一般二項定理による多項式カーネルの拡張と学習

椿真史，Kevin Duh，新保仁，松本裕治
奈良先端科学技術大学院大学 情報科学研究科
自然言語処理学 松本裕治研究室
〒 630-0192 奈良県生駒市高山町 8916-5
{masashi-t,kevinduh,shimbo,matsu}@is.naist.jp

Abstract

本稿では，機械学習の一分野である類似度学習におけるカーネルを用いた非線形拡張の中でも，特に多項式カーネルに着目し，その一般化を行う．多項式カーネルは，アイザック・ニュートンが 1665 年頃に発見した一般二項定理を用いることで，ガウシアンカーネルやシグモイドカーネルと同様に無限級数展開される．我々はこれを，一般多項式カーネルと呼ぶ．この一般多項式カーネルでは，無限級数における各々の項の重みが，複雑なガンマ関数によって制御される．これにより，既存の類似度学習を非線形に拡張するだけでなく，意味や構造などの様々な種類の素性ベクトルを統一的に扱い，それらの適切な組み合わせ（次数）をデータから自動的に獲得できることが期待される．我々の基本的なアイディアは，カーネルを用いた機械学習手法へ様々に適用可能である．

1 はじめに

自然言語処理において，何らかの二つの対象（例えば単語や句，文や文書など）の類似度を計算することは非常に重要である．この類似度に基づき，情報の検索や抽出，文書の分類やクラスタリングを行う．多くの従来手法は，与えられた 2 つの文に含まれる単語や N-gram のマッチング等の表層的な素性に基づくものが多いが，このような手法には言語の意味的な情報は用いられていない．しかし近年では，意味からのアプローチが盛んに研究され，単語ベクトル空間モデルはその基礎を提供する．(Turney, 2013; Collobert et al., 2011).

しかしながら，単語ベクトル空間モデルのみでは，単語の意味的類似度はある程度適切に計算できるものの，文といったより複雑な対象を扱う際には，未だ二つの大きな問題が存在する．

1. 単語ベクトル空間において句や文をどのように表現，あるいは学習するか？
2. 句や文の類似度をどのように計算，あるいは学習するか？

1 については，多くの研究 (Tsubaki et al., 2013) があり，特に近年では Deep Learning の分野で盛んに研究されている (Socher et al., 2012)．一方で 2 については，主に機械学習の分野において，計量，距離，あるいは類似度学習として以前から研究され続けている (Xing et al., 2002; Chechik et al., 2009)．上述の自然言語処理における意味表現と機械学習における類似度学習は密接に関係しているが，これら 2 つの分野を繋ぐような研究は未だ少ない．

そこで我々は，これら 2 つの研究分野を繋ぐため，本稿では 2 に着目し，特にカーネルを用いた類似度学習の非線形拡張に焦点を当てる．そして本稿で最も重要となるのが，自然言語処理で通常用いられる多項式カーネル $K_p(\mathbf{x},\mathbf{y}) = (c+\mathbf{x}^{\mathrm{T}}\mathbf{y})^p$（ただし $c \geq 0, p \in \mathbb{N}$）を一般化した一般多項式カーネル $K_{gp}(\mathbf{x},\mathbf{y}) = (1+\mathbf{x}^{\mathrm{T}}\mathbf{y})^n$（ただし $|\mathbf{x}^{\mathrm{T}}\mathbf{y}| < 1, p \in \mathbb{R}$）である．ガウシアンカーネルやシグモイドカーネルでは無限次元空間を扱うことが可能であるが，多項式カーネルでは有限次元空間である．本稿で提案するこの一般多項式カーネルは，指数部分が $n \in \mathbb{R}$ の実数値であるため，無限級数展開される．本稿では非線形類似度学習において，この一般多項式カーネルを用いることに焦点を当てる．本稿の貢献は以下の通りである．

