2F1-03

ポリマー判別のための共生進化に基づく決定木生成

Generating the Decision Tree based on Symbiotic Evolution for Discrimination of Polymers

大谷紀子[*1] 貝原巳樹雄[*2] 佐伯和光[*3] 志村正道[*1]
Noriko Otani Mikio Kaihara Kazumitsu Saeki Masamichi Shimura

[*1]武蔵工業大学 Musashi Institute of Technology
[*2]一関工業高等専門学校 Ichinoseki National College of Technology
[*3]富山県工業技術センター Toyama Industrial Technology Center

This paper describes how to generate decision trees for discriminating polymers accurately with near-infrared rays spectrum. The former system using symbiotic evolution SESAT can generate simple and accurate trees, but is not useful for data that has a lot of attributes like near-infrared rays spectrum. We designed the structure of the partial solution "sprig" for sufficient learning. In addition, the fitness function of the whole solution "decision tree blueprint" was defined for 2-class discrimination. Based on our method we developed a system called SEPT for generating decision trees. Our experimental results show that SEPT has the ability to generate accurate trees for discrimination of polymers.

1. はじめに

近年，資源枯渇と環境破壊の問題が深刻化する中で，プラスチック等のポリマー製品のリサイクルの必要性が高まっている．リサイクルにあたってはポリマーを種類ごとに分別しなければならないが，1 つのポリマーのみ他とは別の処理を施すことがあるので，特定のポリマーとそれ以外を簡便かつ正確に分別する手法が望まれている．

近赤外スペクトルは赤外スペクトルと比較して透過率が高く，異物の付着や凹凸，着色などの表面状態から受ける影響が少ない．すなわち，近赤外スペクトル波形の測定では，測定対象の汚れを落としたり，変形したりといった前処理が必要ない．従って，近赤外スペクトルは非破壊的かつ迅速なポリマー判別に有用なデータといえる．近赤外スペクトルを用いたポリマー判別法として，ニューラルネットワークや 1/f ゆらぎなどに基づく手法が提案されている [松本 99, 小野寺 99]．しかし，いずれの方法でも判別過程が明確でなく，各ポリマーの特徴がスペクトルのどの部分に表れるのかを知ることは困難である．

判別過程を明確に示す分類規則の表現技法の 1 つに決定木がある．決定木はノードとアークからなる木構造であり，非終端ノードには属性の種類，アークには属性値，終端ノードにはクラスが割り当てられている．事例の各属性値に従って根ノードから終端ノードまで決定木を辿ることで，事例の属するクラスを判定することができる．これまでに，代表的な決定木生成システム CART[Breiman 84] を用いて，特定のポリマーとそれ以外を判別するための決定木の生成が試みられているが，正解率 100%での判別は実現されていない [貝原 03]．

遺伝的アルゴリズムの一種である共生進化 [Moriarty 96] に基づいて決定木を生成する方法が提案され，決定木生成システム SESAT によりその有効性が確認されている [大谷 04]．しかし，ポリマー判別に使用するデータは属性数が非常に多く，属性値間に連続性があるため，SESAT でポリマー判別のための決定木を十分に学習するのは難しいと考えられる．

本研究では，判別過程が明確にわかるポリマー判別システムの構築を目的として，近赤外スペクトルによりポリマーを正確に判別するための決定木生成手法を提案する．本手法は SESAT における決定木生成手法をポリマー分別用に改変したものである．部分解 sprig を近赤外スペクトルデータの特徴に適合した形で表現し，特定の種類のポリマーとその他を分別する 2 分決定木を生成する．また，個体の評価には，正解率の代わりに再現率と適合率の積を用いた適応度を使用する．提案手法に基づくポリマー判別システム SEPT(Symbiotic Evolution for Polymer discriminative Trees) を構築し，実測データによる評価実験で提案手法の効果を確認する．

表 1: ポリマーの略称と試料数

ポリマー名	略称	数
ポリスチレン	PS	44
ポリエチレン	PE	43
ABS 樹脂	ABS	24
ポリプロピレン	PP	31
ポリカーボネート	PC	13
ポリ塩化ビニル	PVC	15
ポリオキシメチレン	POM	18
ポリカーボネート/ABS 樹脂共重合体	PC/ABS	11
アクリロニトリル/スチレン共重合体	AS	13
ポリメタクリル酸メチル	PMMA	8
ポリエチレンテレフタレート	PET	11
ポリブチレンテレフタレート	PBT	3
尿素樹脂	UF	3
メラミン	MF	3
ポリフェノール	PF	3
ナイロン 66	PA66	6
ナイロン 6	PA6	6
セルロース	CEL	3

