Vol.2012-NL-209 No.11
2012/11/22

k近傍法でハブを軽減する類似度尺度

鈴木 郁美[1,a)]　原　一夫[1,b)]　新保　仁[2,c)]

概要：k 近傍法はしばしば分類や検索に用いられる．しかし，データが高次元の場合，他のオブジェクトの k 近傍に頻出するオブジェクト (ハブと呼ばれる) が出現し，結果として k 近傍法の性能は低下する．最近，著者らは，グラフラプラシアンベースのカーネルにハブを軽減する効果があることを報告した．本稿では，より簡易な「センタリング」（原点をデータ中心に移動すること）によっても ハブが軽減できることを示す．分類タスクと検索タスクの実験で，その効果を検証する．

1. 序論

簡便な機械学習法の一つである k 近傍法は，自然言語処理タスクにも幅広く応用され，文書分類，情報検索，語義曖昧性解消等で用いられている．テキストデータは通常，高次元特徴空間上のオブジェクト (ベクトル) として表現され，オブジェクト間の類似度は，しばしばベクトルの内積（あるいはコサイン）を用いて測られる．k 近傍法は，そういった類似度を用いて，分類対象オブジェクト (分類タスクの場合) あるいはクエリオブジェクト (情報検索タスクの場合) と，データセット内のオブジェクトとの類似度を測り，最も類似度の高い k 個のオブジェクトの情報を分類や検索に活用する．

しかしながら，高次元空間上のデータセットには，「ハブ」と呼ばれるオブジェクトが出現することが知られている [10]．ハブとは，データセット内の多くのオブジェクトと類似するオブジェクトを指す．言い換えると, ハブは高頻度で他のオブジェクトの k 近傍に出現するオブジェクトということもできる．

ハブの出現は，k 近傍法に基づいて分類や検索を行うとき，性能を低下させる原因となる．

たとえば，k 近傍法に基づく分類の場合，ラベル未知のオブジェクト（テストオブジェクト）のラベルを，ラベル既知の k 個の近傍オブジェクトのラベルを手掛かりに分類する．しかし，テストオブジェクトによらず，k 近傍に常に含まれるハブオブジェクトが存在すれば，どのテストオブジェクトも，ハブオブジェクトと同じラベルを持つという分類がされやすくなってしまう．

一方，情報検索は，データベースへの問い合わせ（クエリ）に対して関連する（類似する）オブジェクトを提示するタスクである．こちらの場合も，個々のクエリに対していつも同じ (ハブ) オブジェクトが提示されるなら，検索結果は意味のあるものとは言えず，望ましくない．

よって，k 近傍法に基づいて分類や検索を行うとき，ハブの出現を抑えることが望まれる．

本研究は，データセットをセンタリングすることにより，ハブの出現を抑え，結果として，k 近傍法による分類・検索精度が改善することを示す．ここで，データセットのセンタリングは，データセットが定義される特徴空間において，原点をセントロイド（データ中心）に移動することにより，新たな類似度尺度を作ることに相当する．つまり，各々のオブジェクトを表す特徴ベクトルの原点をセントロイドに移動し，その結果得られる新しいベクトル同士の内積としてオブジェクト間の類似度を再定義し，k 近傍法で用いる新たな類似度尺度とする．

本研究の背景として，セントロイドに類似するオブジェクトがハブになるという報告がある [10]．我々は，この報告を踏まえて，セントロイドとの類似度を全てオブジェクトについて等しくするような類似度尺度を定義できれば，ハブの出現を抑えられるのではないか，と考えた．そして，グラフラプラシアンに基づくカーネルがそのような類似度尺度であり，実際にハブを減らすことについては，既に示した [11]．これに対して本稿では，データのセンタリングによって作られる新たな類似度尺度もまた，セントロイドとの類似度を全てのオブジェクトについて等しくし，ハブの出現を抑えることを示す．

また，センタリングは，大規模データに対して適用しや

[1] 国立遺伝学研究所
411-8540 静岡県三島市谷田 1111
[2] 奈良先端科学技術大学院大学
630-0192 奈良県生駒市高山町 8916-5
[a)] suzuki.ikumi@gmail.com
[b)] kazuo.hara@gmail.com
[c)] shimbo@is.naist.jp

すいという利点がある．つまり，データセットのオブジェクト数 n に対して，グラフラプラシアンをベースとするカーネルは $O(n^3)$ の逆行列計算 (あるいは固有値計算) を要するのに対し，センタリングは $O(n^2)$ で計算できる．

本論文の構成は次の通りである．2 章で，センタリングについて簡単に述べる．その後 3 章で，センタリングによってハブの出現を抑制できると我々が考える理由について説明する．そして 4 章で，分類タスク（語義曖昧性解消，マルチクラス文書分類，マルチラベル文書分類）と，検索タスク（シソーラスマッピング）の実験で，センタリングの効果を検証する．関連研究を 5 章で述べ，最後に 6 章で本稿をまとめる．