1. 多項式カーネルの一般化とその有用性の検証に関する研究は，我々の知る限り本稿が初めてである．
2. 一般多項式カーネルは，様々な種類の素性クラス（意味，品詞タグ，係り受け関係ラベルなど）に対する適切な組み合わせ（次数）を，データから自動に学習することが可能であることを示した．

2 背景

2.1 単語ベクトル空間モデルと意味の構成性

単語ベクトル空間モデルとは，単語の意味的な情報をベクトル空間において表現する一般的な枠組みのことである．古くからは，潜在意味解析と呼ばれる手法が存在する (Deerwester et al., 1990)．一方で特に近年では，ニューラルネットワークを用いた手法が注目され

ている (Collobert et al., 2011). しかしながら，単語ベクトル空間モデルのみでは，より長い句や文の意味をどのように表現するのかという大きな問題が残る．これを解決するために，特定の種類の句 (Tsubaki et al., 2013) や，任意の文の意味を計算する様々なモデルが提案されている (Socher et al., 2012).

2.2 距離あるいは類似度学習

距離学習と呼ばれる分野は，すでに得られているデータ間の距離を，タスクに合わせて処理しやすい距離へ変換することを目的とする. 例えば，同じラベルを持つデータ間の距離は近く，異なるラベルを持つデータ間の距離は遠くなるように，ベクトルデータ自体の変換を行う (Xing et al., 2002). また一方で，類似度学習においては, 距離ではなく主に内積を最適化する (Chechik et al., 2009). 近年ではさらに，その非線形拡張が提案され (Kedem et al., 2012)，データ点をカーネル関数 $K(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{y})$ によって写像した後の高次元空間のユークリッド距離 $d_\phi(\mathbf{x},\mathbf{y}) = ||\phi(\mathbf{x}) - \phi(\mathbf{y})||_2$ や，正規化されたカーネルを直接学習するなどの手法も存在する．これらの距離ないしは類似度学習によって，元のデータ点から，解くべき問題に合わせてより適切なベクトル空間を新たに得ることができる.

3 提案手法

訓練データセットは $\{(S_i, S'_i), y_i\}_{i=1}^n$ の形式で与えらるものとする (4.1 節参照). 我々は，二つの文 S と S' の類似度計算についてはシンプルな既存手法を述べ (3.1 節)，その後 2 つの文の類似度である連続値 $y \in C = [-1, +1]$ を予測する学習モデルを提案する (3.2 節). 本稿では，3.2 節の類似度学習において，カーネル法を用いた非線形性拡張の中でも特に，多項式カーネルとその一般化に焦点を当てる.

3.1 文の類似度計算

まず最も単純に、$sim(S, S')$ を文 S と S' の類似度とし，以下のように計算する.

$$sim(S, S') = K\left(\sum_{w\in S} \mathbf{d}(w), \sum_{w'\in S'} \mathbf{d}(w')\right) \quad (1)$$

ここで，$\mathbf{d}(w)$ は n 次元の単語ベクトル表現とする．つまり，文 S に含まれる単語 w のベクトルの総和について，カーネル K を計算する.

次に，もう一つの文の類似度計算について，以下の計算手法を用いる.

$$sim(S, S') = \frac{1}{|S||S'|}\sum_{i=1}^{|S|}\sum_{j=1}^{|S'|} K(\mathbf{d}(w_i), \mathbf{d}(w_j)) \quad (2)$$

ここで，$|S|$ は文の長さ，w_i は i 番目の単語を表す. この手法では，2 つの文 S と S' に含まれる単語ベクトル間のすべてのカーネルの平均を取る．これは (Kanagawa and Fukumizu, 2014) などにおいて提案される確率分布のカーネル埋め込みと同様であり，近年はこれを分類手法として用いる Support Measure Machine が提案されている (Yoshikawa et al., 2014)．我々はこれを，後述する類似度学習に適用する.

