MAGICIAN 養成講座

MAterials Genome/Informatics and Chemo-Informatics Activate Networks

第 7 回　遺伝的アルゴリズムを用いた最適解探索

2018.11.8　横浜国大非常勤講師　山本博志

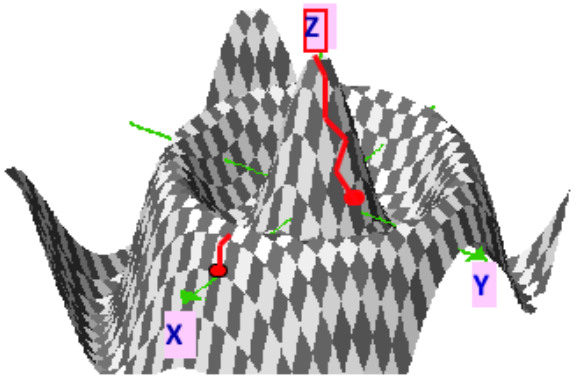
Materials Genome という言い方をした時、Genome:遺伝子はどのように扱うのだろうか？　医薬品でゲノム解析と言った場合には、ある遺伝子特有の病気に対して、オーダーメード医薬品を作るようなイメージで使われる。マテリアルでゲノム解析と使う場合、例えば触媒を考える時に、触媒の性能を司る“遺伝子”を特定して、性能を向上させるためには“どう進化”させたら良いのかを考える。例えば、馬からサラブレッドを作る、美味しいお米を作る、病害虫に強い作物を作るのと同じようなものだ。進化した生物は雌雄を持つものが多い。これは親の遺伝子を組み合わせて、性質の良いとこ取りするのに適している。触媒の良い遺伝子と言った場合に、例えば、転化率が高い、選択率が高い、寿命が長い、コストが安いなどを目指して、処方、プロセスを最適化していく。このように、目的変数が複数ある場合の最適化を考える時、昔から使われているのが、遺伝的アルゴリズム法というやり方だ。これは、簡単にいえば“適者生存の法則”をコンピュータ上に載せたものだ。レースに勝つ馬の遺伝子がどうなっているかは置いておいて、レースに勝つ馬同士を掛け合わせて子供を産ませればサラブレッドができる。触媒の転化率がいくら高くても、選択率が低ければ適者でない。そこで、転化率や選択率を予測する式をニューラルネットワーク法で作成し、遺伝的アルゴリズム法で最適化するという方法が成立する。ニューロンや遺伝などという言葉が入っているので、まるでバイオの世界（実際、昔その手の書籍は生物のコーナーに置いてあった）だが、医薬品の設計から、マテリアルの設計まで広く使われている。今回はこの遺伝的アルゴリズム法を学ぼう。

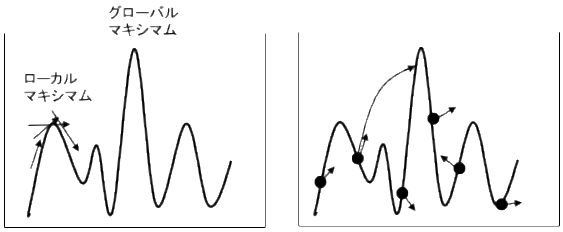
遺伝的アルゴリズム法

遺伝的アルゴリズム法(GA: Genetic Algorithm)は 1975 年に提案された近似解を探索するアルゴリズムである。アルゴリズムというのは、“考え方”のことである。そこで、汎用のニューラルネットワークのシミュレータは存在するが、汎用の GA のソフトは、自分は知らない。“考え方”に従って、自分でプログラムを実装していかなければならないので、広く一般的には使われていないのだろう。

GA の考え方の例題として、一つの山の頂点を探すことを考えて見よう。自分のいる地点の傾斜を見て、登りの方向に常に進めば、何時かは頂点（どの方向へ進んでも下りになる）にたどり着く。

