双極空間への埋め込みを利用した文書識別

Document Classification using Poincare Embeddings

逆瀬川 滉大[1]　峯 恒憲[2]　廣川 佐千男[3]

Kota Sakasegawa[1], Tsunenori Mine[2], and Sachio Hirokawa[3]

[1] 九州大学システム情報情報科学府

[1] Graduate School of Information Science and Electrical Engineering, Kyushu University

[2] 九州大学システム情報科学研究院

[2] Faculty of Information Science and Electrical Engineering, Kyushu University

[3] 九州大学情報開発基盤センター

[3] Research Institute for Information Technology, Kyushu University

Abstract: Document classification is an important task, however document vectorization methods are often in the Euclidean space. In this study, we vectorized documents on Poincare space and evaluated the classification performance.

1. はじめに

人手には負えない膨大な数の文書（テキストデータ）を扱う方法として、機械学習を用いた文書分類が挙げられる。機械学習を用いた文書分類の際には対象の文書をベクトル化することが必要となる。このベクトル化手法には多種の手法が提案されているが、その中でも精度の高いものとして state-of-the-art を記録した Mekala ら[1]の SCDV や、その手法の前身である Vivek ら[2]の BowV がある。これらの手法は最初に単語ベクトルを作成し、クラスタリング、スパース化など複数のステップと経て文書ベクトルを作成する方法だが、いずれもユークリッド空間への埋め込みで作成される文書ベクトルであるといえる。一方で、対象のベクトル化手法として、Nickel ら[3]の ponicare embeddings をはじめとする双曲空間への埋め込みが注目されている。この手法は、従来の多くが対象データをユークリッド空間に埋め込むことでベクトル化していたのに対し、非ユークリッド幾何である双曲空間に埋め込むことでベクトル化する。この手法では双曲空間と階層構造や木構造をもつデータとの親和性により、階層的な構造をもつデータを効率よく取り込むことができる。実際に Nickel[3]は、WordNet で定義される単語の上位下位の階層関係を 5 次元の双曲空間に埋め込むことで、200 次元のユークリッド空間に埋め込んだ場合より高い精度を示した。

Nickel ら[3]が、ponicare embeddings の手法を階層構造を持つ有向グラフである wordnet や無向グラフである ASTROPH などの共著ネットワークに用いたのに対し、双曲空間の埋め込み手法をテキストに用いる研究も複数存在するが、文書識別に利用した例は少ないと言える。本研究では、poincare embeddings を利用した文書ベクトル化手法を提案し、提案手法を用いた際の文書識別性能の評価を 20news group dataset を用いて行った。

2. 関連研究

poincare embeddings は Nickel ら[3]が提案した、双曲空間のモデルの一つであるポアンカレボールモデル上でのベクトル表現獲得手法である。poincare embeddings を用いることで、Nickel ら[3]は wordnet などのグラフデータをユークリッド空間でのベクトル化に比べて高精度かつ低次元で、ベクトル化した。その後も pooincare embeddings に続く形で、双曲空間での埋め込みを利用するような研究が多数行われている。

自然言語処理に利用した研究としては、Tifrea ら[4]や Leimeister ら[5]、Dhingra ら[6]、Chen ら[7]などがある。Tifrea ら[4]や Leimeister ら[5]は双曲空間上での単語の分散表現の獲得を目的としたものであり、文書識別タスクは行っていない。Dhingra ら[6]や、Chen ら[7]では事前に汎用的なコーパスで学習した単語ベクトルを用いて文章識別や文書識別に取り組んでいるが、本研究では、Mekala ら[1]の SCDV や、Vivek ら[2]の bowv と同じく、識別対象の

文書群から単語ベクトル、文書ベクトルを作成して文書識別を行うことを目指す。

3. Poincare embeddings を利用した文書ベクトル化手法

3.1 各単語のベクトル化

はじめにテキストから poincare embeddings を利用して単語のベクトル化を行う。単語のベクトル化にあたっては、Dhingra ら[6]と同様にテキスト中の指定ウィンドウサイズ以内で共起した単語の共起関係のグラフを学習データとして、poincare embeddings の手法を適用する。今回ウィンドウサイズは Dhingra ら[6]と同じく 5 とした。損失関数は Nickel ら[3]や Dhingra ら[6]と同じく、(1)、(２)式で定義されるものを扱う。ただし、学習データに共起する（単語、単語）の組に加えて、(単語、その単語が出現した文書カテゴリ）の関係も学習データに与えて学習させた。