最後に，構文解析した結果として出てくる品詞タグや係り受け関係ラベルを，単語と同様にベクトル表現（初期値はランダムとして後に学習する）した上で，それを用いて文 S と S' の構造的類似度を以下のように計算する.

$$sim(S, S') = K\left(\sum_{(i,j)\in D} \mathbf{t}(w_i, w_j), \sum_{(i',j')\in D'} \mathbf{t}(w_{i'}, w_{j'})\right)$$

ここで D は係り受け関係にある単語のペア集合，$\mathbf{t}(w_i, w_j)$ はそれらの単語の品詞タグと係り受け関係ラベルの各々をベクトルで表現した上で，それらを連結させたものとする．これは (Chen and Manning, 2014) の手法と類似しており，彼らはこのようなベクトルを設計し学習した上で，依存構造解析する手法を提案している．我々は式 (1) とこれを組み合わせ，後述する手法を用いて意味と構造のベクトル表現を最適化する.

3.2 一般多項式カーネルを用いた非線形類似度学習

まず我々は，最も基本となるカーネル K に，正規化された内積である以下のコサイン類似度を用いる．コサイン類似度は，単語ベクトル空間における意味的類似度として幅広く用いられるものである．本稿で述べるすべてのカーネルについても，同様に以下のように正規化する．正規化されたカーネルは，写像された高次元空間 ϕ におけるコサイン類似度と等価となる.

$$\cos(\phi(\mathbf{x}), \phi(\mathbf{x}')) = \frac{K(\mathbf{x}, \mathbf{x}')}{\sqrt{K(\mathbf{x}, \mathbf{x})}\sqrt{K(\mathbf{x}', \mathbf{x}')}} \quad (3)$$

我々は，高次元空間 ϕ に写像されたベクトルを陽に計算することなく，カーネルを通して類似度を計算し学習する.

そして本稿では特に，以下の 2 つの多項式カーネルに着目し，その比較を行う.

$$K_p(\mathbf{x}, \mathbf{x}') = (c + \cos(\mathbf{x}, \mathbf{x}'))^p \quad s.t \quad c \geq 0, \ \ p \in \mathbb{N} \quad (4)$$

$$K_{gp}(\mathbf{x}, \mathbf{x}') = (1 + \cos(\mathbf{x}, \mathbf{x}'))^n \quad s.t \quad |\cos(\mathbf{x}^{\mathrm{T}}\mathbf{x}')| < 1, \ \ n \in \mathbb{R} \quad (5)$$

式 (5) は機械学習で通常用いられる多項式カーネルであり，式 (6) は本稿で提案する一般多項式カーネルである．これは，アイザック・ニュートンが 1665 年頃に発見した一般二項定理を用いた拡張であり，本稿で我々はこれを一般多項式カーネルと呼ぶ．式 (6) の右辺は，ガンマ関数

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \quad (6)$$

を用いて，以下のように無限級数展開される [1].

$$\sum_{k=0}^{\infty} \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \cos(\mathbf{x}, \mathbf{x}')^k \qquad (7)$$

このような拡張により，無限次元空間を扱うことが可能になると共に，3.1 節で述べたような単語の意味ベクトルの他，品詞タグ，係り受け関係ラベルなどの様々な種類の素性の最適な組み合わせ（次数）を学習することが可能となる．

最終的なロス関数は以下の通りである．

$$L(\Theta) = \sum_{i=1}^{n} \frac{1}{2}\{y_i - K(\mathbf{x}_i, \mathbf{x'}_i)\}^2 + \frac{\lambda}{2}||\Theta||^2 \qquad (8)$$

ここで，Θ は学習するパラメータの集合であり，単語の意味，品詞タグ，係り受け関係ラベルのベクトル表現とカーネル内のパラメータである．

4 実験

4.1 文の意味的類似度の評価データセット

本稿の提案手法は，SemEval 2014 の Sentences Involving Compositional Knowledge (SICK) (Marelli et al., 2014) のデータセットを用いて評価した．これは，2 つの文の意味的な類似度を人手でスコア付けしたものである（Table 1 を参照）．評価には，提案手法によって計算された二つの文ベクトルの類似度と，人手の類似度スコアとのピアソンの相関係数を用いる．

