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 １４ 日現在

機関番号：１７３０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2020
課題番号：１６Ｋ００００６

研究課題名（和文）ハブネスの数理基盤

研究課題名（英文）Hubness Analysis

研究代表者
鈴木　郁美（SUZUKI, Ikumi）

長崎大学・情報データ科学部・准教授

研究者番号：２０６３７７３０

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：ビックデータが大きな注目を集める中，大量データから欲しい情報にたどり着くのは未だ難しい課題である．本研究は，大規模高次元データ一面であるハブネス現象に着目し，欲しい情報にたどり着くための，より頑健な数理基盤の確立を目指す．本研究では，研究代表者のこれまでの 研究をより深化させ，特にハブネスの発生原因である次元とデータの大規模性について数理的に解析を行い，関係を明らかにする．

研究成果の学術的意義や社会的意義
大規模高次元データは機械学習や統計科学的に重要な対象であり，低次元の場合とは異る様相を見せることが知られているが，その理解はまったく十分ではない．特に新しい概念であるハブの問題は，高次元空間でハブが出現し問題となる，という指摘に留まっており，特にその解決法は，申請者らの研究以外世界的にもあまり見当たらない．データが増えるにつれ，欲しい情報にたどり着くことは難しい．その要因の一つであるハブの存在が悪影響を与えていると考えられ，妥当な検索結果を得ることができないという障害が一因となっていると考えられる．一般的な大規模高次元データに対する分類・検索などの様々な応用につながるため，実用的な意義は大きい

研究成果の概要（英文）：Lately Big-data gain a lot of attention, however it is still hard problem to get to information we need. In this research, we focus on one of the curse of dimensionality problem, hubness, to establish firm and robust method. We profoundly extend our original hubness reduction method to understand the behind mechanism of hubness.

研究分野：情報科学

キーワード：近傍法　ハブネス　半教師あり学習

１．研究開始当初の背景

事例を，その特徴を要素として持つベクトルとして表現し、特徴空間上で扱う方法が，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．

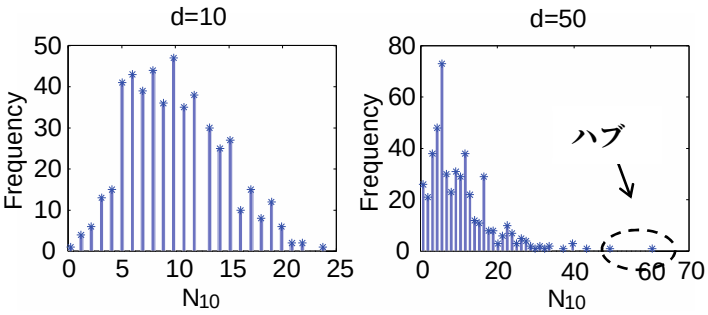

図１：次元が上がるとハブが出現する様子を，人工データ（事例数＝500）を用いて示す．各々の事例について，他の事例の近傍（ここでは，類似度の高い 10 個の事例を近傍と定義する）に入る回数（N_{10}）を調べ，ヒストグラムを作成した．低次元（10 次元，左図）の場合，他のデータの近傍に入る回数は多くても 25 回以下であるのに対し，高次元（50 次元，右図）になると，他の事例の近傍に入る回数が 60 回にもなる事例(ハブ)が現れる．

高次元空間では，「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ハブネスの現象が報告された．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，高次元空間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．

ハブの出現は，非常に多くの事例の近傍に出現するため，近傍情報を利用した高次元データの検索や分類に対し，望ましくない結果をもたらす．実際，

- 商品推薦システムでは，ハブとなる商品が推薦されてしまう（Knees et al. ICMR 2014）
- 音楽検索をする際，常に同じ音楽（ハブ）がランキングの上位に現れる(Schnitzer et al. JMLR 2012)
- 文書分類などの分類タスクにおいて，クラスラベルに関係なく多くの事例の近傍に頻出する問題(Radovanović et al. JMLR 2010)

など，多岐にわたるタスクにおいて，ハブの影響が報告されている．

２．研究の目的

本研究では，我々が開発したハブの発生を抑制する手法をより深化させ，手法開発と応用に向けて研究開発を行う．とくに，これまで開発した手法はラベル情報を利用しない方法であるため，ラベルを活用した手法に我々の方法を適用する方法を考案する．

３．研究の方法

ハブを軽減する方法として，通勤時間カーネル（データを節点とするグラフ上のランダムウォークで，節点間の往復に必要な平均ステップ数として，データ間の類似度を定めるカーネル）を用いると，データセットの中心（重心）との類似度がすべてのデータに関して等しくなるため，ハブが軽減される(Suzuki et al. AAAI 2012)．

一方，グラフを用いた半教師あり学習におけるハブを削減した方法の効果について知られていない．そこで，我々が開発したハブ軽減手法をグラフベースの半教師あり学習を組み込み，これまでの教師なし学習による方法を発展させ，その効果の検証を行った．