カテゴリと単語との関係も学習させるのは、次の工程で文書をベクトル化するにあたって、より文書識別しやすいような単語ベクトルのベクトル表現を得るためである。Nickel [3] らや Dhingra ら[6]の報告にあるように、中心付近に多くの単語と共起するような上位語が、円周付近には少数の単語と共起するようなより具体的な単語が配置されるのであれば、各カテゴリベクトルの近傍かつ円周側に各カテゴリの特徴語が集められ、逆に複数のカテゴリにまたがって出現するようなカテゴリよりも上位と言えるような単語については、各カテゴリベクトルよりも原点付近に配置されるのではないかと考えられる。実際に、予備実験として Reuter21578 データセットのサブセットを用いて単語・カテゴリの関係を学習させ、2 次元の埋め込み表現を可視化したものが図１、図 2 である。図 1 は各カテゴリの文書にのみ出現した単語の単語ベクトルをプロットしたものであり、図 2 は得られた単語ベクトルから 3.2 で述べる手法で作成された文書ベクトルをプロットしたものである。カテゴリごとに特徴語、文書ベクトルの分布が分離していることが確認できる。

また、単語カテゴリ間の学習で得られるベクトル表現について、ポアンカレモデルで定義される距離である(2)式を用いた場合と通常のユークリッド空間で定義される距離を用いた場合との比較が図３にあたる。比較にあたっての評価指標には、Nickel [3] らや Dhingra ら[6]と同じく mean rank を使用した。ベクトルの次元数を変化させたところ、ポアンカレモデルで定義される距離を用いた場合の方が低次元でも精度を保つことが確認できた。これは Nickel [3]らの報告とも一致する。このことからも、poincare embeddings で効率的に単語・カテゴリの関係が学習されることが期待される。

$$\mathrm{L} = -\sum_{(u,v)\in D} log\frac{\exp\left(-d(u,v)\right)}{\sum_{v'\in N(u)\cup\{v\}}\exp\left(-d(u,v')\right)} \quad (1)$$

$$\mathrm{d(u,v)} = \cosh^{-1}\left(1+2\frac{\|u-v\|^2}{(1-\|u\|^2)(1-\|u\|^2)}\right) \quad (2)$$

3.2 各文書のベクトル化

各文書のベクトル化にあたっては、文書内単語の単語ベクトルの中点をもって、その文書のベクトルとした。具体的には (3),(4)式で表されるポアンカレモデル上での中点の式を利用することで文書をベクトル化している[8]。

$$\frac{1}{2}\otimes\sum_{j}^{n}\frac{\gamma(V_j)^2V_j}{\sum_l\gamma(V_l)^2-\frac{n}{2}} \quad (3)$$

$$\mathrm{r}\otimes\mathrm{v} = \tanh(\mathrm{r}*\tanh^{-1}\|v\|)\frac{v}{\|v\|} \quad (4)$$

4. 実験手順

poincare embeddings を利用した文書識別性能を評価するにあたって、文書識別の標準的ベンチマークデータである 20news group datasets を用いた。比較対象の base line としては、skipgram negative sampling で得た単語ベクトルに対して、同じく文書内単語の単語ベクトルの平均による文書ベクトル化を用いている。文書識別にあたっての分類器には baseline には svm を用いるが、poincare embeddings を利用した文書ベクトル化に対しては、Cho ら[9]の hyperbolic svm（以下 hsvm）を使用した。なお、実験にあたっては Agibetov ら[10]の python 実装の hsvm を用いている。なお、単語ベクトルは 200 次元で作成している。

学習・テストデータの分割は 7:3 とし、学習データの 2 割をバリデーション用のデータとしてハイパーパラメータのチューニングの調整に使用した。

5. 実験結果及び考察

識別性能の評価結果を表 1 に示す。skipgram

negative sampling を用いた場合に、及ばない結果となった。Nickel [3]らの報告にもあるように、次元数が多いため、あまり euclidean と poincare での精度差が出なかったことが考えられる。

表１．識別性能の比較

	skipgram negative sampling	ponicare embeddings
macro precision	0.836215	0.825178
macro recall	0.834202	0.815371
macro F1	0.833679	0.816723
accuracy	0.841237	0.821502

6. おわりに

単語同士の共起関係と単語・カテゴリの関係を Poincare embeddings 学習させ、得られた単語ベクトルを用いて文書ベクトル化を行ったが、skipgram negative を用いた場合に対して、文書ベクトルの識別性能は向上されず、わずかに下回る結果となった。今後としては、poincare embeddings がデータによっては低次元で十分精度を出せることに期待して、今回の手法で単語ベクトルの次元数を下げた場合に、維持されるかどうかを確認したいと考えている。