しかし、そうした微分係数を用いた探索方法では、多峰性の山登りには使えない。初期値をどこに取ったかに依存してしまう。また、評価関数が微分できないと傾きがわからないので進む方向がわからない。





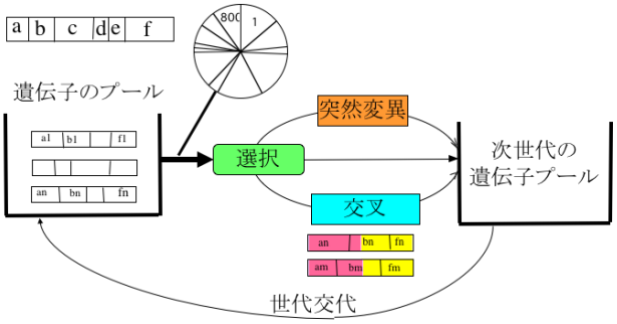
GA では、多次元に散りばめられた多くの評価点から傾きを使わずに探索を行う。

完全にランダムな方向に動かす、ランダム探索では、特に次元数が増えてきたときに近似解の探索が非常に困難に

なる。例えば、宅配便のトラックが 50 個の荷物を配達する事を考えてみよう。各家を回るルートの数は 50！(50*49*　*2*1)存在する。そのルートを最短に（ガソリン・コストを最小に）するために、50！の距離を全部計算して最短距離を探すのはドライバーには不可能であるし、スパコン使っても難しい。ましてや、その１軒が再配達となったら、その１軒に戻るルートも含めて最短距離を再計算しなくてはならなくなる。囲碁や将棋の次の数手先を読むなども同じような問題を抱える。

そこでランダムに動かすのでは無く、”適者生存の法則”で動かすのが GA である。例えば、触媒の組成を遺伝子に見立てる。（本来はスキーマ理論というのがあって遺伝子は 0 か 1 のビットで表す。しかし、化学の現象はビットでは表しにくいので、%の実数で計算を行う。）

そうした遺伝子を多数用意し遺伝子のプールに入れる。



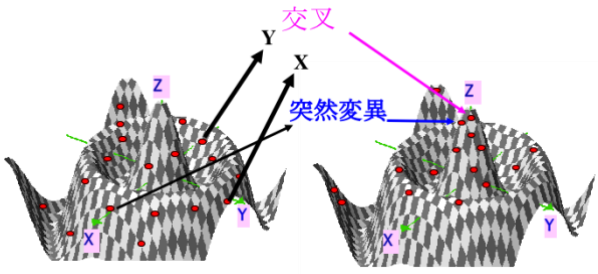
そして、各遺伝子の評価値を計算する。

例えば、次のような評価式を自分のニーズに合わせて作る。

E= 30*転化率 + 50*選択率 + 100*寿命

これらの各係数は自分の考えた“適者”によって様々な値になり得る。転化率や選択率は少し余裕があるので多少落ちても寿命が長くなってほしいなどによって評価式を作る。そして各遺伝子の評価値を計算したら、ルーレットを作成する。このルーレットのポケットのサイズは、評価値が高いものほど広くとる。つまり適者ほど親になる権利が高くなる。そして、ルーレットを回して、遺伝子を二つ選択する。そして、ある確率で突然変異と交叉を行って、次世代の遺伝子プールにためる。次世代の遺伝子プールに初期と同じ数の遺伝子が入ったら、世代交代をして、同じ事を繰り返す。

突然変異とは、ある元素の組成が変わることに相当する。メキシカン・ハットの山登りの例では、Y の値は良いが、X の値が悪いので評価が低いが、X の値が突然変異すると高い評価になる。