2. ポリマーの近赤外スペクトル

本研究で用いたポリマーは，表 1 に示す 18 種類の 258 試料である．近赤外分光装置を用いて各試料の近赤外波長領域 1200〜2400nm の吸光度を 0.5nm おきに測定し，正規化したデータを決定木生成に使用する．各波長に対応する吸光度をそれぞれ $x_1, x_2, \cdots, x_{2401}$ とすると，正規化データ $\vec{x'}$ は測定データ $\vec{x} = (x_1, x_2, \cdots, x_{2401})$ により次の式で求められる．

$$\vec{x'} = \frac{\vec{x}}{|\vec{x}|} \tag{1}$$

決定木では，1200〜2400nm の 2401 種類の波長を属性，各

連絡先: 大谷紀子，武蔵工業大学環境情報学部
〒 224-0015 横浜市都筑区牛久保西 3-3-1，045-910-2938
E-mail: otani@yc.musashi-tech.ac.jp

属性に対応する正規化吸光度 $x'_1, x'_2, \cdots, x'_{2401}$ を属性値，ポリマーの種類をクラスとして試料を分別する．近赤外スペクトルのデータには，SESAT の評価に用いられたデータと比較して，以下のような特徴がある．

1. 属性数が非常に多い．
2. 属性は連続値を一定間隔で抽出した値を表す．
3. 全属性値が吸光度という同一対象を表す連続値である．

従って，決定木の各ノードで分岐の基準とする属性，および分岐の閾値に関して，より多様な解候補からの探索を行なう必要がある．

3. ポリマー分別のための共生進化

共生進化は Moriarty ら [Moriarty 96] により提案された遺伝的アルゴリズムの一手法であり，部分解と全体解をそれぞれ個体とする 2 つの集団を並行して進化させる点に特徴がある．全体解は部分解の組み合わせにより表現する．全体解の評価に基づいて部分解を評価し，その評価値に従って進化した部分解を全体解に反映する．両者を相互に関係付けながら進化させることで集団内の個体の多様性が維持され，局所解への収束を回避した効率的な最適解探索を可能としている．本節では，提案手法に基づく決定木生成システム SEPT の詳細について，SESAT との相違点を中心に説明する．

3.1 SESAT との相違点

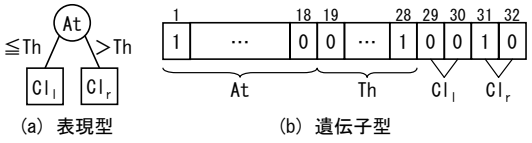
SESAT は多クラス分別が可能な多分木を生成するが，学習過程および生成可能な決定木に関して次のような問題点を持つ．

1. 交叉によって属性が変化することはない．
2. 子ノードの種類と閾値を独立に変化させられない．
3. 1 つのノードが持つ子ノード数の上限値を M とするとき，分岐の閾値となる値が $2M-3$ 通りしかない．

2. 節で述べた近赤外スペクトルデータの特徴を考慮すると，SESAT における sprig の構造では，より正確にポリマーを分別する決定木の生成は難しいと考えられる．そこで SEPT では，先行研究 [貝原 03] と同様に，ある特定のポリマーとそれ以外のポリマーを分別するための 2 分決定木を生成することとし，上記の問題点を解決すべく SESAT における sprig の構造を変更する．

特定のポリマーとそれ以外を分別するための決定木を正解率に基づいて学習する際には，両者の事例数の偏りが問題となる．1 つの事例が不正解から正解に変わったときの正解率の変化量は，その事例のクラスに依らず一定であるため，単に事例数の多い方のクラスに分別するだけで高正解率が得られ，局所解に陥りやすい．SESAT の評価実験でもベンチマークデータの 2 クラス分別を行ない，既存システムと同程度の正解率が得られているが，各クラスの事例数比は大きくても 1:2 であった．ポリマー分別では 2 クラスの事例数の偏りが大きいため，決定木構成子の適応度算出法にも工夫が必要である．

以上より，SEPT では SESAT の sprig の構造と決定木構成子の適応度算出法を改変し，決定木構成子の構造と進化方法，世代交代モデルについては SESAT と同様とする．分別対象とする特定のポリマーをクラス 2，それ以外のポリマーをクラス 1 として分別を行なう．



