3F7-3

根つき最小全域木(MST)を利用した化学構造の類似性検索

Similar Structure Searching for Chemicals using Rooted Minimal Spanning Tree

熊谷 正雪 Masayuki Kumagai　藤島 悟志 Satoshi Fujishima　高橋 由雅 Yoshimasa Takahashi

豊橋技術科学大学 工学部 知識情報工学系

Department of Knowledge-based Information Engineering, Toyohashi University of Technology

In this paper, we proposed an new approach to similar structure searching for drug molecules. The method is based on finding a rooted minimal spanning tree (MST). A query structure is regarded as the root vertex for finding the rooted MST that consists of a specified number of vertexes (i.e. drug molecules). This approach gives us an alternative searched result that is different from the result that obtained by the conventional k-nearest neighbors method. The present method allows us to take vicinal relations of the data at the searching. The detail of the algorithm is described with an illustrative example, then the utility of the method is discussed with several artificial data sets and a real chemical data set.

1. はじめに

構造類似性検索においては注目するクエリに対し、対象とするデータベース中の全ての構造との類似性を評価し、その類似度の高いものから任意の数のデータを参照する方法（KNN 法）が一般的である。この方法ではクエリを中心に、ある半径内に位置するデータを参照することになる。

本研究では、検索対象のデータ空間に対して、クエリを基点とした MST を生成しながら構造類似性検索を行うことで、データの近接性が反映された探索結果を得られる手法を考案し、その有用性を検討した。

2. 最小全域木(Minimal Spanning Tree,MST)

あるグラフ G の全域木とは G の全ての頂点を含んだ木グラフである。全域木の中で、その重み(本研究では距離)が最小のものを MST という。本研究では MST の探索には Prim のアルゴリズムを用いた。Prim のアルゴリズム[Prim 57]を以下に示す。

＜Prim のアルゴリズム＞

(1) グラフ中の任意の点を選び、MST 部分木の構成点とする。
(2) MST 部分木の構成点からの距離が最短の点を選び、MST 部分木の構成点に含める。
(3) グラフの全ての点が MST 部分木の構成点に含まれるまで(2)を繰り返す。

3. MST を利用した類似性検索手法

本手法は上記の MST 探索アルゴリズムをもとに、以下に示す手順で類似性探索を行う。ただし、今回は入力値として構造データではなく 2 次元の座標データを用いた。

(1) クエリを MST 部分木の根として、クエリと各サンプル間の類似度（または非類似度）を計算する。
(2) クエリに対する最近隣（最も類似度の高い）パターンを選択し、2 頂点からなる MST 部分木を構成する。
(3) 構成された部分木の頂点集合の各要素に対する最近隣パターンを探索し、そのうちで、最も類似度の高いパターンを選択し、新たな部分木を構成する。
(4) 部分木のサイズが予め指定された大きさに達していれば探索を終了。さもなければ、(3)、(4)の操作を繰り返す。

上述の MST 部分木探索による方法は、KNN 法と比べた場合、一般に探索の結果、取り出される順序が異なることが予想される。図 1に従来法と MST 法の探索順序の概念図を示す。

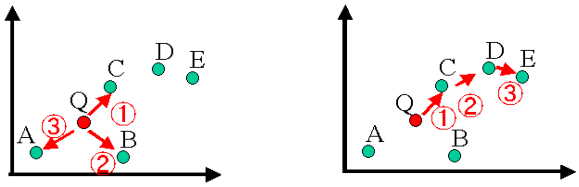


図 1 KNN 法(左)と MST 法(右)の探索順序

KNN 法は常にクエリ自身との比較から近い順に取り出される。一方、MST 部分木法はクエリを含む部分木構成点のいずれかとの最近隣点がその都度取り出される。そのため、MST 法を用いることによって、データ空間における局所的な近接性をより詳細に反映した検索結果が期待できる。

4. 合成データを用いた計算機実験

2 次元の合成データ上で KNN 法と MST 法による検索を行い、結果の比較を行った。4.1に示す合成データに対し、クエリの位置を変えて検索を行った結果を 4.2 に示す。なお、類似度評価関数にはユークリッド距離を用いた。

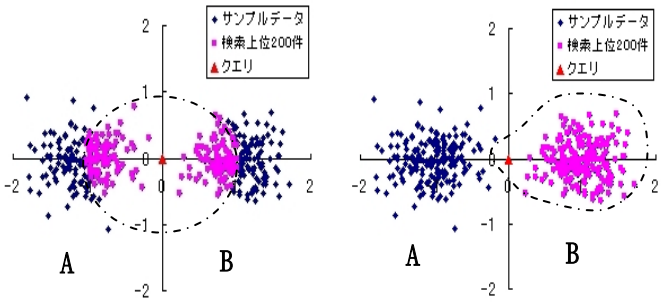
4.1 合成データセットの説明

作成した合成データは 2 つのクラスタ(クラスタ A,B)を含んでいる。各クラスタに含まれるサンプルの数は 200 で、その分布は分散(0.15,0.15)の正規分布である。クラスタ A の中心点は(-1,0)、クラスタ B の中心点は(1,0)である。

4.2 実験結果

初めにクエリの位置を(0,0)として、KNN 法と MST 法による検索を行った。図2は検索結果上位 200 件を図示したものである。破線部が上位に検索されたサンプルである。

連絡先：高橋由雅，〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1 豊橋技術科学大学 知識情報工学系，Tel: 0532-44-6878, taka@mis.tutkie.tut.ac.jp