交叉は、遺伝子のある部分を切ってもう一つの遺伝子とつなぎ変えてあげる。X○Y✖と X✖Y○を繋ぎかえると X○Y○と X✖ Y✖の 2 つの遺伝子になる。

どのくらいの頻度で突然変異と交叉を行うか？は難しい問題である。基本的には突然変異と交叉の操作を行うと遺伝子的には悪い遺伝子になる。生物の進化が何百万年もかかるゆっくりとした進化なことからも明らかであろう。詳しい事は自分もわからないが、癌細胞のような悪い遺伝子を持った細胞が優性遺伝子なら人類はもっと早く滅亡していただろう。そこで、一番性能の高い遺伝子は、突然変異と交叉を行わず必ず次世代に残すという操作を取り入れることがある。これをエリート保存というが、必ずしも良い結果になるという保障は無い。そのエリートに引きずられてローカルミニマムから抜け出せない事もある。

本来、ビットで表せば

突然変異　11100101 → 11110101

交叉　10100101 → 11110101

1111 0100 → 10100100

となり余計な操作は必要ないのだが、組成を実数の%で表しているので、突然変異や交叉の後、%に直す。その際には他の元素の組成も変わってしまうので、収束は遅くなる。世代を繰り返すごとに動かす絶対量を小さくするなどの工夫が必要であるが、余り気にしなくても、計算負荷は大きくないので、取り敢えずこれで良しとする。

この部分にこだわるより、この元素は高価なので、使用量は 30%まで（30%を超えたら評価値を下げる）とか、レアーアースは使わないでとか、化学者ならではの評価関数の作成に労力を集中した方が良い。

言葉で表すと以上のように難しいので、実例を示そう。本来“遺伝的アルゴリズム”と言ったときに、アルゴリズムとは考え方を示すものなので、汎用の遺伝的アルゴリズム（GA）・ソフトウエアーは無い。現象ごとに GA ソフトを自分で組み立てることになる。これが GA を勉強する上でネックになる。