(a) 表現型　　(b) 遺伝子型

図 1: sprig

3.2 sprig

部分解 sprig を図 1(a) に示すような高さ 1 の部分木で表し，根ノードを属性ノード，葉ノードをクラスノードと呼ぶ．事例に出現する属性数が A であるとき，属性ノードには属性番号 $At(=1,2,\cdots,A)$，左右 2 つのクラスノードにはそれぞれクラス番号 Cl_l，$Cl_r(=0,1,2)$ が入る．sprig が決定木に組み込まれる際，クラス番号が 0 のクラスノードは非終端ノードとなり，別の sprig の属性ノードが接続される．

sprig には閾値 Th が与えられている．事例の属性値に従って属性ノードからクラスノードへと走査するときは，属性値が閾値以下の場合には左部分木へ，閾値より大きい場合には右部分木へと進む．

sprig を表す染色体は，図 1(b) に示すような 32 ビットの 0 と 1 の並びで表現する．第 i ビット目の値を g_i，属性値の最大値を v_{max}，最小値を v_{min} と表すと，At，Th，Cl_l，Cl_r は次式で求められる．%は整数除算の剰余を求める演算子とする．

$$At = \sum_{i=1}^{18} 2^{18-i} g_i \ \% \ A + 1 \tag{2}$$

$$Th = \frac{v_{max} - v_{min}}{2^{11} - 1} \sum_{i=19}^{28} 2^{28-i} g_i + v_{min} \tag{3}$$

$$Cl_l = g_{29}(g_{30} + 1) \tag{4}$$

$$Cl_r = g_{31}(g_{32} + 1) \tag{5}$$

最初の第 1〜18 ビットで表される数を属性数で割った余りにより属性番号を算出する．第 19〜28 ビットは，事例に出現する属性値の範囲のどのあたりに閾値を設定するかを表している．第 29，31 ビットが 0 のときはそれぞれ左右の子ノードが非終端ノードであることを表し，1 のときは終端ノードであることを表す．子ノードが終端ノードのとき，第 30，32 ビットに 1 を加えた値がクラス番号となる．

ランダムに設定した指定個のビット列を初期集団の個体とし，SESAT と同様に適応度を求めて 1 点交叉と突然変異により進化させる．

3.3 決定木構成子の適応度

SESAT では決定木構成子 T の適応度 $tfit(T)$ を正解率と正解局在率 $bias(T)$ から算出する．しかし，SEPT では 2 クラスに分別する決定木を生成するため，クラス 1 とクラス 2 の訓練事例数に偏りがある場合，正解率を用いると訓練事例数が少ないクラスの特徴が決定木に反映されにくくなる．

両クラスの特徴を反映した決定木を生成するために，SEPT では正解率の代わりにクラス 2 の再現率 $rec(T)$ と適合率 $pre(T)$ を用いて適応度を算出する．再現率はクラス 2 の事例をクラス 2 として判断できた割合を表し，適合率はクラス 2

表 2: パラメータ

パラメータ	値
突然変異確率	0.01
sprig 集団の個体数	800
決定木構成子集団の個体数	800
世代交代回数	50000
木の高さの上限値	5

表 3: 全事例分類の正解率と平均ノード数

データ	SEPT			SESAT		
	正解率 [%]		ノード数	正解率 [%]		ノード数
	最高	平均		最高	平均	
PS	99.6	98.9	11.8	98.8	97.0	7.9
PE	99.6	99.6	11.6	99.6	96.3	10.1
ABS	97.3	96.7	15.0	94.2	85.2	3.7
PP	100.0	99.5	10.6	99.6	95.6	9.1
PC	100.0	100.0	13.4	97.7	96.9	4.6
PVC	99.6	99.3	13.0	96.5	95.7	2.4
POM	100.0	100.0	7.0	99.6	99.6	6.6
PC/ABS	99.6	99.3	9.6	96.5	90.7	4.6
AS	100.0	99.7	13.4	95.0	94.1	1.4
PMMA	100.0	100.0	9.2	99.6	98.1	4.4
PET	99.6	99.2	7.8	99.2	98.6	5.8
PBT	100.0	100.0	7.6	98.8	98.8	1.0
UF	100.0	100.0	5.0	100.0	99.0	2.0
MF	100.0	100.0	3.0	100.0	99.9	5.6
PF	100.0	100.0	5.0	100.0	100.0	4.4
PA66	100.0	99.9	10.8	97.7	97.7	1.4
PA6	100.0	99.9	10.0	99.6	99.3	7.6
CEL	100.0	100.0	5.0	100.0	99.1	2.0