4.2 実装の詳細

単語ベクトル表現には，SENNA[2] を用いた．SENNA はニューラルネットワークを用いて Wikipedia をコーパスとして学習された，50 次元の単語ベクトル表現である．また，構文解析には Enju[3] を使用した．

最終的なコスト関数は $L(\Theta)$ であり，これを最小化する．非線形類似度学習における単語ベクトルとカーネル内パラメータの最適化には，AdaGrad (Duchi et al., 2011) を用いた．単語ベクトル表現の学習率は $\alpha = 1.0$, カーネル内パラメータの学習率は $\beta = 10^{-2}$，正則化項については $\lambda = 10^{-6}$ とした．データセットに対してはイテレーションを 100 に統一し実験を行い，比較検証した．そして一般多項式カーネルの次数の学習は，初期値 1.0 から開始し，その推移を見た．また，比較する既存研究として，SICK の上位 3 チーム（詳細は 6 章）の結果と比較した．

5 結果と考察

5.1 線形 vs 多項式カーネル

Table 2 を見るとわかるように，線形であるコサイン類似度よりも，非線形である多項式カーネルを用いた場合に，相関係数の大幅な上昇が見られた．ただし，一般多項式カーネルで次数を最適化する場合と次数を固定した場合とでは，性能の差異は見られなかった（次数はほぼ 3 に収束した）．また，文に含まれるすべての単語ベクトル間のカーネルの平均、あるいは品詞や係り受けの文構造の情報を用いたモデルでは，著しく相関係数が低い結果となった．これは，類似度の取り方や構造の入れ方が適切でない等の理由が考えられ，今後の重要な課題となった．しかし，意味表現のベクトル，品詞タグと係り受け関係ラベルのベクトルとでは，多項式カーネルの次数は異なるように学習されるのを確認した．

カーネル	r (ADD)	r (MAT)	r (POS+DEP)
コサイン	0.740	0.448	0.575
多項式 (p=3)	**0.817**	0.574	0.630
一般多項式	**0.817**	0.588	0.632

Table 2: 様々なカーネルを用いた場合の相関係数の比較．ADD は単純な単語ベクトルの総和，MAT は単語ベクトル行列間の類似度の平均，POS + DEP は意味ベクトルの他，それらの構造ベクトルも同時に最適化するモデルを指す．

Models	r (Rank)
Zhao et al., 2014	**0.828 (1)**
Jerva et al., 2014	0.827 (2)
Our best model	**0.817 (3)**
Jimenez et al., 2014	0.804 (4)

Table 3: 我々の手法と SICK の上位チームの結果との比較．

5.2 提案手法 vs 既存研究

Table 3 を見るとわかるように，我々の提案手法は SemEval 2014 ランキングで 3 位に位置している．我々の手法において強調したいのは，シンプルな文の意味ベクトル表現とその非線形類似度学習のみで，高い相関係数を達成している点である．比較する既存手法では大量の素性を用いている点を考えても（6 章参照），我々の手法に大きな優位性がある．しかし，本稿で着目し提案した一般多項式カーネルについては，5.1 節と同様の課題は残る．

6 関連研究

従来研究における文の意味的類似度計算の主なアプローチは，本稿とはまったく異なる．従来研究では例えば，2 つの文に含まれる単語や N-gram のマッチング，品詞や木構造のアライメント，さらには WordNet などの外部知識や機械翻訳に用いられる評価指標などの様々な素性を考え，それらを用いてサポートベクター回帰で学習するものがほとんどである．SemEval2014 の SICK に関しても，同様の手法に基づいたアプローチが多数を占めている (Zhao et al., 2014; Bjerva et al.,