2. データセットのセンタリング

データセットとして，m 次元特徴空間上のベクトルの集合を考える．すなわち，データセットは m 次元ベクトルで表された n 個のオブジェクト $\mathbf{x}_1,\cdots,\mathbf{x}_n \in \mathbb{R}^m$ からなるとする．ベクトル間の内積 $\langle \mathbf{x}_i,\mathbf{x}_j\rangle$ により，i 番目のオブジェクトと j 番目のオブジェクトの類似度を定義し，これを (i,j) 要素としてもつ $n\times n$ の類似度行列を $\mathbf{K}$ とする．すなわち，$m\times n$ の行列 $\mathbf{X}=[\mathbf{x}_1,\cdots,\mathbf{x}_n]$ を用いて，

$$\mathbf{K}=\mathbf{X}^{\mathrm{T}}\mathbf{X} \tag{1}$$

である．

このとき，本研究では，特徴空間の原点をセントロイドベクトル

$$\bar{\mathbf{x}}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

に移すことを，データセットのセンタリングと呼ぶ．すなわち，各オブジェクトのベクトル $\mathbf{x}_i$ を，センタリングしたベクトル：

$$\mathbf{x}_i^{\mathsf{cent}}=\mathbf{x}_i-\bar{\mathbf{x}}$$

に置き換えることを行う．そして，i 番目のオブジェクトと j 番目のオブジェクトの類似度を新たに $\langle \mathbf{x}_i^{\mathsf{cent}},\mathbf{x}_j^{\mathsf{cent}}\rangle$ として定義する．

2.1 セントロイドとの類似度が一定になること

センタリングしたベクトル間の内積 $\langle \mathbf{x}_i^{\mathsf{cent}},\mathbf{x}_j^{\mathsf{cent}}\rangle$ を (i,j) 要素とする $n\times n$ の新たな類似度行列 $\mathbf{K}^{\mathsf{cent}}$ は，$\mathbf{K}$ を用いて次のように計算できる．

$$\mathbf{K}^{\mathsf{cent}}=\mathbf{Z}^{\mathrm{T}}\mathbf{K}\mathbf{Z}. \tag{2}$$

ここで，

$$\mathbf{Z}=\mathbf{I}-\frac{1}{n}\mathbb{1}\mathbb{1}^{\mathrm{T}}=\frac{1}{n}\begin{bmatrix} n-1 & -1 & \cdots & -1\\ -1 & n-1 & & \\ \vdots & & \ddots & \\ -1 & -1 & \cdots & n-1\end{bmatrix}$$

であり，$\mathbf{I}$ は $n\times n$ の単位行列，$\mathbb{1}$ はすべての要素を 1 とする n 次元ベクトルである．このとき，$\mathbf{Z}\mathbb{1}=\mathbf{0}$ ゆえ，$\mathbf{K}^{\mathsf{cent}}\mathbb{1}=\mathbf{0}$ であるから，$\mathbf{K}^{\mathsf{cent}}$ は $\mathbb{1}$ を固有ベクトルとして持つ．したがって，$\mathbf{K}^{\mathsf{cent}}$ はセントロイドとの類似度を一定にする行列になっている．

また，式 (2) には，オリジナルのベクトルからなる行列 $\mathbf{X}$ が現れないことに注意する．つまり，このことは，（カーネル行列として）類似度行列 $\mathbf{K}$ だけが存在する状況，すなわち，オブジェクトのベクトル表現が明示的に与えられない場合でも，特徴空間上でデータセットをセンタリングすることにより得られる新たな類似度行列 $\mathbf{K}^{\mathsf{cent}}$ を計算できることを意味する．

なお，式 (2) は以下のように導かれる．センタリングした n 個のベクトルからなる行列 $\mathbf{X}^{\mathsf{cent}}=[\mathbf{x}_1-\bar{\mathbf{x}},\cdots,\mathbf{x}_n-\bar{\mathbf{x}}]$ を用い，

$$\mathbf{K}^{\mathsf{cent}}=(\mathbf{X}^{\mathsf{cent}})^{\mathrm{T}}\mathbf{X}^{\mathsf{cent}}$$

である．一方，

$$\mathbf{X}^{\mathsf{cent}}=\mathbf{X}\mathbf{Z}$$

と表せるから，式 (1) を用い，

$$\mathbf{K}^{\mathsf{cent}}=(\mathbf{X}\mathbf{Z})^{\mathrm{T}}\mathbf{X}\mathbf{Z}=\mathbf{Z}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{Z}=\mathbf{Z}^{\mathrm{T}}\mathbf{K}\mathbf{Z}$$

となる．

3. センタリングによりハブの発生を抑制できることの理論的考察

平均 μ，分散 σ^2 の正規分布を $N(\mu,\sigma^2)$ で表す．いま，m 次元ベクトル空間から独立同分布により生成されたオブジェクト集合 (データセット) を考える．具体的には，各オブジェクト $\mathbf{x}=[x_1,\cdots,x_m]^{\mathrm{T}}$ の要素 x_k $(k=1,\ldots,m)$ がそれぞれ，平均 μ_k，分散 σ_k^2 の正規分布から独立に生成されたとする．すなわち

$$\mathbf{x}=\begin{bmatrix}x_1\\ \vdots\\ x_m\end{bmatrix}\sim\begin{bmatrix}N(\mu_1,\sigma_1^2)\\ \vdots\\ N(\mu_m,\sigma_m^2)\end{bmatrix}. \tag{3}$$

このとき，ある特定のオブジェクト $\mathbf{x}'$ と任意のオブジェクト $\mathbf{x}$ との内積は，

$$\langle \mathbf{x}',\mathbf{x}\rangle=\sum_{i=1}^{m}x_i'x_i \tag{4}$$

であり，この値は次の正規分布に従う．

$$\langle \mathbf{x}',\mathbf{x}\rangle\sim N(\sum_{i=1}^{m}x_i'\mu_i,\sum_{i=1}^{m}(x_i')^2\sigma_i^2). \tag{5}$$

今，データセットから二つのオブジェクトを選ぶ．セントロイドとの類似度（内積）が大きいオブジェクト $\mathbf{x}^h$ と，小さいオブジェクト $\mathbf{x}^l$ である．すなわち，

$$\langle \mathbf{x}^h,\mu\rangle=\sum_{i=1}^{m}x_i^h\mu_i\geq\langle \mathbf{x}^l,\mu\rangle=\sum_{i=1}^{m}x_i^l\mu_i \tag{6}$$

