G-010

グラフカーネルを用いた SVM による代謝物の機能予測
Function Prediction of Metabolite by SVM with Graph Kernel

田口 雄大† 蓬莱 尚幸†
Yudai Taguchi Hisayuki Horai

1. はじめに

代謝とは生命活動を維持するために生体内で化合物を合成することであり、その代謝によって合成される化合物を代謝物と言う。代謝物は一次代謝産物と二次代謝産物に大きく２種類に分類される。一次代謝産物とは細胞成長や生殖、発生などの生命活動に直接関わる代謝物の総称であり、代表的な一次代謝産物としてアミノ酸や核酸、脂質などが挙げられる。二次代謝産物とは一次代謝産物のように生物の細胞成長や生殖、発生などに直接関係はしないが、植物や微生物の中で生成される代謝物の総称である。二次代謝産物には生物学的機能を有しているものが多い、例えば、植物の感染防御などの役割を果たすものがある。しかし、これらの二次代謝産物の多くは機能未知である。実際、生物学的実験にて発見された二次代謝産物の数がおよそ５万種類であるのに対して、そのうち機能既知である二次代謝産物の数は発見された二次代謝産物の数の 10%以下である。

そこで本研究は機能的に未知である二次代謝産物の生物学的機能を解明するために、代謝物の機能予測手法の開発を目的とする。今回は代謝物関連データベースを用いて、生物学的機能の一つである細胞毒性を持つ二次代謝産物を Support Vector Machine により予測する。そして、その予測性能を交差検証の一つである Leave-One-Out を用いて評価し、その結果を考察する。

2. 予測手法

2.1 実験データ

二次代謝産物の化学情報を格納するデータベースとして代表的なものが KNApSAcK[1]データベースである。この KNApSAcK データベースは生物種と二次代謝産物の関係を蓄積した二次代謝産物データベースシステムであり、50,048 種の代謝物と 101,500 対の生物種とその生物の代謝物の関係を収録している。本研究では KNApSAcK に登録されている二次代謝産物の中で 51 個の細胞毒性を持つ化合物と 76 個のそれ以外の化合物のあわせて 127 個の化合物の化学構造を実験データとした。

2.2 Support Vector Machine

Support Vector Machine(SVM)とは２クラスのパターン識別のための教師あり機械学習法である。SVM では超平面から最も近い距離にあるトレーニングサンプルとの間の距離、即ちマージン(Margin)を最大化するような超平面を計算することで学習が行われる。しかし、代謝物の機能予測においては学習データとして化合物の構造データを扱わなければならなく、超平面で分類しただけでは正しく分類が行われない。そこで学習データ$\boldsymbol{x}$を特徴ベクトル$\boldsymbol{\Phi(x)}$に写像し、カーネル法を適応する。カーネル法ではカーネル関数$\boldsymbol{K(x,y)}$を計算することによって学習する。カーネル法を用いて SVM を行うことで学習データとして構造データを扱うことが可能となる。

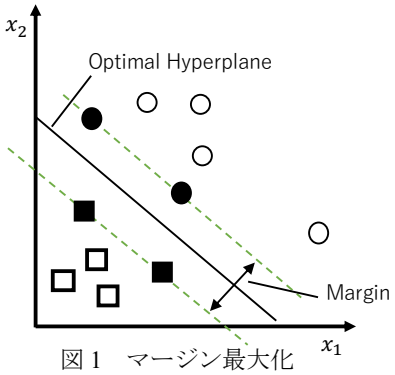

図 1　マージン最大化

2.3 Marginalized Graph Kernels

化合物データを SVM で計算するためには化合物のカーネル関数を定義する必要がある。しかし、構造データである化合物のカーネル関数を定義するのは難しい。そこで、化合物間のカーネルを定義するために化合物をグラフとして表し、グラフ間の類似度を測る Marginalized Graph Kernels が用いられる。ここで、カーネル関数を次のように定義する。

$$K(G_1,G_2)=\sum_{h\in v_1^*}\sum_{h\in v_2^*}p(h)p(h')K'(l(h),l(h'))$$

h,h' はそれぞれグラフG_1,G_2におけるパス、$l(h)$はパスのラベル(原子名)の列、$K'(x,y)$はラベル列間のカーネル関数を示す。これをすべての化合物のペアについて計算し、カーネル関数を定義する。

2.4 Morgan index

実際にグラフカーネルを計算するためには膨大な時間がかかり、Marginalized Graph Kernel を用いるとランダムウォークの計算により更に計算量が増加する。そこで、その問題を解決するために Morgan index が用いられる。Morgan index は以下の反復プロセスによって定義される

１．すべての原子に整数１をラベル付けする
２．すべての原子に結合している原子の整数の総和をその原子に再ラベル付けする
３．２を繰り返す

この処理をグラフカーネルを生成する前に施すことによって、グラフ間の共通ラベル経路の数が減少し、グラフカーネルの計算時間を短縮することができる。また、パ

† 茨城工業高等専門学校専攻科情報工学コース
Advanced Course of Information Engineering, National Institute of Technology, Ibaraki College

スがこのラベル付けによって変化するため、グラフ間の類似度を測る性能が向上する可能性がある。

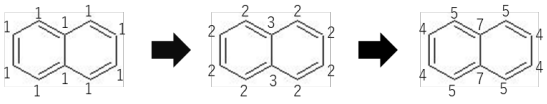

図 2　Morgan index の過程

3. 予測性能評価

3.1 Leave-One-out 交差検証

代謝物の機能予測手法の性能を評価するために、本研究では交差検証の一つである Leave-One-Out を用いる。Leave-One-Out では全てデータセットのうち、1 個のデータをモデル評価のためのテストデータとして、その他の全てのデータを学習するためのトレーニングデータセットとして予測を行う手法であり、テストデータとトレーニングデータセットを全データの個数の回数分変化させながら、全データについてそれを繰り返すことによって、予測手法を評価する。

