

令和元年6月18日現在

機関番号：14603

研究種目：基盤研究(B) (一般)

研究期間：2015～2018

課題番号：15H02749

研究課題名(和文) データ中のハブを考慮した距離尺度構成法の新展開

研究課題名(英文) New development in hubness-aware metric learning in high dimensional data

研究代表者

新保 仁 (SHIMBO, Masashi)

奈良先端科学技術大学院大学・先端科学技術研究科・准教授

研究者番号：90311589

交付決定額(研究期間全体)：(直接経費) 12,400,000円

研究成果の概要(和文)：高次元空間における距離・類似度尺度の設計について理論的分析を行い、尺度の頑健性を確保するための方法を検討した。また、それらの実問題への応用を試みた。なかでも特に、高次元空間におけるハブの出現現象に着目した。主に、(i) 複数の高次元ベクトル空間におけるオブジェクト対応付け問題に関する分析と解法の提案 (ii) 上記(i)の成果を応用した、単一空間におけるk近傍分類の簡易な精度向上法の提案。(iii) 知識グラフのベクトル空間への埋込みの研究(特に複素空間への埋め込み法の分析・精度向上) という3分野で大きな成果を得た。

研究成果の学術的意義や社会的意義

成果の概要で述べた貢献(i)については、従来の手法の欠点を理論的分析に基づいて明らかにし、単純だが効果的な手法を提案した。従来法と計算コストは変わらないが、精度を高めることができる。(ii)は(i)の応用であるが、やはり簡単かつ効率的であり、距離学習に比較してはるかに小さな計算量で、精度向上が可能である。(iii)は知識グラフ埋め込みに関する研究である。最近、知識グラフを有効活用した自然言語処理での様々な応用が試みられており、われわれの分析および精度向上は、これら実応用の精度向上の一助になると考えている。

研究成果の概要(英文)：A theoretical and empirical investigation was carried out on the similarity/dissimilarity measures in high-dimensional space. In particular, a special focus was placed on analyzing the hubness phenomenon. On the basis of this analysis, we have proposed approaches for several applications. Our major contributions are (i) the analysis of regression-based object-matching over two domains, and the proposal of a simple yet effective method derived from the results of the analysis; (ii) application of our analysis for (i) to improving (single-domain) k-nearest neighbor classification; and (iii) analysis of knowledge graph embedding methods (in particular those embed objects in complex-valued space).

研究分野：機械学習

キーワード：ベクトル空間モデル 高次元空間 近傍検索 表現学習 知識グラフ 多変量回帰

1 研究開始当初の背景

最近のニューラルネットワークの有効性の再発見により、それと親和性の高いベクトル空間モデル(ベクトル空間上の点として、さまざまなオブジェクトを表現し、ベクトル演算によってオブジェクト間の近さ・類似度を捉えることを目的とする)の研究が盛んになっている。このうち、もっとも基本的であり、主流として用いられる演算は、高次元ベクトル空間における距離や内積の計量である。さらに、それら演算に基づく近傍検索(nearest neighbor search)は、多くの処理のために不可欠な技術である。

2 研究の目的

高次元空間における距離・類似度尺度の設計について理論的分析を行い、尺度の頑健性を確保するための方法を確立するとともに、それらの実問題への応用を試みる。

3 研究の方法

研究目的の達成にあたっては、特に、高次元空間において近傍検索が困難な要因として、データセット中に「ハブ(hub)」事例が出現することがある点に着目する。この影響を受けにくい距離尺度・手法を作るために、理論的な分析およびそれに基づく設計指針が必要になるが、本研究では、この分野における研究代表者らのこれまでの研究を発展させる。

適用する応用問題としては、自然言語や各種グラフデータのベクトル埋め込みタスクを用いる。

4 研究成果

本研究によって得られた主要な成果を以下に要約する。

(1) 複数の高次元ベクトル空間におけるオブジェクト対応付け問題に関する分析と解法

高次元空間における「ハブの出現」現象は、最近になって発見された新しい「次元の呪い」であり、高次元空間上のデータセットで事例間の距離を測ると、ごく少数の事例(「ハブ」と呼ばれる)が多くの他の事例の最近傍となる、という現象である。「次元の呪い」と言えば、距離集中といった現象が有名であるが、それとは異なる現象であり、近傍検索の妨げになることから注目を集めている。

われわれは、異なる(高次元)ベクトル空間の間の対応付け問題においても、ハブは無視できない問題であることを示した。

この問題では、二つの空間(以下、 X と Y と書く)のオブジェクトの対応を取るために、双方のオブジェクトを同一の空間上に写像する必要がある。一旦、同一の空間に配置されれば、その空間上の近傍検索によって、対応を取ることができるからである。対応を取りたい方向が X から Y に固定されている場合、同一の空間の取り方として三通りが考えられる。

- (1) 順方向($X \rightarrow Y$)の写像を求め、その写像を用いて全てのオブジェクトを Y に配置する。
- (2) 逆方向($Y \rightarrow X$)の写像を求め、その写像を用いて全てのオブジェクトを X に配置する。
- (3) X, Y の全てのオブジェクトの、別の共通空間 Z への写像($X \rightarrow Z, Y \rightarrow Z$)を求め、その写像を用いて全てのオブジェクトを Z に配置する。

