情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

# 木構造および化学構造に対する特徴ベクトル：埋め込み、検索、構造推定

# Feature Vectors for Trees and Chemical Structures: Embedding, Search and Pre-Image

阿久津　達也*
Tatsuya Akutsu

**Abstract:** In this short article, we briefly review our recent results on feature vectors on tree structures. For edit distance between trees, it is shown that the edit distance between two ordered trees can be approximated within a factor of $O(h)$ by using the edit distance between the corresponding Euler strings, and the edit distance between two unordered trees can be approximated within a factor of $O(h)$ by using feature vectors consisting of the numbers of occurrences of subtrees (each induced by a node and its descendants), where $h$ is the minimum height of input trees. For the pre-image problem on trees (i.e., inferring a tree from a feature vector consisting of the numbers of occurrences of vertex-labeled paths), it is shown that the problem can be solved in polynomial time in the size of an output graph if the graphs are trees whose maximum degree is bounded by a constant and the lengths of given paths and alphabet size are bounded by constants, but is NP-hard even for trees of bounded degree if the maximum length of paths is not bounded. A practical branch-and-bound algorithm for the pre-image problem is also reviewed.

**Keywords:** feature vector, kernel methods, embedding, tree edit, graph pre-image

## 1　まえがき

特徴ベクトルはカーネル法を始めとする機械学習手法などにおいて幅広く利用されている。一方、高速検索などを目的に文字列の編集距離の特徴ベクトルを用いた $L_1$ 空間への埋め込みが理論計算機科学分野で研究されてきた [4]。これらを背景に、筆者はいくつかの共同研究を通じて、木構造の編集距離の特徴ベクトルを用いた $L_1$ 空間への埋め込み、および、パスの出現頻度に基づく特徴ベクトルからの木構造の推定という問題に取り組んできた。本稿ではその結果について概観する。

## 2　木の編集距離の埋め込み

木の編集距離は文字列の編集距離を（根付き）木に拡張したものであり、木 $T_1$ を木 $T_2$ に変換するのに必要な頂点の挿入、削除、置換の最小回数（もしくは最小コスト）として定義される。順序木の編集距離は $O(n^3)$ 時間で計算でき、無順序木の編集距離の計算は NP 困難であることが知られている [6]。ここで、木 $T_1$ と木 $T_2$ の間の編集距離を $D_T(T_1, T_2)$ とする。この距離を特徴ベクトル間の距離により近似するのが目標である。

順序木の場合には、木を深さ優先探索した結果得られるオイラー文字列を用いることにより、順序木の情報を文字列に変換することができる。木 $T$ に対するオイラー文字列を $s(T)$ とし、$T_1, T_2$ のうち、低い方の木の高さを $h$ とすると、

$$\frac{1}{2}D_S(s(T_1), s(T_2)) \leq D_T(T_1, T_2) \leq (2h+1)D_S(s(T_1), s(T_2))$$

が成立する [2]。なお、$D_S(s_1, s_2)$ は文字列 $s_1, s_2$ に対する編集距離である。この結果は木の高さが低い場合には、オイラー文字列間の編集距離により順序木間の編集距離が良い精度で近似できることを示しており、文字列の編集距離に対する埋め込みを用いると、順序木間の編集距離が効率的に埋め込める事を示唆している。なお、

*京都大学　化学研究所　バイオインフォマティクスセンター, 〒 611-0011　京都府宇治市五ヶ庄, tel. 0774-38-3015, e-mail takutsu@kuicr.kyoto-u.ac.jp,
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.