参考文献

[１] Mekala, Dheeraj and Gupta, Vivek and Paranjape, Bhargavi and Karnick, Harish, "SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations", Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017

[２] Vivek Gupta Harish Karnick Ashendra Bansal Pradhuman Jhala, "Product classification in ecommerce using distributional semantics" In Proceedings of COLING, 2016

[３] M. Nickel and D. Kiela, “Poincar´e embeddings for learning hierarchical representations,” inAdvances in Neural Information Processing Systems 30, I. Guyon, U. V. Luxburg, S. Bengio,H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp.6338 - 6347. [Online]. Available: http://papers.nips.cc/paper/7213-poincare-embeddings-for-learning-hierarchical-representations.pdf

[４] A. Tifrea, G. B´ecigneul, and O. Ganea, “Poincar´e glove: Hyperbolic word embeddings,” CoRR, vol.abs/1810.06546, 2018. [Online]. Available: http://arxiv.org/abs/1810.06546

[５] M. Leimeister and B. J. Wilson, “Skip-gram word embeddings in hyperbolic space,” CoRR, vol.abs/1809.01498, 2018. [Online]. Available: http://arxiv.org/abs/1809.01498

[６] B. Dhingra, C. Shallue, M. Norouzi, A. Dai, and G. Dahl, “Embedding text in hyperbolic spaces,”01 2018, pp. 59–69.

[７] B.Chen, X.Huang, L.Xiao, Z.Cai, and L.Jing, “Hyperbolic interaction model forhierarchical multi-label classification,” CoRR, vol. abs/1905.10802, 2019. [Online]. Available:http://arxiv.org/abs/1905.10802

[８] Ungar, Abraham, "A Gyrovector Space Approach to Hyperbolic Geometry", Morgan & Claypool Publishers, "Synthesis Lectures on Mathematics and Statistics", 2009

[９] H. Cho, B. DeMeo, J. Peng, and B. Berger, “Large-margin classification in hyperbolic space,”in Proceedings of Machine Learning Research, ser. Proceedings of Machine Learning Research,K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 1832–1840. [Online].Available: http://proceedings.mlr.press/v89/cho19a.html

[１０] A. Agibetov, G. Dorffner, and M. Samwald, “Using hyperbolic large-margin classifiers for biologicallink prediction,” in Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5).Macau, China: Association for Computational Linguistics, Aug. 2019, pp.26–30.[Online].Available: https://www.aclweb.org/anthology/W19-5805

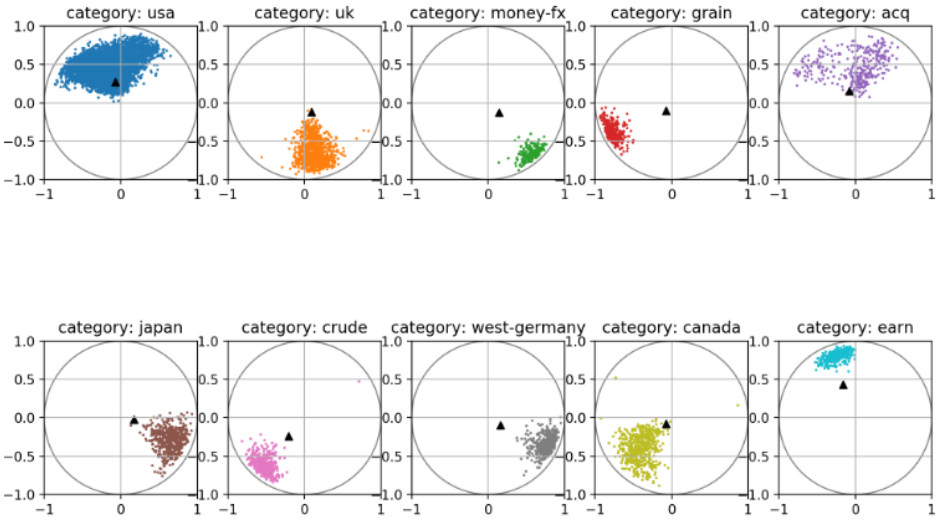

図 1　Reuter21578 上位 10 カテゴリでの 2 次元 poincare disc 上での各カテゴリの特徴語の分布。▲がカテゴリベクトルを、そのほかの点が各カテゴリにのみ出現した単語の単語ベクトルを表す。