われわれは、このうち手法2が最も良い方法であることを理論的に解明した。具体的には、手法2が最もハブを作りにくい方法であり、一見自然に見える(実際先行研究でもつばら使われていた)手法1は、逆にハブの形成を促進する手法である、ということがわかった。

本研究は、内容を高く評価され、欧州の機械学習に関する国際学会 ECML PKDD 2015 において Best Student Paper Runner-Up を受賞した。

(2) k 近傍分類の簡易な精度向上法

上記の成果は、異なるベクトル空間上の対応付け問題に関するものであるが、その分析結果は、単一空間上の分類タスクにも応用できる。すなわち、単純な線形変換を用いて、近傍検索の対象となる点を予め移動しておくだけで、 k 近傍分類の精度が向上することを発見した。これまで k 近傍分類の精度向上には、距離学習 (metric learning) を用いるのが有効とされていたが、学習時間が膨大であることが欠点であった。これに対して、われわれの提案手法は単なる線形写像を求めるだけであり、計算時間は遙かに小さい。また、実験によると、われわれの手法は、距離学習の分類精度を凌駕した。

(3) 知識グラフのベクトル空間への埋込みに関する研究

近年、知識をネットワークとして表した「知識グラフ」のベクトル埋め込みに基づく手法が注目を集めている。これらの手法は、ベクトル空間上に、エンティティ (事物) に加え、それらの間の二項関係までも埋め込み、ベクトル演算を用いて、未知の関係の存在を予測する。

とくに、われわれは複素空間への埋め込み法 (ComplEx) に着目し、複素空間における類似度の持つ性質、他手法との関連性の分析、および、その改善のための研究を行った。ここでの主要な成果は、複素空間への埋め込み法が、実数空間でベクトル間の相互相関 (畳み込み演算と深いつながりがある) を用いる別の埋め込み法 (HolE) と実質的に等価であることを、フーリエ変換の双対性に基いて示したことである。また、同時に、二つの埋込空間 (フーリエ変換の時間領域・周波数領域に対応する) における内積が保存されることから、近傍検索を行う際のハブの出現といった現象は複素空間においても存在することが示唆される。

上記を含め、この分野でのわれわれの成果は、2017 年から 2018 年にかけて、ACL 2017, IJCAI 2017, AAAI 2018, といった人工知能・自然言語処理の難関国際会議に採択された。

5 主な発表論文等

[雑誌論文] (計 3 件)

(1) Graph neural network を用いた未知エンティティの表現獲得について。

濱口 拓男, 大岩 秀和, 新保 仁, 松本 裕治。

人工知能学会論文誌, Vol. 33, No.2, pp. F-H72_1-10. 2018. DOI: 10.1527/tjsai.F-H72. 査読あり。

(2) Protein fold recognition with representation learning and long short-term memory.

Masashi Tsubaki, Masashi Shimbo, and Yuji Matsumoto.

IPSJ Transactions on Bioinformatics, Vol. 10, pp. 2–8. 2017. DOI: 10.2197/ipsjtbio.10.2. 査読あり。

(3) 意味構成のための非線形類似度学習。

椿 真史, 新保 仁, 松本 裕治。

人工知能学会論文誌, Vol. 31, No.2, pp. O-FA2_1–10. 2016. DOI: 10.1527/tjsai.O-FA2. 人工知能学会 2016 年度論文賞. 査読あり。

[学会発表] (計 8 件)

(1) Ranking-based automatic seed selection and noise reduction for weakly supervised relation extraction.

Van-Thuy Phi, Joan Santoso, Masashi Shimbo, and Yuji Matsumoto.

In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL '18), Vol. 2: Short Papers, Melbourne, Australia. 2018. pp. 89–95. 査読あり。

- (2) Citation recommendation using distributed representation of discourse facets in scientific articles.
Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto.
In *Proceedings of The ACM/IEEE Joint Conference on Digital Libraries (JCDL '18)*, pp. 243–251, 2018. 査読あり.
- (3) Data-dependent learning of symmetric/anti-symmetric relations for knowledge base completion.
Hitoshi Manabe, Katsuhiko Hayashi, and Masashi Shimbo.
In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI '18)*, pp. 3754–3761, 2018. 査読あり.
- (4) Knowledge transfer for out-of-knowledge-base entities: a graph neural network approach.
Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto.
In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI '17)*, pp. 1802–1808, 2017. 査読あり.
- (5) On the equivalence of holographic and complex embeddings for link prediction.
Katsuhiko Hayashi, and Masashi Shimbo.
In *Proceedings of the 47th Annual Meetings of the Association for Computational Linguistics (ACL '17), Vol. 2: Short Papers*, pp. 554–559, 2017. 査読あり.
- (6) A fast and easy regression technique for k -NN classification without using negative pairs.
Yutaro Shigeto, Masashi Shimbo, and Yuji Matsumoto.
In *Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '17), Part 1*, pp. 17-29, Lecture Notes in Computer Science, Vol. 10234, Springer. 査読あり.
- (7) Ridge regression, hubness, and zero-shot learning.
Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto.
In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD '15)*, pp. 135–151, 2015. Best student paper runner-up 受賞—学生が第一著者の Scientific Track 論文 (採録 64 件, 投稿 260 件) 中 2–4 位に該当. 査読あり.
- (8) Coordination-aware dependency parsing (preliminary report).
Akifumi Yoshimoto, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto.
In *Proceedings of the 14th International Conference on Parsing Technologies (IWPT '15)*, pp. 66–70, 2015. 査読あり.

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。