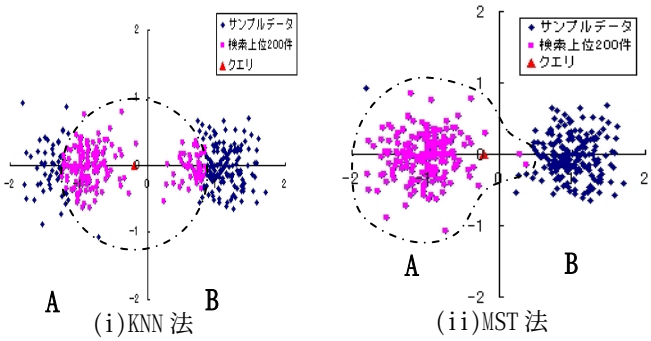


(i)KNN 法　　(ii)MST 法

図 2 クエリ位置(0,0)の場合の検索結果

KNN 法ではクエリを中心とした円状に検索範囲を広げている。そのため、クラスタ A,B のほぼ同数のサンプルを同時に上位に検索している。一方、MST 法では検索上位 200 件全てがクラスタ B のサンプルであった。これはクエリの最近傍にクラスタ B のサンプルが位置しており、そのデータの近接性によってクラスタ B が優先的に検索された結果である。

次に、クエリの位置をクラスタ A 側に近づけた例を示す。図 3 はクエリ位置を(-0.2,0)に置き、KNN 法と MST 法による検索を行った結果である。



(i)KNN 法　　(ii)MST 法

図 3 クエリ位置(-0.2,0)の場合の検索結果

この例でも KNN 法はクエリを中心とする円状に検索していることが分かる。一方、MST 法では検索上位 200 件中、左上のサンプルを除く 199 件がクラスタ A 側のサンプルであった。クエリの最近傍にクラスタ A のサンプルが位置しており、そのデータの近接性によってクラスタ A が優先的に検索された結果である。

5. 実データを用いた計算機実験

実データに対して KNN 法と MST 法による検索を行い、結果の比較を行った。5.1 に示す実データに対して、検索を行った結果を 5.2 に示す。なお、類似度評価関数にはユークリッド距離を用いた。

5.1 データセット

実データを用いて KNN 法との比較を試みた。実験には治験薬構造データベース MDDR[MDL 02]から抽出した薬物構造データ 1364 件を用いた。

化合物の構造特徴の記述子には筆者らの提案するトポロジカルフラグメントスペクトル（Topological Fragment Spectra; TFS[Takahashi 98]）を用いた。TFS とは化学物質の構造式から可能な部分構造を列挙し、その数値的な特徴付けに基づいて化学物質のトポロジカルな構造プロフィールを多次元数値ベクトルとして表現しようとするものである。ここでは，結合サイズ 5 までの部分構造を列挙し、特徴付けには各部分構造の質量数を用いた。結果として、各化合物の構造特徴は 163 次元のベクトルとして記述された。

5.2 結果

図4、図5は同一のクエリ構造(図中の左上)を用いて KNN 法と MST 法で類似構造探索（10 件）を行った結果を示したものである。2 位までの構造は両者で共通であるが、3 位の構造以降は両者で異なる検索結果を与えている。例えば、KNN 法で 3 位に挙げられた構造は MST 法では 276 位に挙げられている。また、MST 法の結果では 7 位までにクエリと類縁の構造が並んで検索されていることが分かる。すなわち、両者が異なる探索結果を与えることは明らかであり、特に後者はクエリ構造に対する近接性に加え、探索過程で検出される類似パターンの局所近傍を探索することによって、より研究者の直感に近い類似構造探索が可能であることを示唆している。

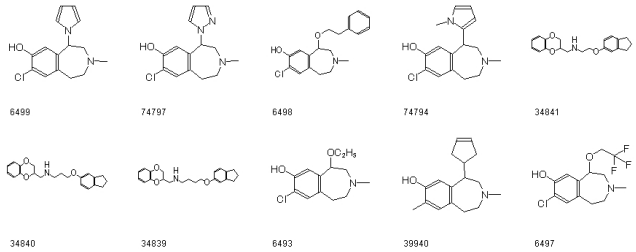


図 4　KNN 法の探索結果（上位 10 件）

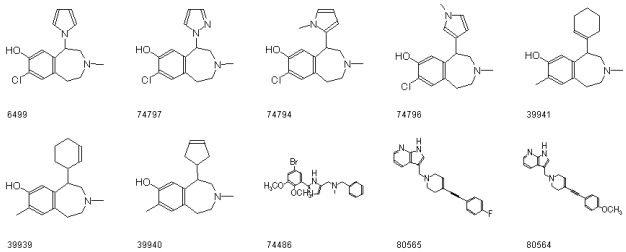


図 5　MST 法の探索結果　（上位 10 件）

6. おわりに

以上、一般的に用いられている KNN 法と本研究での提案手法である MST 法との構造類似性検索結果の比較を行った。結果として、両者の検索法では検索される構造の順位が大きく異なることを示した。さらに MST 法は、クエリの近傍に存在するクラスタを検出し、そのクラスタを優先的に検索することが可能な検索手法であることが示唆された。これらの結果から、本研究で提案した根付き MST 法は従来の検索方法とは異なる視点から構造類似性を捉えることができるといえる。

参考文献

[Prim 57] Robert Prim, “Shortest connection networks and some generalizations”, Bell System Technical Journal, vol.36, pp.1389-1401, 1957.

[MDL 02] MDL Drug Data Report, http://www.mdli.com/

[Takahashi 98] Y.Takahashi, H.Ohoka and Y.Ishiyama, “Structural Similarity Analysis Based on Topological Fragment Spectra”, Advances in Molecular Similarity, Vol.2, pp.93-104,1998.