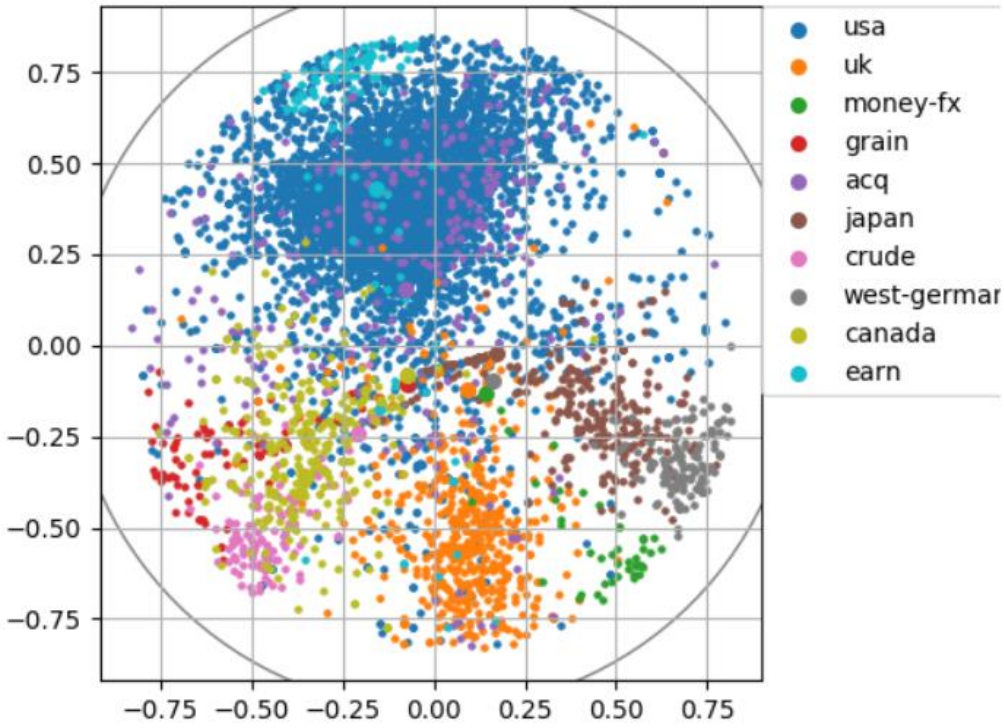

図 2　Reuter21578 上位 10 カテゴリでの 2 次元 poincare disc 上での文書ベクトルの分布。中心付近の大きい丸は各カテゴリベクトルを、そのほかの小さい点は単語ベクトルから作成された各文書ベクトルを表す。

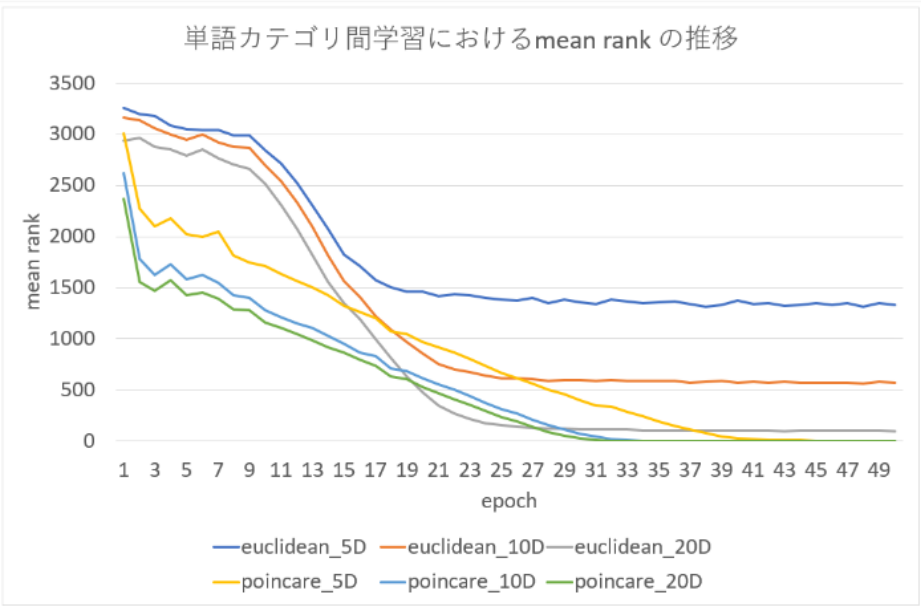

図 3　単語･カテゴリ間の学習時の mean rank の比較。5,10,20 次元でのベクトル化にあたって、ポアンカレモデル、ユークリッドそれぞれで定義される距離を用いた場合の mean rank の推移を表す。