と判断した事例のうちクラス 2 の事例が占める割合を表す．ここで，決定木構成子 T でクラス 2 と判定された事例数を $f_2(T)$，クラス 2 の事例数を r_2，決定木構成子 T でクラス 2 と判定されたクラス 2 の事例数を $rf_2(T)$ とする．

$$rec(T) = \frac{rf_2(T)}{r_2} \tag{6}$$

$$pre(T) = \frac{rf_2(T)}{f_2(T)} \tag{7}$$

$$tfit(T) = rec(T) \cdot pre(T) \cdot (1 - 0.2 \cdot bias(T)) \cdot 100 \tag{8}$$

適応度を正解率を用いて算出する場合と比較すると，再現率と適合率の積を用いる場合には，正解局在率の変動が適応度に及ぼす影響が小さくなる．従って，上式の適応度に基づいて個体を進化させると，簡素な決定木の生成を目指した SESAT とは異なり，事例数に偏りのある 2 クラスを正確に分別することに重点を置いた学習を行なうことができる．

4. 評価実験

SEPT および SESAT において，表 1 の試料を分別する決定木を各種類ごとに 10 回ずつ生成した．設定したパラメータを表 2 に示す．生成された決定木の最高正解率，平均正解率，平均ノード数を調べたところ，表 3 のような結果が得られた．

いずれの種類のポリマーに関しても，SEPT の正解率は平均，最高ともに SESAT 以上となっている．クラス 2 に属する試料が極端に少ない場合は，クラス 1 を表す 1 つの終端ノードのみからなる決定木によって高正解率が得られるため，正解率から適応度を求める SESAT では，PBT，UF，PA66，CEL などでノード数が 1 の決定木が多く生成された．一方，クラス 1 の終端ノードのみからなる決定木は再現率が 0 なので，SEPT ではそのような決定木の生成を抑制することができ，より正解率の高い決定木を生成することができた．従って，再現率と適合率の積による適応度は，事例数に偏りのある 2 クラスを分別する決定木の生成に有効であるといえる．

また，最高正解率と平均正解率の差からわかるように，SEPT の正解率は SESAT に比べて分散が小さく，高正解率の決定木を安定して生成できた．sprig の設計を変更することでより多様な解が生成され，適応度の高い解が見つけられたものと考えられる．

次に，試料数が 30 以上の PS，PE，PP について，試料の 10 分の 9 を訓練事例，10 分の 1 をテスト事例とする 10-fold クロスバリデーションを行なった．各テスト事例に対して試行を 10 回ずつ繰り返したときの平均正解率と平均ノード数を表 4 に示す．この結果，提案手法は未知事例分類に関しても有効であることが確認された．

表 4: 未知事例分類の平均正解率と平均ノード数

データ	SEPT		SESAT	
	正解率 [%]	ノード数	正解率 [%]	ノード数
PS	95.4	10.2	94.1	7.3
PE	96.0	12.3	94.7	10.8
PP	96.9	11.2	93.0	7.4

5. おわりに

属性数が非常に多く，事例数に偏りのあるデータを 2 クラスに分別するための共生進化に基づく決定木生成手法を提案した．実測データを用いた評価実験により，必要以上に 1 つのクラスに分別するような決定木生成を回避し，正解率の高い決定木が生成できることが確認された．今後は，さらなる正解率向上，および多クラス分別のための手法について検討していく．

参考文献

[Breiman 84] Breiman, L., Friedman, J., Olshen, R., and Stone, C.: *Classification and Regression Trees*, Wadsworth & Brooks (1984)

[貝原 03] 貝原, 樋口：決定木を用いたポリマーの判別, 日本化学会第 26 回情報科学討論会要旨集, J08 (2003)

[松本 99] 松本 他：近赤外分光測定とニューラルネットワーク解析を組み合わせたプラスチック廃棄物の非破壊判別, 分析化学, Vol. 48, No. 5, pp. 483–489 (1999)

[Moriarty 96] Moriarty, D. and Miikkulainen, R.: Efficient Reinforcement Learning through Symbiotic Evolution, *Machine Learning*, Vol. 22, pp. 11–32 (1996)

[小野寺 99] 小野寺 他：スペクトル揺らぎを利用したプラスチックの近赤外反射スペクトルの特徴抽出, 化学ソフトウェア学会論文誌, Vol. 5, No. 3, pp. 93–102 (1999)

[大谷 04] 大谷, 志村：共生進化に基づく簡素な決定木の生成, 人工知能学会論文誌, Vol. 19, No. 5, pp. 399–404 (2004)