である．二つのオブジェクトそれぞれについて，データセット内の他のオブジェクトとの類似度（内積）を考えると，

$$\langle \mathbf{x}^h, \mathbf{x} \rangle = \sum_{i=1}^{m} x_i^h x_i \sim N(\sum_{i=1}^{m} x_i^h \mu_i, \sum_{i=1}^{m} (x_i^h)^2 \sigma_i^2) \quad (7)$$

$$\langle \mathbf{x}^l, \mathbf{x} \rangle = \sum_{i=1}^{m} x_i^l x_i \sim N(\sum_{i=1}^{m} x_i^l \mu_i, \sum_{i=1}^{m} (x_i^l)^2 \sigma_i^2) \quad (8)$$

となり，それぞれ，平均 $\sum_i^m x_i^h \mu_i$ と $\sum_i^m x_i^l \mu_i$ の正規分布に従う．$\sum_i^m x_i^h \mu_i \geq \sum_i^m x_i^l \mu_i$ なので，セントロイドとの類似度 (内積) が大きいオブジェクト $\mathbf{x}^h$ のほうが，小さいオブジェクト $\mathbf{x}^l$ より，他のオブジェクトとの類似度（内積）が大きい傾向にあることがわかる．

一方, 元データをセンタリングした二つのオブジェクト $\mathbf{x}^h - \bar{\mathbf{x}}$，$\mathbf{x}^l - \bar{\mathbf{x}}$ と，同様にセンタリングした任意のオブジェクト $\mathbf{x} - \bar{\mathbf{x}}$ との内積は以下のようになる．

$$\langle \mathbf{x}^h - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle = \langle \mathbf{x}^h, \mathbf{x} \rangle - \langle \mathbf{x}^h, \bar{\mathbf{x}} \rangle - \langle \mathbf{x}, \bar{\mathbf{x}} \rangle + \|\mathbf{x}\|^2 \quad (9)$$

$$\langle \mathbf{x}^l - \bar{\mathbf{x}}, \mathbf{x} - \bar{\mathbf{x}} \rangle = \langle \mathbf{x}^l, \mathbf{x} \rangle - \langle \mathbf{x}^l, \bar{\mathbf{x}} \rangle - \langle \mathbf{x}, \bar{\mathbf{x}} \rangle + \|\mathbf{x}\|^2. \quad (10)$$

これら内積の分布の違いについて考るため，式 (9)，(10) 右辺の各項を比較する．まず，第 3 項（$\langle \mathbf{x}, \bar{\mathbf{x}} \rangle$）と第 4 項（$\|\mathbf{x}\|^2$）は両式に共通であり分布に違いをもたらさないため，以下では考えない．

第 1 項の $\langle \mathbf{x}^h, \mathbf{x} \rangle$，$\langle \mathbf{x}^l, \mathbf{x} \rangle$ は式 (7)，(8) より，それぞれ平均 $\sum_i^m x_i^h \mu_i$ と $\sum_i^m x_i^l \mu_i$ 正規分布に従う．一方，第 2 項（$\langle \mathbf{x}^h, \bar{\mathbf{x}} \rangle$ と $\langle \mathbf{x}^l, \bar{\mathbf{x}} \rangle$）は定数項である．ここで，サンプル平均 $\bar{\mathbf{x}}$ が真の平均に十分近い ($\bar{\mathbf{x}} \approx [\mu_1, \ldots, \mu_m]$) と仮定すると，これらの定数項は $\langle \mathbf{x}^h, \bar{\mathbf{x}} \rangle \approx \sum_i^m x_i^h \mu_i$，$\langle \mathbf{x}^l, \bar{\mathbf{x}} \rangle \approx \sum_i^m x_i^l \mu_i$，となり，第 1 項で与えられる分布（式 (7)，(8)）の平均 $\sum_i^m x_i^h \mu_i$ と $\sum_i^m x_i^l \mu_i$ に等しくなる．よって，第 1 項と第 2 項の差によって与えられる正規分布の平均はともに 0 であり一致する．

よって，センタリングした二つのオブジェクト $\mathbf{x}^h - \bar{\mathbf{x}}$，$\mathbf{x}^l - \bar{\mathbf{x}}$ とデータセット内の他のオブジェクト $\mathbf{x} - \bar{\mathbf{x}}$ との内積の分布の平均に差はない．すなわち，多くのオブジェクトと類似度（内積）が高くなるようなハブの出現が軽減されることがわかる．

4. 実験

本章では，データセットのセンタリングにより，実際にハブを解消できるかどうかについて調べ，以前 ([11]) 有効性を確認したグラフラプラシアンに基づくカーネル (正則化ラプラシアン・通勤時間カーネル) と比較する．さらに，k 近傍法に基づく分類や検索の性能が向上するかどうかについて，3 つの分類タスク（語義曖昧性解消，マルチクラス文書分類，マルチラベル文書分類）と，1 つの検索タスク（シソーラスマッピング）で調査する．

本章を通して，実験結果の表中では，ベースとなる類似度尺度（行列）を $\mathbf{K}$ と表し，これをセンタリング操作により変換した行列を $\mathbf{K}^{\mathrm{cent}}$，正則化ラプラシアンとして変換した行列を $\mathbf{L}_{\mathrm{RL}}$，通勤時間カーネルとして変換した行列を $\mathbf{L}_{\mathrm{CT}}$ と表す．なお，正則化ラプラシアン $\mathbf{L}_{\mathrm{RL}}$ は，正値のパラメタ β をゼロに近づけると（対角要素を除き）変換元となる類似度行列 $\mathbf{K}$ に，無限大に近づけると通勤時間カーネル $\mathbf{L}_{\mathrm{CT}}$ に近づくが，それら両極をカバーするように β を段階的に選び，精度が最も良くなるときの β による結果を，実験結果の表中に記した．