[1] 無限級数における係数は，n 次を過ぎた直後急激に減衰し，正と負の微小な値の両方を取るため，厳密にこれは正定値カーネルではないことに注意されたい．

[2] http://ronan.collobert.com/senna/

[3] http://www.nactem.ac.uk/enju/index.ja.html

文 A と B	人手による類似度スコア
A : *A man is jumping into an empty pool.* B : *There is no biker jumping in the air.*	1.6
A : *Two children are lying in the snow and are making snow angels.* B : *Two angels are making snow on the lying children.*	2.9
A : *The young boys are playing outdoors and the man is smiling nearby.* B : *There is no boy playing outdoors and there is no man smiling.*	3.6
A : *A person in a black jacket is doing tricks on a motorbike.* B : *A man in a black jacket is doing tricks on a motorbike.*	4.9

Table 1: データは Amazon Mechanical Turk で複数のアノテータらによって作られ，類似度スコア（高いほど意味的類似度が高い）はそれらの平均値である．訓練データとテストデータは各々5000 文対から成る．

2014; Jimenez et al., 2014)．特に，単語ベクトル空間を用いた意味的なアプローチのみでは，相関係数が 0.7 程度に留まるという報告がある (Marelli et al., 2014)．これらの手法と比較して，我々の手法が特に異なる部分は，単語ベクトル表現をベース とした意味的なアプローチのみを用いている点であり，それによって最高性能に迫る結果を達成することに成功している．

7 結論と今後の課題

本稿で我々は，単語ベクトル空間から文の構成に伴って生じる新たな意味空間の類似度学習について，非線形拡張の中でも特に多項式カーネルの一般化に焦点を当てた．提案手法の優位性は確認されなかったものの，機械学習手法における素性の組み合わせ最適化という問題について，新たな知見を提供する手法だと考えている．今後の課題は以下の通りである．

1. 意味と構造のデータの双方をより適切にカーネル埋め込みする枠組みを提案する．
2. 深層カーネル（Deep Kernel）を用いて，非線形類似度学習手法をより拡張させる．

References

Johannes Bjerva, Johan Bos, Rob van der Goot, and Malvina Nissim. 2014. The meaning factory: Formal semantics for recognizing textual entailment and determining semantic similarity.

Gal Chechik, Uri Shalit, Varun Sharma, and Samy Bengio. 2009. An online algorithm for large scale image similarity learning. In *NIPS*.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 12:2493–2537.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12:2121–2159.

Sergio Jimenez, George Duenas, Julia Baquero, Alexander Gelbukh, Av Juan Dios Bátiz, and Av Mendizábal. 2014. Unal-nlp: Combining soft cardinality features for semantic textual similarity, relatedness and entailment.

Motonobu Kanagawa and Kenji Fukumizu. 2014. Recovering distributions from gaussian rkhs embeddings. In *AISTATS*.

Dor Kedem, Stephen Tyree, Fei Sha, Gert R Lanckriet, and Kilian Q Weinberger. 2012. Non-linear metric learning. In *NIPS*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *SemEval*.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *EMNLP-CoNLL*.

Masashi Tsubaki, Kevin Duh, Masashi Shimbo, and Yuji Matsumoto. 2013. Modeling and learning semantic co-compositionality through prototype projections and neural networks. In *EMNLP*.

Peter D. Turney. 2013. Domain and function: Dual-space model of semantic relations and compositions. *CoRR*, abs/1309.4035.

Eric P Xing, Michael I Jordan, Stuart Russell, and Andrew Y Ng. 2002. Distance metric learning with application to clustering with side-information. In *NIPS*.

Yuya Yoshikawa, Tomoharu Iwata, and Hiroshi Sawada. 2014. Latent support measure machines for bag-of-words data classification. In *NIPS*.

Jiang Zhao, Tian Tian Zhu, and Man Lan. 2014. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment.