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

令和　３ 年　 ６ 月　 ５ 日現在

機関番号：１１５０１
研究種目：基盤研究(B)（一般）
研究期間：2016～2019
課題番号：１６Ｈ０２８２１

研究課題名（和文）大規模データセットに生じるハブ現象の解明とその医療生命系データへの応用

研究課題名（英文）Elucidation of hub phenomenon occurring in large-scale data and its application to bio-medical data

研究代表者
原　一夫（Hara, Kazuo）

山形大学・理学部・准教授

研究者番号：３０４６７６９１

交付決定額（研究期間全体）：（直接経費）　　13,100,000円

研究成果の概要（和文）：データ密度が一様になるようデータ間の距離を変換することによりハブの出現を抑制する方法を開発した．特に，グラフベースの半教師あり学習において重要なグラフ構築方法として，ハブがない，かつ，過剰にエッジの数を減らさないで済む方法を提案した．さらに，ハブネスと呼ばれる現象が，バイオ配列データのデータセットにおいても生じているか（特定の配列が，他の多くの配列と類似するという現象が起きているか）を調べたところ，ハブネスが生じていることを確認できた．

研究成果の学術的意義や社会的意義
Radovanovic et al. JMLR 2010によって「（グラフ上ではなく）空間上のハブ」に起因する問題が提起されて以来，ハブを解消または利用する方法（例えば新たなグラフィカルモデルやクラスタリング法）の開発，および，各ドメインタスクへの適用は，国際的な競争となりつつある．取り分け，音楽情報検索におけるハブを取り除く研究は，オーストリアの研究グループOFAIが世界をリードしている（Schnitzer et al. JMLR 2012）．本研究の成果は，医療生命系データにおけるハブの問題を，世界に先駆けて解決する土台となるものである．

研究成果の概要（英文）：We have developed a method to suppress the appearance of hubs by converting the distance between data so that the data density is uniform. In particular, as an important graph construction method in graph-based semi-supervised learning, we proposed a method that does not have hubs and does not require an excessive reduction in the number of edges. Furthermore, when we investigated whether hubness occurred in bio-sequence data i.e., whether a specific sequence was similar to many other sequences, we confirmed that hubness occurred.

研究分野：情報学

キーワード：近傍検索

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

高次元空間では，三次元空間で生活する人間の直感に反する「次元の呪い」と呼ばれる現象が起きることが知られている．本研究のテーマである「ハブ現象」は，近年発見された「次元の呪い」の一つである（Radovanović et al. SIGIR 2010 および Radovanović et al. Journal of Machine Learning Research 2010）．「ハブ現象」は，高次元空間におけるデータセットに生じる現象である．データ（事例）の空間が高次元であるほど，他の数多くの事例と類似する事例（「ハブ」と呼ばれる事例）が出現することが発見された．さらに，データセットにおける平均的な事例がハブ事例となることもまた発見された．

ハブの存在は，次の問題を引き起こすため，データセットの類似検索の価値を低下させる．（１）ハブによる検索結果上位の独占：クエリによらず検索結果の上位を少数のハブが占めること，すなわち，検索結果の上位にハブとなる事例がいわばスパムのように出現すること．（２）検索結果上位に現れにくい事例数の増大：（ハブとなる事例とは反対に）ほとんどアクセスされない事例の数が増大すること．

実際，医療生命系データセットにおいて，上記ハブの問題が生じている．例えば，アミノ酸配列のデータセットである CATH95（事例数 11373；Pearl et al. Nucleic Acids Research 2005）を調べると，検索結果の上位 100 位以内に入る事例の半分は，約 13%の事例（1480 事例）で占められる．病気の診断に用いるマイクロアレイのデータセットである Kent Ridge Biomedical Dataset においても，同様の問題が生じる．また，音楽，画像，テキストのデータセットにおいても，上記のハブの問題が生じることを，Schnitzer et al. ECIR 2014 は報告している．

２．研究の目的

ハブ現象の発生を抑制する既存手法（たとえば，Schnitzer et al. Journal of Machine Learning Research 2012）はいずれも，以下の欠点を持っている．（１）ハブ現象の発生のメカニズムに迫ることなく，既存手法は対症療法的なものに留まっている．（２）タスクによっては，ハブは都合の良い役割を果たすにもかかわらず，既存手法はハブとなる事例をすべて消去しようとする．

すなわち，既存手法には大幅な改良の余地がある．実際，どの既存手法を用いても，コントロールできないハブが生じるデータセットが存在する．また，ハブはデータセットにおける平均的な事例（典型事例）であるため，典型事例が役に立つタスク（たとえば，2 クラス分類など，粒度の粗い分類タスク）では必ずしも悪者ではない．

さらに，既存研究の成果のほとんどは，データは，ベクトル空間上の点として表されるベクトル値データであること，という前提の上に成り立つ限定的なものであった．しかし，近年，社会やビジネスにおいてその活用が期待されるビッグデータの多くは，単なるベクトル値データではなく，時系列データ，空間データのような，構造を持ったデータである．

本研究は，構造を持ったデータのなかでも，高齢化社会を迎える我が国にとって重要度の高い，医療生命系分野のビッグデータに生じるハブをコントロールする方法を開発する．

３．研究の方法

ハブ抑制のためのこれまでの手法は，ハブを発生させる原因として，「データ間の近傍関係が対称でないこと」を指摘し，近傍関係を対称化することによってハブを抑制してきた（Schnitzer et al. JMLR 2012）．しかし，データ間の近傍関係を非対称にする原因は何であるのかについての議論は，ほとんどされてこなかった．本研究では，データ密度の観点から，ハブが発生する原因を明らかにすることを試みた．具体的には，データ間の近傍関係が非対称となる（ゆえにハブが生じる）理由は，データセットを生成する分布の密度勾配である，すなわち，密度勾配を持つ分布（一様分布以外の分布）から生成されたデータセットにはハブが生じる，とわれわれは考えた．

これを説明するための例として，図１，図２は密度勾配を持つ正規分布，および，密度勾配を持たない一様分布から生成したデータセットに関して，最近傍関係を有向枝で表した図である．図中の数字は各データが他のデータの最近傍となった回数であり，この数字が大きいデータがハブである．

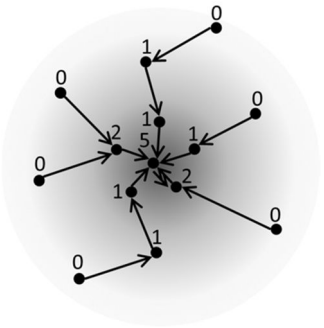

図１：正規分布

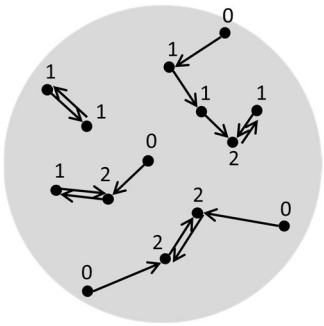

図２：一様分布

正規分布の場合（図１）は，中心に近いほど高い密度を持つため，中心付近に多数のデータが生成される．このことは，各データの最近傍として選ばれやすいのは，（各データから見て）中心から遠いデータではなく，中心に近いデータであることを意味する．つまり，有向枝は中心に向かう方向を持ちやすく，このため，中心付近のデータには多数の有向枝が集まりやすくなる（近傍関係が非対称となる）．その結果として，中心付近のデータ，すなわち，密度が高い場所にあるデータはハブになりやすい．他方，密度に濃淡がない一様分布の場合（図２）は，有向枝の方向はランダムである．よって，多数の有向枝を集めるようなデータ，すなわち，ハブは出現しにくいと考えられる．

４．研究成果

データ密度が一様になるようデータ間の距離を変換することによりハブの出現を抑制する方法を開発した．特に，グラフベースの半教師あり学習において重要なグラフ構築方法として，ハブがない，かつ，過剰にエッジの数を減らさないで済む方法を提案した．バイオ配列データについては，RefSeq と呼ばれる核酸データのデータベースから，７つの生物種の合計約１３万のアミノ酸配列（mRNA 配列）を利用した．これらの配列について，総当たりで（blastx を用いて）類似スコアを計算した．ハブネスと呼ばれる現象が，バイオ配列データのデータセットにおいても生じているか（特定の配列が，他の多くの配列と類似するという現象が起きているか）を調べたところ，ハブネスが生じていることを確認できた．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Suzuki Ikumi、Hara Kazuo	４．巻 SIGIR '17 Proceedings
２．論文標題 Centered kNN Graph for Semi-Supervised Learning	５．発行年 2017年
３．雑誌名 SIGIR '17 Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval	６．最初と最後の頁 857-860
掲載論文のDOI（デジタルオブジェクト識別子） 10.1145/3077136.3080662	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	鈴木　郁美 (Suzuki Ikumi) (20637730)	長崎大学・情報データ科学部・准教授 (17301)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関