4.1 語義曖昧性解消

4.1.1 タスクとデータセット

語義曖昧性解消のベンチマークデータとして Senseval-3 English Lexical Sample (ELS) task [7] で使用されたデータを用いる．多義性のある 57 種類のターゲット単語（appear, argument, bank など）に対して，それぞれ約 50〜400 個の事例が与えられる．各々の事例は，文脈（数センテンスからなるパラグラフ）の中に埋め込まれている．また，各事例には 1 つまたは複数の正解語義ラベルが与えられる．このベンチマークデータは訓練用事例とテスト用事例に分けられており，タスクの目的は，文脈をヒントにテスト用事例の語義ラベルを正しく予測することである．

文脈を手掛かりに語義を予測する標準的な手順は，まず，文脈を単語の集まり（bag-of-words）と見なし，ストップワード[*1]を除く単語について，文脈内での出現頻度を tf-idf で重み付けした値を要素とするベクトルを，各々の事例に対して作成する [6], [8]．つまり，事例集合を，高次元ベクトル空間上のデータセットに作り換える．そして，語義曖昧性解消タスクを，機械学習分野での教師付き学習としての高次元データ分類問題と捉える．すなわち，何らかの分類モデルとカーネル（あるいは類似度尺度）を選択した上で，訓練用事例の語義ラベルを正しく分類できるようにモデルパラメタを学習し，分類器を作成する．そして，作成した分類器を用いて，テスト用事例の語義ラベルを予測する．

4.1.2 比較手法

語義曖昧性解消の state-of-the-art の解決方法は，分類モデルとして Support Vector Machine（SVM）や k 近傍分類法を用い，これを適切なカーネル（類似度尺度）と組み合わせ，語義ラベルを予測する方法である．代表的なカーネルとしては，線形カーネル，多項式カーネル，コサインカーネルが挙げられる．事例 i，事例 j を表すベクトルを $\mathbf{x}_i$，$\mathbf{x}_j$ とすると，事例間の類似度 $k(\mathbf{x}_i, \mathbf{x}_j)$ は，次のように計算される．線形カーネル：

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \quad (11)$$

2 次の多項式カーネル：

[*1] [4] の on-line appendix として公開されているストップワードリストを用いた．

$$k(\mathbf{x}_i,\mathbf{x}_j) = (\langle \mathbf{x}_i,\mathbf{x}_j\rangle + 1)^2, \tag{12}$$

コサインカーネル：

$$k(\mathbf{x}_i,\mathbf{x}_j) = \frac{\langle \mathbf{x}_i,\mathbf{x}_j\rangle}{\sqrt{\langle \mathbf{x}_i,\mathbf{x}_i\rangle}\sqrt{\langle \mathbf{x}_j,\mathbf{x}_j\rangle}}. \tag{13}$$

これらカーネルを類似度尺度として用い，SVM [*2] あるいは k 近傍分類法と組み合わせる方法を，本実験の比較手法とする．

なお，k 近傍分類法によるテスト用事例の語義ラベルの予測は，近傍 k 個の訓練用事例の語義ラベルを，類似度で重み付けして投票することにより行う．また，SVM のソフトマージンパラメタ C と，k 近傍分類法のパラメタ k は，訓練用事例のクロスバリデーションにより調整する．

4.1.3 提案手法

ここで，本研究の動機を思い出すと，カーネルをそのまま類似度尺度として用いると，ハブオブジェクトが存在するために k 近傍分類の精度が高くならないという状況を，ハブの出現を抑えるように類似度尺度を変換することにより，解決することであった．ハブの出現を抑えることを見込む変換方法には，著者らが [11] で報告したグラフラプラシアンベースのカーネルとして変換する方法と，本稿で新たに提案するセンタリングによる方法がある．

そこで，上記の 3 つのカーネル（線形カーネル，2 次の多項式カーネル，コサインカーネル）をグラフラプラシアンベースのカーネル（正則化ラプラシアン，通勤時間カーネル）として変換する，あるいは，センタリング操作により変換することを行った上で，k 近傍分類法と組み合わせる．そして，カーネルを変換せずにそのまま類似度尺度として用いた場合と比較する．

なお，これらの変換は，ラベル情報を必要としないため，訓練用，テスト用を問わず，全事例を用いて行う．また，本稿で提案するセンタリング操作による変換は，SVM と組み合わせた場合，分類結果を変化させないことに注意されたい．

4.1.4 評価方法

まず，上記の各種カーネル行列で与えられる類似度尺度それぞれに対して，データセット中の各オブジェクト $\mathbf{x}$ が他オブジェクトの k 近傍に出現した回数 $N_k(\mathbf{x})$ を求め，次式で与えられる N_k 分布の歪度（skewness）を計算する．

$$S_{N_k} = \frac{\mathbb{E}[N_k - \mu_{N_k}]^3}{\sigma_{N_k}^3} \tag{14}$$

ここで $\mathbb{E}[\cdot]$ は期待値を表し，μ_{N_k} と $\sigma_{N_k}^2$ はそれぞれ N_k の平均と分散である．ハブが出現すると，N_k 分布は右に歪み (skew to the right)，歪度 S_{N_k} は大きな値を持つ．言い換えると，歪度が小さいほど望ましい類似度尺度であると言える．

