] http://web.ist.utl.pt/acardoso/datasets/

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[3] https://code.google.com/p/word2vec/

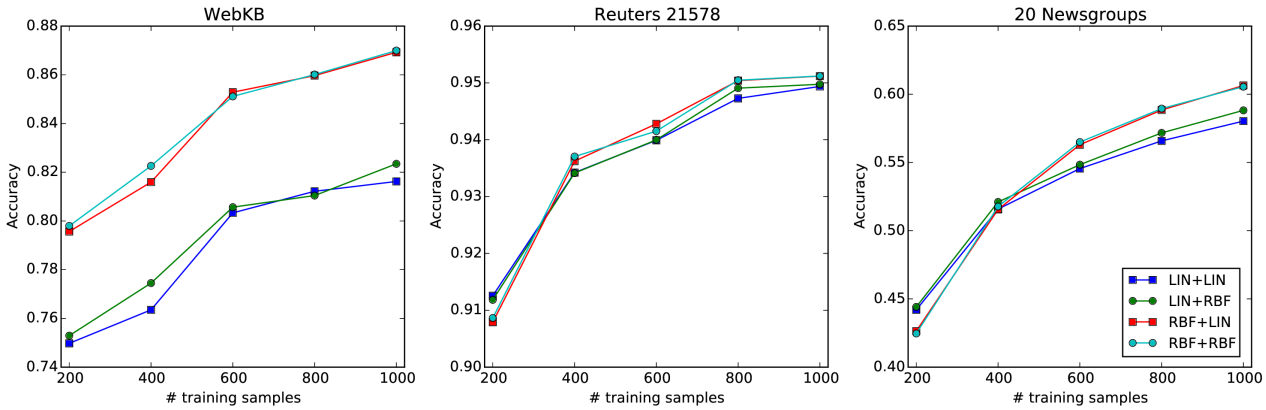

図 2: カーネルの比較実験

がない線形カーネルを用いるという組み合わせが最も有効である．

3.3 訓練データから学習した単語ベクトルの比較実験

訓練データ自体から単語ベクトルを得る方法としては，特異値分解 (SVD) と潜在的ディリクレ配分法 (LDA)[2] が考えられる．SVD では，bag-of-words データから作られる単語-文書行列に対して特異値分解を施すと得られる単語数×次元数の大きさの行列の各行を単語ベクトルと見なせる．LDA では，各単語について得られるトピックの分布を単語ベクトルと見なせる．

実験では，埋め込みカーネルに RBF カーネル，レベル 2 カーネルに線形カーネルを使用した．訓練データとして 600 文書を使用した．また，式 (5) より，$\gamma \to \infty$ の極限では，文書間のカーネルは単語の一致度にのみ基づき計算される．したがって，ランダムな単語ベクトルを使用する場合でも，γ を十分大きくとれば SMM はある程度の分類精度を示す．SVD と LDA による単語ベクトルが，ランダムな単語ベクトルに比べて分類に有効であることを確かめるために，各次元の値を $[-1, 1]$ の一様分布から得たランダムな単語ベクトル (300 次元) と比較した．図 3 に各手法で得られる単語ベクトルの次元数を 20 次元から 400 次元まで変化させたときの分類精度を示す．

SVD，LDA で得た単語ベクトルを使用した場合，ランダムな単語ベクトルを使用した場合に比べて，分類精度が向上する場合があった．特に 20 Newsgroups では，SVD と LDA 両方の手法で 1%以上の精度向上があった．

3.4 コーパスから学習した単語ベクトルの比較実験

訓練データ以外のコーパスから学習した単語ベクトルを SMM に使用した場合の精度の比較を行った．

表 2: 実験で使用した単語ベクトル

学習手法	学習コーパス	単語数	次元数
word2vec	Google News corpus	100B	300
GloVe	Wikipedia 2014 + Gigaword 5	6B	300
GloVe	Common Crawl	42B	300
GloVe	Common Crawl	840B	300
(Random)	(なし)	(0)	300

カーネルの比較実験で使用した word2vec による単語ベクトルと，GloVe[7] による単語ベクトル [4] を比較した．また，300 次元のランダムな単語ベクトルを使った場合 (SMM(Random)) との比較も行った．表 2 に実験で使用した単語ベクトルの詳細を示す．

埋め込みカーネルは RBF カーネル，レベル 2 カーネルは線形カーネルを使用した．図 4 に，訓練に使用する文書数を 200 文書から 1000 文書まで変えたときの，各単語ベクトル学習法での分類精度を示す．図 4 の SMM(word2vec) は，図 2 の RBF+LIN と同じである．

Reuters 21578 と 20 Newsgroups で，コーパスから学習した単語ベクトルを使用することによる精度向上が見られた．一方で， WebKB では，ランダムな単語ベクトルを使った場合と比べて精度の向上は小さかった．これは，WebKB が大学の授業や教員のウェブページを集めたものであり，単語ベクトルの学習に使用したコーパスとドメインが大きく異なることが原因であると推測される．WebKB と Reuters 21578 では，SVM(SVD) がコーパスから学習した単語ベクトルを使用しないにも関わらず高い分類精度を示した．一方で，20 Newsgroups では，GloVe による単語ベク