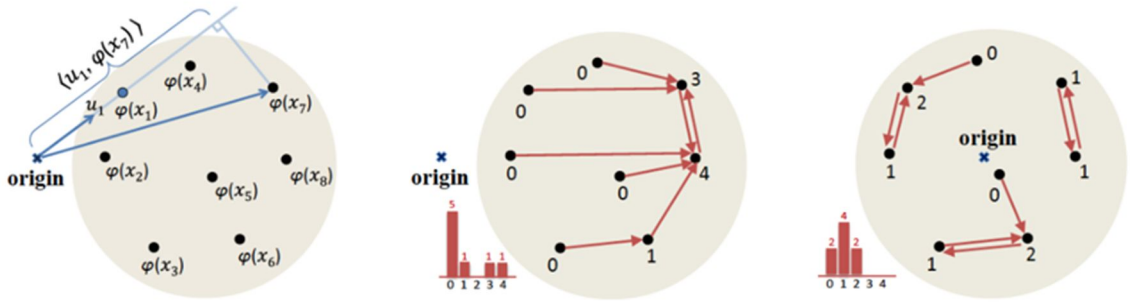

図 2.：ハブ軽減法のハブを減らす直感的な説明の図．サンプルがデータ中心付近にあるとき，その多くのサンプルの 1 近傍に入る．一方我々のハブ軽減法を適用すると，多くのサンプルの近傍に入るサンプルがなくなる．

４．研究成果

グラフベースの半教師あり学習としてグラフラプラシアンを元にしたラベル伝搬法を用いた(Zhou et al, NIPS 2003)．各サンプルのクラス予測のための行列を F，クラスラベル行列を Y，サンプル間の類似度行列を S，伝搬パラメタを α としたとき，以下のようにラベル予測を行う．

$$F = (I - \alpha S)^{-1} Y$$

この際に，類似度行列 S の与え方として，一般的にはコサイン類似度を用い，枝刈りを行う．我々はこの類似度行列 S の与え方にハブを軽減する尺度を適用した．また，適用した類似度を元にグラフのスパース化のために近傍グラフを構築した．近傍グラフの k は図３に示すように 1,5,..40 としてグラフを構築し，それぞれの精度や歪度を比較した．

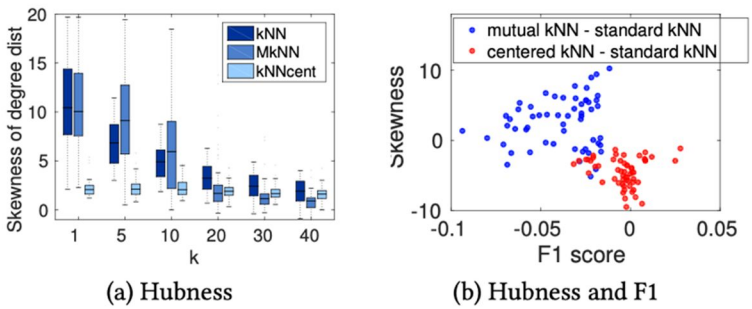

図３：(a) 近傍グラフ，相互近傍グラフ，ハブを軽減した kNN グラフのハブ度合いを頂点次数の歪度により図った結果の分布を示す．(b) ５７個の WSD データセットを用いて相互近傍グラフとハブを軽減した kNN グラフの F1 スコアと歪度を比較した散布図．

われわれの手法との比較手法として近傍グラフ，相互近傍グラフと比較を行い，我々のハブを軽減する手法を組み込んだ結果が，ハブの影響を軽減し，文書分類などの精度を向上させることがわかった．ハブ軽減法の効果の検証を行った結果，近傍グラフはハブを軽減しないため，データに存在するハブの影響をラベル推定の際に受け，相互近傍グラフは二重にスパース化われるためグラフが疎になる度合いが強いことがわかった（図３）．我々の提案手法は，ハブを軽減しつつ，グラフのスパース化を実現できると考えられる．

今後の展望としては，グラフ構築においてハブを軽減するグラフと軽減しない近傍グラフについて，高次元由来のハブを軽減する影響以外に，グラフの性質の違いが観察される．それらの性質を解明することが，グラフ構築におけるデータの高次元性の影響を理解する上で重要と考えられる．

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 Ikumi Suzuki
２．発表標題 Elimination of Spatial Centrality; Hubness in High-Dimensional Problem and Its Reduction Method
３．学会等名 Asia Pacific Society for Computing and Information Technology 2018 Annual Meeting（招待講演）（国際学会）
４．発表年 2018年

１．発表者名 Ikumi Suzuki, Kazuo Hara
２．発表標題 Centered kNN Graph for Semi-Supervised Learning
３．学会等名 SIGIR 2017（国際学会）
４．発表年 2017年

〔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

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