[*2] $SVM^{multiclass}$: http://svmlight.joachims.org/svm_multiclass.html を使用した．

表 1 語義曖昧性解消の実験結果：歪度（skewness）．値が小さいほど良い．**K** は基本となる類似度尺度をそのまま使用した場合を表し，これをセンタリング操作により変換した行列を $\mathbf{K}^{\mathrm{cent}}$，正則化ラプラシアンとして変換した行列を $\mathbf{L}_{\mathrm{RL}}$，通勤時間カーネルとして変換した行列を $\mathbf{L}_{\mathrm{CT}}$ と表す．

類似度尺度	**K**	$\mathbf{K}^{\mathrm{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$
線形カーネル	3.118	**0.838**	2.882	2.473
2 次の多項式カーネル	3.118	**0.936**	1.940	1.907
コサインカーネル	4.554	**1.193**	4.485	4.484

表 2 語義曖昧性解消の実験結果：精度（F1 スコア）．値が大きいほど良い．SVM のカラムを除き，kNN を分類モデルとして使用した．なお，SVM の結果は，類似度尺度を（変換することなく）そのまま使用したときの結果である．

類似度尺度	**K**	$\mathbf{K}^{\mathrm{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$	SVM
線形カーネル	0.612	**0.635**	0.620	0.567	0.622
2 次の多項式カーネル	0.609	**0.628**	0.625	0.571	0.610
コサインカーネル	0.602	**0.641**	0.610	0.525	0.631

なお，今回は [10] に倣い，$k = 10$ とした．

さらに，提案手法，および，比較手法による語義曖昧性解消の精度（F1 スコア）を，Senseval-3 ELS task で配布されたスコア計算のためのスクリプトを用いて計算する．

4.1.5 結果

前節で述べた歪度（skewness）および精度（F1 スコア）を，57 種類のターゲット単語の事例集合に対してそれぞれ計算し，それらの平均（精度はマクロ平均）を表 1, 表 2 に報告する．

歪度は，線形カーネル，2 次の多項式カーネル，コサインカーネルのいずれを用いた場合でも，グラフラプラシアンベースのカーネルによる変換と比較して，センタリングにより顕著に下がった．これに伴い，センタリングを kNN と組み合わせると，分類精度は大きく改善され，SVM を用いるよりも大きい F1 スコアを得る結果となった．

4.2 シソーラスマッピング

4.2.1 タスクの概要

シソーラスマッピングは，シソーラス辞書（既登録語をノードとする木構造をなしている）に未登録語（クエリ）を追加しようとするとき，追加すべきシソーラス木上の位置を推定するタスクである．クエリに対して既登録語との類似度を正確に測ることが，このタスクの精度を向上させる鍵となるが，ここでは類似度を，コーパス（文書データベース）から抽出する文脈の類似度として測る．

4.2.2 データセット

シソーラスとしてバイオ医療専門用語を集めた MeSH (2009 年版)[*3] を用い，コーパスとして生命科学の文献情報

[*3] http://www.nlm.nih.gov/mesh/2009/introduction/introduction.html

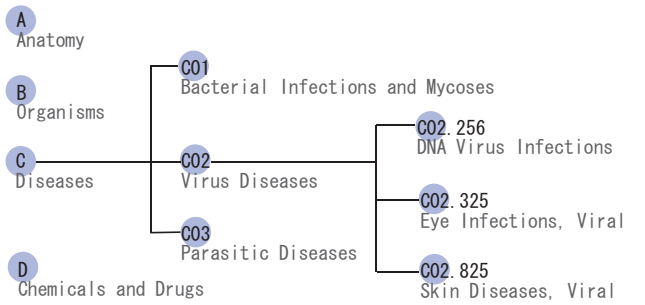

図 1 MeSH tree structure：Diseases をトップノードとする木の一部を示す．上位（C：Diseases）と比較して下位（例えば C02.256：DNA Virus Infections）になるほど抽象度の低い専門用語となる．

を収集した MEDLINE [*4]を使用した．MeSH シソーラスを説明する例として，Diseases（C）をトップノードとする木の一部を図 1 に示す．MeSH の各ノードには，1 つの専門用語がノード ID と共に割り振られている．ただし，1 つの専門用語に複数のノード（語義）が割り当てられることがあるが，ここでは語義曖昧性解消を興味の対象としないため，1 つの語義をもつ専門用語のみを実験に使用する．また，MeSH は 16 のトップノード (A, B, C,...) を持つが，本実験では登録専門用語数が多い 4 つのカテゴリ (A, B, C, D) をトップノードとする 4 つの木に属する専門用語を実験に用いる．

実験に使用する専門用語は，MeSH に登録されており，かつ MEDLINE コーパスに出現する専門用語である．その結果，専門用語オブジェクト数は，カテゴリ (A, B, C, D) 別に，それぞれ 833，2,098，1,347，1,961 となった（異なる文脈で現れても同一の専門用語は 1 つと数える）．

4.2.3 比較する手法

専門用語オブジェクト間の類似度を，文脈類似度により測るために，文脈情報を MEDLINE コーパスから抽出し，各専門用語オブジェクトを特徴ベクトルに変換する．ここで文脈情報とは，専門用語が出現する文の周辺単語のことである．各専門用語に対し，特徴ベクトルを周辺単語の bag-of-words として作成し，tf-idf による重み付けした後，ベクトルの長さを 1 に正規化する．特徴ベクトルの次元数は，カテゴリ (A, B, C, D) 別に，それぞれ 274,064，228,522，200,339，212,614 となった．この特徴ベクトルを用いてコサイン類似度行列を作り，さらに，これを元にセンタリングによる変換，および，グラフラプラシアンベースのカーネルを計算し，これらを類似度尺度として専門用語間の類似度を測った．

4.2.4 評価方法

シソーラスマッピングは，クエリとなる専門用語が与えられると，既に登録されている専門用語をクエリとの類似度の高い順にランキングするタスクと見なせる．正解専門用語をより上位にランクしてくれる類似度が，望ましい類似度である．ここで，正解専門用語とは，シソーラス木でクエリ専門用語と近い場所に位置する専門用語のことである．本実験では正解専門用語を，クエリ専門用語の親（一つ上のノード），子（一つ下のノード），兄弟（親ノードを共有し自分と同じ階層にあるノード）にあたる専門用語とする．たとえば，図 1 の Virus Dieases（C02）がクエリであるとき，C（親），C02.256，C02.325，C02.825（子）そして C01，C03（兄弟）が正解の専門用語となる．1 つのクエリに対して複数の正解があり得ることに注意する．

今回の実験では，使用する全ての専門用語のシソーラス木での位置は既知であるから，評価のために 1 つ専門用語を取り出してクエリとして用いても，そのクエリのシソーラス木での位置は分かっている．ゆえに，そのクエリにとっての正解専門用語を予め列挙しておくことができる．

よって，各々の類似度尺度に対する評価値は，それぞれの類似度を用いて出力される，クエリと類似する専門用語のランキングにおける，正解専門用語の最上位の順位とする．最上位とする理由は，1 つのクエリに対して複数の正解があり得るためである．

実験に用いる全ての専門用語を交互に一つずつクエリとして用いて，上記の過程を繰り返し行い，各クエリに対して正解専門用語の最上位の順位を得る．この平均順位（以下，平均最上位ランクと呼ぶ）を，ランキング出力に用いた類似度の最終的な評価値とする．なお，良い類似度であるほど，平均最上位ランクの値は小さくなる．

4.2.5 結果

比較する類似度尺度を，コサイン類似度行列（**K**），そ

表 3 シソーラスマッピングの実験結果：歪度（skewness）．値が小さいほど良い．**K** は基本となる類似度尺度としてコサイン類似度行列をそのまま使用した場合を表し，これをセンタリング操作により変換した行列を $\mathbf{K}^{\mathrm{cent}}$，正則化ラプラシアンとして変換した行列を $\mathbf{L}_{\mathrm{RL}}$，通勤時間カーネルとして変換した行列を $\mathbf{L}_{\mathrm{CT}}$ と表す．

データセット	**K**	$\mathbf{K}^{\mathrm{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$
カテゴリ A	6.620	2.115	1.193	**0.919**
カテゴリ B	9.911	3.341	1.766	**1.746**
カテゴリ C	7.311	2.254	**1.210**	1.215
カテゴリ D	9.005	3.589	1.578	**1.575**

表 4 シソーラスマッピングの実験結果：精度（平均最上位ランク）．値が小さいほど良い．

データセット	**K**	$\mathbf{K}^{\mathrm{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$
カテゴリ A	14.7	**11.7**	13.4	13.7
カテゴリ B	42.6	**35.7**	38.5	38.5
カテゴリ C	42.0	**35.6**	37.4	37.4
カテゴリ D	119.0	109.1	105.9	**105.7**

[*4] 2009 年版の中で Publication Year が 2000 年以降のデータ：http://www.nlm.nih.gov/archive/20100419/bsd/licensee/2009_stats/baseline_med_filecount.html を使用した．

してこれを元に作成するセンタリング行列（$\mathbf{K}^{\mathrm{cent}}$），正則化ラプラシアン（$\mathbf{L}_{\mathrm{RL}}$），通勤時間カーネル（$\mathbf{L}_{\mathrm{CT}}$）とし，MeSH の各カテゴリ（A，B，C，D）ごとに，ハブオブジェクトの出現度合いを示す歪度と精度を調べ，表 3，表 4 に報告する．

MeSH のどのカテゴリ（A，B，C，D）についても，コサイン類似度行列をそのまま用いる場合と比較して，センタリング，あるいは，グラフラプラシアンベースのカーネルに変換すると，歪度が下がり，精度は改善された．

4.3 マルチクラス分類

4.3.1 タスクの概要

文書（オブジェクト）を複数のトピック（クラス）のいずれかに分類する．クラス数がオブジェクトの数だけ存在するとも言えるシソーラスマッピングタスクと異なり，クラス数はオブジェクト数より小さく，1 つのクラスに複数オブジェクトが属する．

4.3.2 データセット

クラス分類された文書集合である Reuters 21578（新聞記事コーパス）および TDT2（新聞・テレビ・ラジオの記事からなるコーパス）の，それぞれ 52 クラスのサブセット（以下，Reuters-52）と 30 クラスのサブセット（以下，TDT2-30）を実験に用いる．文書オブジェクト数は，Reuters-52 が 9,100，TDT2-30 が 9,394 である．文書の特徴ベクトルへの変換は，Reuters-52 については CSMINING GROUP が公開するもの[*5] を使用し，TDT2-30 については Deng Cai ら [1] が変換処理したものを利用する[*6]. 特徴ベクトルの次元数は，Reuters-52 が 19,241，TDT2-30 が 36,771 である．コサイン類似度行列を計算する際には，シソーラスマッピングタスクと同様に，特徴ベクトルを tf-idf で重み付けし，長さ 1 に正規化する前処理を行う．

4.3.3 評価方法

文書を訓練データとテストデータに分け，テストデータについて，真のクラスラベルと予測ラベルが一致する割合（以下，accuracy）を測り，評価値とする．訓練・テストのスプリットは，Reuters-52 については，コーパスに付随するオリジナルのスプリットを用いる．TDT2-30 はスプリットが与えられていないため，各クラスに属するオブジェクトが均等に分配されるように 2 分割した．

4.3.4 比較する手法

コサイン類似度行列（$\mathbf{K}$），そしてこれを元に作成するセンタリング行列（$\mathbf{K}^{\mathrm{cent}}$），正則化ラプラシアン（$\mathbf{L}_{\mathrm{RL}}$），通勤時間カーネル（$\mathbf{L}_{\mathrm{CT}}$）を類似度尺度として用い，比較する．テストデータのラベルの予測は次のように行う．まず，テストデータに対して，これらの類似度尺度をもとに訓練データのみからなる k 近傍リストを作り，このリストに含まれるオブジェクトが持つクラスラベルを類似度の値で重み付けして投票する．投票により合計得点（類似度の重み付き和）が最大になるクラスを，テストデータの予測ラベルとする．なお，k の値の決定は，訓練データの 10 分割交差検定により行う．

表 5　マルチクラス文書分類の実験結果：歪度（skewness）．値が小さいほど良い．

データセット	$\mathbf{K}$	$\mathbf{K}^{\mathrm{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$
Reuters-52	14.815	11.037	**6.727**	6.934
TDT2-30	3.629	3.180	**2.560**	4.203

表 6　マルチクラス文書分類の実験結果：精度（accuracy）．値が大きいほど良い．

データセット	$\mathbf{K}$	$\mathbf{K}^{\mathrm{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$
Reuters-52	0.847	0.889	0.898	**0.900**
TDT2-30	0.964	**0.968**	0.964	0.963

4.3.5 結果

表 5，表 6 に歪度と精度（accuracy）を示す．まず，ハブオブジェクトの出現についてであるが，Reuters-52, TDT2-30 のいずれのデータセットについてもコサイン類似度では歪度が高く，ハブが出現している．特に，Reuters-52 では歪度が高く，顕著にハブが出現していると見られる．これに対し，センタリングでは歪度の下がり幅は大きくないものの，正則化ラプラシアンでは歪度は大きく下がる．精度については，Reuters-52 では，センタリング，グラフラプラシアンベースのカーネルともに精度を改善するが，TDT2-30 では，センタリングのみ精度を向上させる結果となった．

4.4 マルチラベル分類

4.4.1 タスクの概要

文書（オブジェクト）をクラスに分類するが，単一のオブジェクトが複数のクラスに属しうる点で，マルチクラス分類タスクと異なる．一つのオブジェクトが属すクラスの集合をスーパークラスと呼ぶと，クラス数 C に対してスーパークラスは最大で 2^C 通り存在する．スーパークラスの数は，オブジェクト数 N より通常は小さいが，マルチクラス分類タスクのクラス数よりは大であり，精度良く分類するためには高い解像度が求められる．

4.4.2 データセット

enron データ[*7] と bibtex データ[*8] を用いる．文書の特徴ベクトルへの変換は，Tsoumakas ら [13] による変換結果[*9] を利用する．特徴ベクトルの次元数は，enron は 1,001，bibtex は 1,836 である．特徴ベクトルは tf-idf 変換し，長さ 1 に正規化して類似度行列の計算に用いる．

[*5] http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html
[*6] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html
[*7] UC Berkeley Enron Email Analysis Project が作成した．
[*8] ECML PKDD Discovery Challenge 2008 で使用された．
[*9] http://mulan.sourceforge.net/index.html

表 7　マルチラベル文書分類の実験結果：歪度（skewness）．値が小さいほど良い．

データセット	$\mathbf{K}$	$\mathbf{K}^{\mathsf{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$
enron	6.465	**2.005**	2.574	3.038
bibtex	2.473	2.694	**2.431**	5.735

表 8　マルチラベル文書分類の実験結果：精度（loss）．値が小さいほど良い．

データセット	$\mathbf{K}$	$\mathbf{K}^{\mathsf{cent}}$	$\mathbf{L}_{\mathrm{RL}}$	$\mathbf{L}_{\mathrm{CT}}$
enron	2.80	2.67	**2.61**	2.67
bibtex	1.93	**1.91**	1.93	1.97

4.4.3　評価方法

文書を訓練データとテストデータに分け，まず，各々のテストデータが属するクラスの集合 S_{gold} に対し，分類器が出力するクラスの集合 $S_{predict}$ との不一致クラス数：$|S_{gold} \cup S_{predict}| - |S_{gold} \cap S_{predict}|$（以下，loss）を測る．そして，この loss の全てのテストデータに対する平均を取り，最終的な評価値とする[*10]．なお，訓練・テストのスプリットは，データセットに付随するオリジナルのスプリットを用いる．

4.4.4　比較する手法

コサイン類似度行列（$\mathbf{K}$），そしてこれを元に作成するセンタリング行列（$\mathbf{K}^{\mathsf{cent}}$），正則化ラプラシアン（$\mathbf{L}_{\mathrm{RL}}$），通勤時間カーネル（$\mathbf{L}_{\mathrm{CT}}$）を類似度尺度として用い，比較する．各テストデータに対して，これらの類似度をもとに訓練データだけを用いて k 近傍リストを作り，この k 近傍リストに対して ML-KNN 法 [16] を適用して，予測ラベル集合 $S_{predict}$ を出力する．また，k の値の決定は訓練データの 10 分割交差検定により行う．

4.4.5　結果

表 7，表 8 に歪度と精度（loss）を示す．まず，enron データでは，コサイン類似度で歪度が高くハブが出現しているが，センタリング，グラフラプラシアンベースのカーネルにより，歪度が下がり，ハブが減るとともに精度は向上している．他方，bibtex データでは，センタリング，グラフラプラシアンベースのカーネルにより歪度が下がらず，精度もほとんど向上しない．