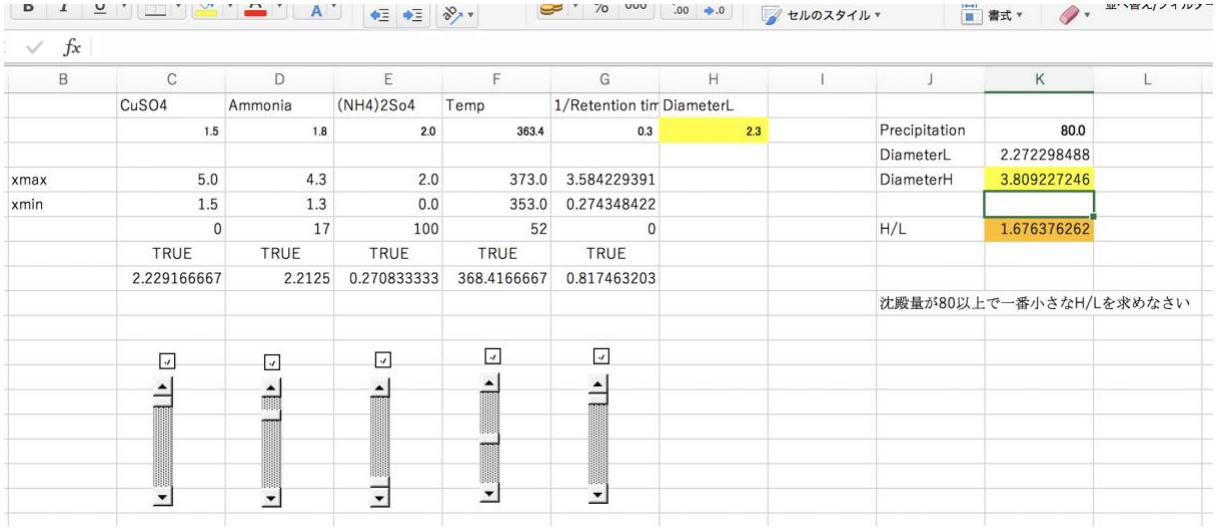
GA 検討の実例

今回は、まず、酸化銅のナノ粒子の合成条件の最適化を、GA を用いて行ってみよう。これは、次の反応で酸化銅のナノ粒子を作成する。

$$Cu(NH_3)_n(H_2O)_{6-n}^{2+} = CuO + nNH_3 + 2H^+ + (5-n)H_2O$$

日本金属学会誌, 62, no.6, 1998, 519-526

様々な実験条件の時に、ナノ粒子の収量と粒子径が記載されている。粒子径は範囲で与えられているので、小さい方を L、大きい方を H として、沈殿量、粒径 L、粒径 H の 3 つを予測する式を作成する。式の作成自体は授業でも取り上げじっくりやったのでここでは省略する。そのような式ができると、ある実験条件の時に目的変数である沈殿量、粒径 L、粒径 H がどうなるかを予測できるようになる。



	B	C	D	E	F	G	H	I	J	K	L
		CuSO4	Ammonia	(NH4)2So4	Temp	1/Retention tim	DiameterL				
		1.5	1.8	2.0	363.4	0.3	2.3		Precipitation	80.0	
									DiameterL	2.272298488	
	xmax	5.0	4.3	2.0	373.0	3.584229391			DiameterH	3.809227246	
	xmin	1.5	1.3	0.0	353.0	0.274348422					
		0	17	100	52	0			H/L	1.676376262	
		TRUE	TRUE	TRUE	TRUE	TRUE					
		2.229166667	2.2125	0.270833333	368.4166667	0.817463203					
									沈殿量が80以上で一番小さなH/Lを求めなさい		

授業の際には上記のようなシミュレータを構築した。実験条件をエクセルのスライドバーで動かし、“沈殿量が 80 以上で一番小さな H/L を求めなさい”が課題だった。H/L が一番小さいというのは、粒径が揃っている事の指標である。こうした酸化銅のナノ粒子は、インクジェットプリンターなどでナノ回路を印刷し水素で還元して金属銅にして使われる。粒径が揃っているというのは、ノズルが詰まらないために重要な指標だ。粒径が揃うと沈殿量が少なくなったりするので、手作業でやるのは大変だろうというのを実感してもらうために学生に競ってもらったのだが、自分が事前に見つけておいた実験条件と結果はあっさりクリアーされてしまった。

そうすると、GA をわざわざ組む必要も無いので、今回はさらに、粒径 L が 2.5μm 付近という条件をつけよう。

実験条件 5 つを動かし、粒径 L が 2.5μm で、沈殿量 80 以上、L/H がなるべく 1 に近い実験条件を求めなさいが課題だ。（実際の製造では、製造コストが最小に、とかさらに色々な制約条件がかかってくる。）

目的変数が 3 つ以上になると手作業で行うのは段々と難しくなる。（それにしても、真面目で優秀な横浜国大の学生はできてしまうのが実に問題だ。GA などは、無精者の科学で、めんどくさいという感覚を持たない真面目な学生には縁遠いのかもしれない。）

それでは、実際に GA ソフトをエクセル上に組んでみよう。まず、最初の遺伝子のプールを作成する。遺伝子を何個用意したらいいだろうか？　さしあたって、論文に記載されている実験条件 48 種類を初期遺伝子にする。実験条件範囲内で、乱数で初期遺伝子を作成しても良い。探索する実験条件の上限値、下限値は定めておいたほうが良い。それは、推算式の適用範囲の問題があるからだ。基本的には推算式を構築した際の実験条件を上下 20%以上逸脱する場合には余り高い推算精度は期待しないほうが良い。実際には個々の説明変数の上限、下限だけでなく、組み合わせの上限、下限の問題もあるので複雑ではあるが、ここでは論文にある実験条件の範囲内で動かすことにする。