[4] http://nlp.stanford.edu/projects/glove/

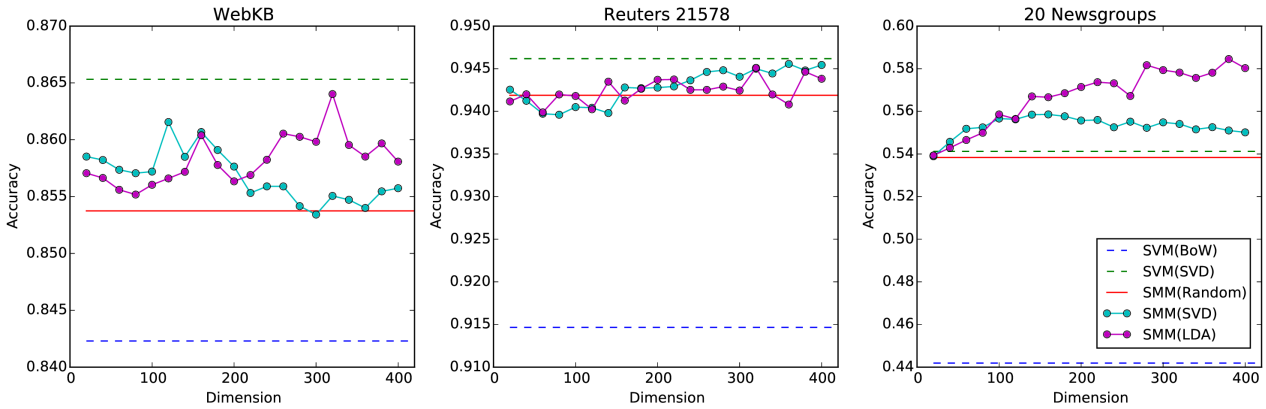

図 3: 訓練データから学習した単語ベクトルの比較実験

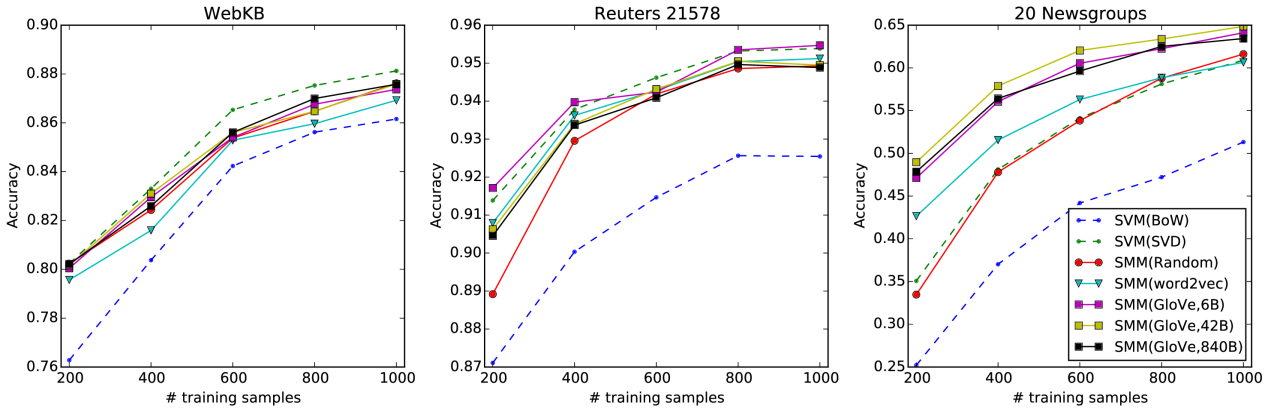

図 4: コーパスから学習した単語ベクトルの比較実験

トルを用いた場合の精度が最も高かった.

4 おわりに

本研究では，サポートメジャーマシン (SMM) の文書分類における有効性を調査するにあたり，カーネルと単語ベクトルの比較を行った.

結果として，カーネルの組み合わせとしては，埋め込みカーネルで RBF カーネルを使い，レベル 2 カーネルでは線形カーネルを使用することが妥当であることがわかった．単語ベクトルについては，SVD や LDA で得た単語ベクトルを SMM に利用することで，ランダムな単語ベクトルを使用する場合と比較して，分類精度の向上が可能であることがわかった．また，ドメインによっては，訓練データ以外の外部のコーパスから学習した単語ベクトルを利用することにより，さらに分類精度の向上が可能であることがわかった.

20 Newsgroups では，既存手法 (SVM(SVD)) に比べて SMM が有効であることが示された．今後の課題として，SMM が有効なデータの種類を解明したい.

参考文献

[1] Silvio Amir, Miguel B. Almeida, Bruno Martins, João Filgueiras, and Mario J. Silva. TUGAS: Exploiting unlabelled data for twitter sentiment analysis. In *SemEval 2014*, pp. 673–677, 2014.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[3] Ana Cardoso-Cachopo. *Improving methods for single-label text categorization*. PhD thesis, 2007.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119, 2013.

[6] Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *NIPS*, pp. 10–18, 2012.

[7] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: global vectors for word representation. *EMNLP*, Vol. 12, , 2014.

[8] Yuya Yoshikawa, Tomoharu Iwata, and Hiroshi Sawada. Latent support measure machines for bag-of-words data classification. In *NIPS*, pp. 1961–1969, 2014.