5.　関連研究

本稿では，原点をデータ中心に移動するセンタリングはハブの出現を抑え，よって，k 近傍法の性能を改善する可能性について論じた．

センタリングの効果は，類似度ベースの手法を用いるときに現れるのであり，距離ベースの分類や検索を行う場合は，オブジェクト間の距離は原点を移動しても変わらないため，結果に影響しない．これに関して [5] が論じており，センタリングにより，SVM の分類結果は理論的には変わらないが，最適解を求める数値計算の安定性が増す，と述べている．

原点を（データ中心に限らず）移動することによって作られる，新たな類似度尺度を利用する研究としては，[12] がある．ただ，そこでの目的はクラスタリングの精度向上に限定されており，「高次元データに現れるハブを減らすために原点を移動する」という本研究とは，異なる動機に基づいている．

なお，k 近傍法を改良する研究としては，類似度尺度（または距離尺度）を訓練データを用いて学習する [9], [15]，k 近傍オブジェクトに対する重み割り当てを工夫する [2]，k を固定せずテストオブジェクトに応じて変更する [3], [14]，など，数多くの取り組みが報告がされているが，ハブを抑えるという観点からの研究は，著者らの知る限り，存在しない．

6.　結論

本稿では，センタリングによって作られる類似度尺度は，セントロイドとの類似度を全てのオブジェクトについて等しくするため，ハブの出現を抑えることについて論じた．実際，分類タスクと検索タスクの実験で，bibtex を除く全てのデータセットでセンタリングによりハブは減少し，精度が改良されることを確認した．特に，語義曖昧性解消タスクで SVM よりも高い精度を得たことは，今後に期待を抱かせる．どのような性質をもつデータセットに対してセンタリングが強く有効になるかを調べることは，今後の課題である．

[*10] これをクラス数で割ったハミングロスがマルチラベル分類タスクの評価値として用いられることがあるが [13]，ここでは真のクラス集合と予測されたクラス集合の要素の喰い違い数をそのまま評価値として用いる．

参考文献

[1] Cai, D., Wang, X. and He, X.: Probabilistic dyadic data analysis with local and global consistency, *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, pp. 105–112 (2009).
[2] Chen, Y., Garcia, E. K., Gupta, M. R., Rahimi, A. and Cazzanti, L.: Similarity-based Classification: Concepts and Algorithms, *J. Mach. Learn. Res.*, Vol. 10, pp. 747–776 (2009).
[3] Guo, R. and Chakraborty, S.: Bayesian adaptive nearest neighbor, *Statistical Analysis and Data Mining*, Vol. 3, No. 2, pp. 92–105 (2010).
[4] Lewis, D. D., Yang, Y., Rose, T. G. and Li, F.: RCV1: a new benchmark collection for text categorization research, *Journal of Machine Learning Research*, Vol. 5, pp. 361–397 (2004).
[5] Meil, M.: Data centering in feature space, *Ninth International Workshop on Artificial Intelligence and Statistics* (Bishop, C. M. and Frey, B. J., eds.), AISTATS (2003).
[6] Mihalcea, R.: Co-training and self-training for word sense disambiguation, *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL '04)* (Ng, H. T. and Riloff, E., eds.), Boston, Massachusetts, USA,

pp. 33–40 (2004).

[7] Mihalcea, R., Chklovski, T. and Kilgarriff, A.: The Senseval-3 English lexical sample task, *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)* (Mihalcea, R. and Edmonds, P., eds.), Barcelona, Spain, pp. 25–28 (2004).

[8] Navigli, R.: Word sense disambiguation: A survey, *ACM Computing Surveys*, Vol. 41, pp. 10:1–10:69 (2009).

[9] Qamar, A. M., Gaussier, É., Chevallet, J.-P. and Lim, J.-H.: Similarity Learning for Nearest Neighbor Classification, *ICDM*, pp. 983–988 (2008).

[10] Radovanović, M., Nanopoulos, A. and Ivanović, M.: Hubs in space: popular nearest neighbors in high-dimensional data, *Journal of Machine Learning Research*, Vol. 11, pp. 2487–2531 (2010).

[11] Suzuki, I., Hara, K., Shimbo, M., Matsumoto, Y. and Saerens, M.: Investigating the Effectiveness of Laplacian-Based Kernels in Hub Reduction, *Proceedings of AAAI-12* (2012).

[12] Thang, N. D., Chen, L. and Chan, C. K.: Clustering with Multiviewpoint-Based Similarity Measure, *IEEE Trans. Knowl. Data Eng.*, Vol. 24, No. 6, pp. 988–1001 (2012).

[13] Tsoumakas, G., Katakis, I. and Vlahavas, I. P.: Mining multi-label data, *Data Mining and Knowledge Discovery Handbook* (Maimon, O. and Rokach, L., eds.), Springer, 2nd edition, pp. 667–685 (2010).

[14] Wang, J., Neskovic, P. and Cooper, L. N.: Neighborhood size selection in the *k*-nearest-neighbor rule using statistical confidence, *Pattern Recognition*, Vol. 39, No. 3, pp. 417–423 (2006).

[15] Weinberger, K. Q. and Saul, L. K.: Distance Metric Learning for Large Margin Nearest Neighbor Classification, *Journal of Machine Learning Research*, Vol. 10, pp. 207–244 (2009).

[16] Zhang, M.-L. and Zhou, Z.-H.: ML-KNN: a lazy learning approach to multi-label learning, *Pattern Recognition*, Vol. 40, pp. 2038–2048 (2007).