3.2 正解率、適合率、再現率および F 値

予測結果を評価するために、適合率(Precision)、再現率(Recall)及び F 値を求める。二値分類における結果は Table1 のように 4 つの状態に分けられる。予測結果が正で実際のクラスも正の場合の TP(True Positive)、予測結果が正で実際のクラスが負の場合の FP(False Positive)、予測結果が負で実際のクラスが正の FN(False Negative)、予測結果が負で実際のクラスも負の TN(True Negative)である。

表 1　予測結果の分類

		予測されたクラス	
		Positive	Negative
実際のクラス	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

そして、正解率、適合率、再現率および F 値は以下のように表される。

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Fmeasure = \frac{2 \cdot Precision \cdot \mathrm{Recall}}{Precision + Recall}$$

正解率とは予測したデータのうち予測が合っていたデータの割合、適合率とは実際に正と予測したデータのうち、実際に正である割合、再現率は実際に正であるデータのうち、正であると予測されたものの割合を示す値である。F 値はその適合率と再現率の調和平均を取ったものであり、トレードオフの関係にある適合率と再現率の性能を評価するための値である。

4. 結果

本研究では、SVM を実装するためのツールとして scikit-learn を用いた。また、予測精度の高い、最適な学習モデルを選択するため、コストパラメータCとカーネル関数のパラメータ$p_q, iteration$の値を変化させることでチューニングを行った。最適なコストパラメータ上での各カーネル関数のパラメータにおける正解率を表 2 に示す。

表 2　各パラメータにおける正解率

p_q	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1st iteration	**63.77**	**63.77**	**63.77**	**63.77**	**63.77**	**63.77**	**63.77**	**63.77**	**63.77**
2nd iteration	67.71	66.92	67.71	66.92	66.92	65.35	66.92	**65.35**	**66.14**
3rd iteration	69.29	70.07	68.50	67.71	67.71	**66.14**	68.50	69.29	**66.14**
4th iteration	**71.65**	**71.65**	**71.65**	70.07	0.69	71.65	68.50	67.71	**66.14**
5th iteration	70.86	**71.65**	**72.44**	**72.44**	68.50	70.86	68.50	67.71	**66.14**
6th iteration	69.29	70.86	**72.44**	**72.44**	68.50	70.86	68.50	67.71	**66.14**
7th iteration	70.86	70.86	**72.44**	**72.44**	68.50	70.86	68.50	67.71	**66.14**
8th iteration	70.07	70.07	**72.44**	**71.65**	68.50	70.86	68.50	67.71	**66.14**
9th iteration	70.86	70.07	**72.44**	**71.65**	68.50	70.86	68.50	67.71	**66.14**

次にカーネル関数の最適パラメータにおける予測結果を示すと、TP=34, FP=18,FN=17,TN=58 であり、適合率は 0.65、再現率は 0.67、F 値は 0.66 だった。

5. 考察

表 2 を見ると、$iteration$が 1, 2 のとき、他の iteration と比べ全体的に正解率が低く、それ以上の iteration の時のほうが正解率が高くなっていることが分かる。これは Morgan index を繰り返すことで、グラフのパスが変化し、精度が向上したことを示している。最適な Marginalized Graph Kernel のパラメータはp_q =0.3, 0.4 であり、iteration が 5 回以降のものが正解率、つまり予測された全ての化合物のうち予測が正解だったものの割合が比較的ほかの正解率と比べて高くなっていることが分かる。また、p_q = 0.9のとき、どの iteration でも正解率が最低値を記録している。これはp_qが大きな値のため Marginalized Graph Kernel によって生成されるパスが短くなってしまい、比較的上手く予測ができていないと考えられる。

最適パラメータにおける予測結果を見ると、適合率は 0.65、再現率は 0.67 であることから、約 65%の確率で細胞毒性を持つ化合物を予測でき、また細胞毒性を持つ化合物の約 67%が正しく分類できたことを示している。

6. 結論

本研究では機能的に未知である二次代謝産物の生物学的機能を解明するために、代謝物の機能予測手法の開発を行った。結果として、適合率は 0.65、再現率は 0.67、F 値は 0.66 が得られ、あまり高い予億性能を得ることができなかった。そのため、今後の課題としてはより予測性能の高い機能予測手法の開発が挙げられ、それを達成するために化合物データへの工夫が必要となる。具体的には今回、実験データとして 127 個の化合物を用いて開発を行ったが、この数のデータを学習データとして扱うには少なすぎるため、データを増やす必要がある。また、化合物データとして扱っているのが代謝物であり、代謝物は構造が複雑なものが多く予測精度が高くないのはそれが原因の一つであると考えられる。よって、その問題を解決するような工夫も必要となってくるだろう。

参考文献

[1] S Kanaya, A Hirai, H Takahashi, etc, “Species- Metabolite relation database KNApSAcK : toward, Comprehensive understanding of metabolites in edible/medicinal plants around the world”, SIS (2010)
[2] [1] H. Kashima, K. Tsuda, A. Inokuchi, “Marginalized Kernels Between Labeled Graphs”, Proc.20th Intern. Conf. Machine Learning, ICML, pp321-328 (2003).
[3] P Mahé, N Ueda, T Akatsu, L Vert J, P, Extensions of Marginalized Graph Kernels, In Proceedings of the Twenty-First International Conference on Machine Learning, pp552-559 (2004)