木の高さが高い場合には意味のある近似精度とはならないので、その場合に対する拡張についても研究を行っている [3]。

無順序木の場合には、各頂点とその子から誘導される部分木で同型なものの個数を並べた特徴ベクトルを $\phi(T)$ とする。この特徴ベクトルに対し、

$$\frac{1}{2h+2}\|\phi(T_1)-\phi(T_2)\|_1 \leq D_T(T_1,T_2) \leq \|\phi(T_1)-\phi(T_2)\|_1$$

が成立する [7]。この結果は木の高さが低い場合には、特徴ベクトル間の $L_1$ 距離により、無順序木間の編集距離が良い精度で近似できることを示している。

## 3　特徴ベクトルからの構造推定

サポートベクターマシン (SVM) などに代表されるカーネル法においては、カーネル関数は特徴ベクトルの内積として定義される。SVM などを用いて精度の高い予測を行うためには、もとのデータの性質を反映した特徴ベクトルを設計することが重要であり、数多くの研究が行われてきた。一方、与えられた特徴ベクトル $\boldsymbol{v}$ からもとの構造 $\phi^{-1}(\boldsymbol{v})$ を計算する問題は pre-image 問題とよばれ、機械学習分野でもいくつかの研究が行われてきた [5]。しかしながら、計算量の観点からはほとんど研究されていなかったため、木構造を対象にラベル付きパスの出現回数からなる特徴ベクトルを用いた場合についての研究を行い、以下の結果を得た [1]。

- パスの最大長、頂点の最大次数、（頂点の）ラベルの種類のいずれもが定数である場合には、$\boldsymbol{v}$ から $\phi^{-1}(\boldsymbol{v})$ の計算（有無の判定を含む）は多項式時間で可能。
- 頂点の最大次数、ラベルの種類のいずれもが定数であってもパスの最大長に制限がなければ NP 困難。
- パスの最大長が定数であっても、頂点の最大次数、ラベルの種類に制約がなければ NP 困難。

化学構造においては、ベンゼン環などを 1 個の頂点とみなしてしまえば木構造を持つものが多い。そこで、上記の pre-image 問題に対するアルゴリズムを新規化学構造の設計に応用することが考えられる。しかしながら、上記アルゴリズムは多項式の次数が高すぎて実用的でない。そこで、木構造の効率的枚挙といくつかの枝刈り手法を組み合わせた、より実用的なアルゴリズムを開発している [8]。現時点では、水素原子を除いて 20〜30 原子程度までのサイズの木状構造を持つ化合物に対して現実的な時間で $\phi^{-1}(\boldsymbol{v})$ が計算可能となっている。

## 4　おわりに

木の編集距離の埋め込みについては木の高さに制限のない場合については未解決であり、高速検索その他の応用とともに、今後の課題となっている。

一方、グラフの pre-image 問題については、理論面からは一般的なグラフ構造の場合の計算複雑度の解析、また、応用面からは枚挙と枝刈りに基づく手法の更なる効率化、および、化学構造設計への具体的応用が今後の課題となっている。

## 謝辞

本稿で述べた研究の共同研究者各位に感謝の意を表する。

## 参考文献

[1] T. Akutsu and D. Fukagawa, “Inferring a graph from path frequency”, *Lecture Notes in Computer Science*, 3537:371–382, 2005.

[2] T. Akutsu, “A relation between edit distance for ordered trees and edit distance for Euler strings”, *Information Processing Letters*, 100:105–109, 2006.

[3] T. Akutsu, D. Fukagawa, and A. Takasu, “Approximating tree edit distance through string edit distance”, *Algorithmica*, to appear.

[4] A. Andoni and K. Onak, “Approximating edit distance in near-linear time”, *Proc. ACM Annual Symposium on Theory of Computing*, 199–204, 2009.

[5] G.H. Bakir, A. Zien, and K. Tsuda, “Learning to find graph pre-images”, *Lecture Notes in Computer Science*, 3175:253–261, 2004.

[6] P. Bille, “A survey on tree edit distance and related problem”, *Theoretical Computer Science*, 337:217–239, 2005.

[7] D. Fukagawa, T. Akutsu, and A. Takasu, “Constant factor approximation of edit distance of bounded height unordered trees”, *Lecture Notes in Computer Science*, 5721:7–17, 2009.

[8] Y. Ishida, L. Zhao, H. Nagamochi, and T. Akutsu, “Improved algorithms for enumerating tree-like chemical graphs with given path frequency”, *Genome Informatics*, 21:53–64, 2008.