様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２９ 年　 ６ 月 １３ 日現在

機関番号：１１５０１
研究種目：若手研究(B)
研究期間：2013～2016
課題番号：２５７３０１４２

研究課題名（和文）大規模高次元データの近傍検索・分類に適した類似度尺度の研究

研究課題名（英文）Similarity Measures for Nearest Neighbor Search and Classification Methods in High Dimensional and Large Number of Data

研究代表者
鈴木　郁美（Suzuki, Ikumi）

山形大学・大学院理工学研究科・助教

研究者番号：２０６３７７３０

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：データが高次元であるとき「次元の呪い」と言われる現象が起こる．ごく最近新たな現象として，データ中心に近い事例が，次元が高くなると非常に多くの事例と距離が近くなる現象が報告された．この事例はハブと呼ばれ，ハブは他の事例のk 近傍に頻出するため，近傍情報を利用した分類や検索，グラフ構築に悪影響を及ぼす．生命情報学分野における塩基配列データや文書（PubMed）をはじめ，大規模高次元データは増える一方であるが，活用法は十分に開発されていない．本研究では，大規模高次元データの問題の一面であるハブに注目し，類似度・距離尺度を工夫することで，分類・検索の改善を行った．

研究成果の概要（英文）：Recently, hubness, a phenomenon occurring in high-dimensional datasets as a result of curse of dimensionality has attracted the attention of researchers in the artificial intelligence community, especially for data mining and machine learning.
In this work, we pointed out that the hubness influences the performance of k-nearest neighbor (k-NN) methods. We reported that subtracting mean vector from each sample (centering) is a simple, yet very effective for improving k-NN classification. Also, we proved that centering is effective for k-NNs, because centering reduces hubs in a dataset.

研究分野：統計的機械学習

キーワード：ハブネス　ハブの軽減　センタリング　近傍法

１．研究開始当初の背景

事例（文書データなど）は，事例の特徴を要素として持つ，特徴空間上のベクトルとして表現される方法がデータマイニングや機械学習分野で広く活用されている．特徴の数が多い（つまり，ベクトルが高次元になる）ほど，事例を表現する情報量が豊かになり，データの解析精度も高まるように思われるが，良いことばかりではない．

高次元空間では，我々が低次元空間での理解がそのまま通じる訳ではなく，「次元の呪い」として知られる現象が起こる．例えば，空間の縁にデータが集中する現象は，次元の呪いの一つとして以前から知られていたが，最近，新たな次元の呪いとして，高次元データにはハブが出現する現象が報告された．ハブは，データ中心（セントロイド）に距離が近い/類似度が高い事例であるために，高次元で多くの事例と距離が近くなる/類似度が高くなる事例である．図１に，次元が上がるだけで多くの事例と類似度が高くなるハブが出現する例を人工データで示す．

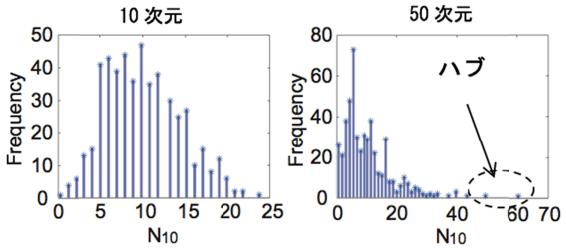

図１：次元が上がるとハブが出現する様子を，人工データ（事例数＝500）を用いて示す．各々の事例について，他の事例の近傍（ここでは，類似度の高い 10 個の事例を近傍と定義する）に入る回数（N_{10}）を調べ，ヒストグラムを作成した．低次元（10 次元，左図）の場合，他のデータの近傍に入る回数は多くても 25 回以下であるのに対し，高次元（50 次元，右図）になると，他の事例の近傍に入る回数が 60 回にもなる事例(ハブ)が現れる．

ハブの出現は，高次元データの検索や分類，グラフ構築に対し，望ましくない結果をもたらす．

- ✓ 検索では，問い合わせクエリに関わらず，検索結果の上位にいつも同じ事例(ハブ)が登場する，
- ✓ 分類では，様々な機械学習のタスクに用いられている k 近傍法(類似度の高い k 個の事例が属するクラスに基づいて分類などを行う方法）を適用するとき，分類対象のテスト事例に関わらず，ハブが属するクラスに分類されやすくなる，
- ✓ グラフ構築を行うとき，近傍にある事例をエッジで繋ぐ k 近傍グラフでは，データ中心に近いハブに多くの事例がエッジをはり，本来似ている事例同士を繋ぐ目的が果たされない．

また，ハブをデータセットから取り除いても，別の事例が新たにハブとなるため，問題は解決されない．よって，簡単に解決できる問題ではなく，本質的にハブを減らす方法を考える必要がある．

２．研究の目的

生命情報学分野における塩基配列データや文書（PubMed）をはじめ，大規模高次元データは増える一方であるが，活用法は十分に開発されていない．本研究では，大規模高次元データの問題の一面であるハブに注目し，距離尺度および類似度を工夫することで，分類・検索の改善を目指す．

３．研究の方法

課題①「セントロイドとの類似度が全データで一定になること」を満たす類似度の理論的背景の調査

「セントロイドとの類似度を全データに対して一定にする」条件を満たすラプラシアンベースのカーネルが，ハブの出現を抑えることを発見した．まず，この成果を追加実験を加え学術論文誌に投稿する．我々はまた，類似度行列をセンタリングする，すなわち，特徴空間上でベクトルの原点をセントロイドに置き換えることによっても，上記条件が満たされることに着目し，ハブの出現が抑えられることを確認した．国際会議，論文投稿での発表を行う．さらに，センタリング法の理論的背景を調べ，それをラプラシアンカーネルのパラメタ調整によるハブの出現度合の変化との関連を調査することより，理論的背景を深め，より一般性の高い理解を求める．

課題②「セントロイドとの類似度が全データで一定になること」以外にも存在すると考えられる，ハブの出現を減らす条件を見つける．

課題③ データセットの次元数及び事例数と，ハブの出現/解消との関係の調査

リンク解析の観点を手掛かりにデータ数とハブの関係を解明する．von Luxburg らは，ラプラシアンベースカーネルを元にした通勤時間距離について，事例数が大きいと不自然な距離尺度を与えることをリンク解析の観点から指摘した．一方 Radovanović らは，ハブの出現と次元数に関しての考察は行っ

たものの，データ数との関連には言及していない．しかし，我々が行った人工データを用いた予備実験の結果，ハブの出現は，データセットの次元数のみならず，事例数とも関係があることが判明した．このことから，ハブの出現/解消にはデータ数も重要な要因であると考えられ，リンク解析の観点を手掛かりに，事例数とハブの関係を解明する．

課題④：データの疎/密（sparse/dense）と，ハブの出現/解消との関係の調査
　ラプラシアンベースのカーネルが類似度尺度として有効となるデータセットの性質を解明する．ラプラシアンベースのカーネルは機械学習の分野でその有効性が知られ，よく使われる．しかし，我々の実験から，データセットの性質(sparse/dense)が異なると，ラプラシアンベースのカーネルにより，ハブの出現を軽減できる度合いが違うことがわかった．
データセットが疎/密かは，事前に簡単に調べることができる．よって，ラプラシアンカーネルが類似度尺度として，どのようなデータセットに対して有効となるのかを解明できれば，複雑なアルゴリズムやモデルに組み込む前に，その有効性を予め予測できる．

４．研究成果

課題①に関して，類似度尺度を「centering」することにより，ハブを減らすことができるか，理論的な説明を行い，国際会議 EMNLP 2013 にて，「Centering Similarity Measures to Reduce Hubs」 として発表を行った．

課題②に関して，これまで検討を行ってきた (global) セントロイド以外にも，local セントロイドとの関係でハブの出現と関係があることがわかった．また，local セントロイドとの関係から，類似度尺度を localized centering を行うことで，ハブを軽減できることがわかり，国際会議 AAAI 2015 にて，「Localized Centering: Reducing Hubness in Large-Sample Data」 として発表を行った．

課題③に関して，ハブの出現には，データセットの次元数のみならず，事例数も関係することがわかった．すなわち，次元数が小さくても，事例数が大きいとハブは出現する．なお，このようなハブについて，我々がこれまで報告してきた global centering ではハブを軽減できなかった．そこで，課題②で提案した localized centering では，ハブを軽減できることがわかり，国際会議 AAAI 2015 にて，「Localized Centering: Reducing Hubness in Large-Sample Data」として発表を行った．

課題④に関して，「データの density gradient」の観点からハブの出現について検討を行った．
本研究では，density gradient を考慮することで，距離尺度に関して，ハブを軽減する方法の提案をした．そして，提案手法によりハブが軽減できる理由を解析的に示した．すなわち，我々の手法を距離尺度に適用することで，density gradient を一定にし，ハブを軽減できることを解析的に示した．
成果は，AAAI 2016 にて「Flattening the Density Gradient for Eliminating Spatial Centrality to Reduce Hubness」で報告した．

５．主な発表論文等

〔雑誌論文〕（計　1 件）

① ラプラシアンカーネルによるハブの解消．鈴木 郁美，原 一夫，新保 仁，松本 裕治．人工知能学会論文誌，28(3)，pp. 297-310，2013．査読あり．

〔学会発表〕（計　5 件）

① Flattening the Density Gradient for Eliminating Spatial Centrality to Reduce Hubness Kazuo Hara，Ikumi Suzuki，Kei Kobayashi，Kenji Fukumizu and Miloš Radovanović．In Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)，pp.1659-1665，Arizona，Phoenix，USA，February 12-17，2016.査読あり（採択率 26%）．

② Reducing Hubness for Kernel Regression. Kazuo Hara，Ikumi Suzuki，Kei Kobayashi，Kenji Fukumizu and Miloš Radovanović.
In Proc. the 8th International Conference on Similarity Search and Applications (SISAP ’15)，pp.339-344，Glasgow，UK，October 12-14，2015．査読あり．

③ Reducing Hubness: A Cause of Vulnerability in Recommender Systems. Kazuo Hara，Ikumi Suzuki，Kei Kobayashi and Kenji Fukumizu.
In Proc. the 38th Annual ACM SIGIR Conference (SIGIR ’15)，pp. 815-818，Santiago de Chile，August 9-13，2015．査読あり（採択率 31%）．

④ Localized Centering: Reducing Hubness in Large-Sample Data. Kazuo Hara, Ikumi Suzuki, Masashi Shimbo, Kei Kobayashi, Kenji Fukumizu and Miloš Radovanović. In Proc. the 29th AAAI Conference on Artificial Intelligence (AAAI'15), pp. 2645-2651, Texas, Austin, USA, January 25-30, 2015. 査読あり（採択率 27%）.

⑤ Centering Similarity Measures to Reduce Hubs. Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, Marco Saerens and Kenji Fukumizu.In Proc. the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13), pp. 613-623, Seattle, USA, October 18-21, 2013. 査読あり（採択率 28%）.

６．研究組織

(1)研究代表者

鈴木　郁美（SUZUKI, Ikumi）
山形大学・大学院理工学研究科・助教
研究者番号：20637730