C	D	E	F	G	H	I	J	K	L	M	N	O
CuSO4	Ammonia	(NH4)2SO4	Temp	1/ks	DiameterL	Precipitation	DiameterH	L/H	pre/80	(DL-2.5)^2	Evaluation	
2.000	3.000	0.000	373.000	0.274	3.17	88.49	5.75	0.551	1.106	0.449	2.745	
2.000	3.000	0.000	373.000	0.278	3.17	88.45	5.76	0.551	1.106	0.456	2.742	
2.000	3.000	0.000	373.000	0.366	3.29	87.52	5.98	0.551	1.094	0.626	2.660	
2.000	2.500	0.000	373.000	0.547	2.57	90.20	5.13	0.500	1.128	0.005	2.796	
4.000	3.400	0.000	373.000	0.278	2.70	89.21	5.65	0.478	1.115	0.040	2.701	
2.000	2.500	0.000	373.000	1.067	3.25	82.35	6.40	0.509	1.029	0.568	2.483	
2.000	2.500	0.000	353.000	0.278	2.56	74.76	5.03	0.509	0.935	0.004	2.601	
2.000	2.000	0.000	373.000	1.613	3.01	76.15	6.44	0.468	0.952	0.262	2.393	
4.000	3.400	0.000	353.000	0.278	3.05	58.32	6.19	0.492	0.729	0.299	2.195	
2.000	3.000	0.000	373.000	0.543	3.52	85.44	6.41	0.550	1.068	1.049	2.459	
2.000	2.500	0.000	373.000	0.275	2.21	93.28	4.48	0.494	1.166	0.083	2.790	

黄色でマークした部分が実験条件になる。その時の粒径 L、沈殿量、粒径 H（緑でマーク）は予測値になる。差し当たってこの予測値は十分な精度を持つとする。すると、今回評価したい物性値（オレンジ色でマーク）L/H(粒径 L/粒径 H), Pre/80(沈殿量/80), (DL-2.5)^2(粒径 L から 2.5 を引いた値の 2 乗)が計算できる。L/H は 1 に近いほど優秀である。Pre/80 は 1 以上であるほど良い。(DL-2.5)^2 は粒

径Lが2.5μmから外れるほど急に大きな値になるので、なるべく小さな値になれば良い。この3つの評価値を組み合わせてEvaluation(青でマーク)を定義する。おそらくこのEvaluationの定義が一番難しく、一番面白いところだ。3つの評価値をどう捉えるかは実験を行う者の主観に基づく。Pre/80は1以上であればそれ以上大きな値を探索する必要がなく、L/Hを重視したいとか、粒径Lが2.5μmに近いものを優先したいとか主観に基づいて評価関数を定義する。

ここでは、

3*L/H+1.15*Pre/80-0.4*(DL-2.5)^2

で評価する。自分で高く評価したい評価値の係数を調整する。(DL-2.5)^2は小さいほど良いので係数はマイナスになる。

ここまで出来たら、Evaluationの列をキーにソートをかける。大きい順にソートをかけると、自分の目的にあった実験操作条件が上位にくる。本来のGAでは、このEvaluationの値をルーレットのポケットの面積に変換して、ルーレットを回して次世代の遺伝子のプールに入れていく遺伝子を決める。エクセルでそれを行うのは大変なので、そこは簡略化する。

=$C$2

C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
CuSO4	Ammonia	(NH4)2SO4	Temp	1/ks	DiameterL	Precipitation	DiameterH	L/H	pre/80	(DL-2.5)^2	Evaluation		procedure	CuSO4	Ammonia	(NH4)2SO4	Temp	1/ks
2.000	2.500	0.000	373.000	0.547	2.57	90.20	5.13	0.500	1.128	0.005	2.796		elete	2.000	2.500	0.000	373.000	0.547
2.000	2.500	0.000	373.000	0.278	2.22	93.26	4.48	0.494	1.166	0.081	2.791		CrossOver	2.000	2.500	2.000	353.000	0.278
2.000	2.500	0.000	373.000	0.275	2.21	93.28	4.48	0.494	1.166	0.083	2.790		CrossOver	2.000	1.300	0.000	373.000	0.547
2.000	3.000	0.000	373.000	0.274	3.17	88.49	5.75	0.551	1.106	0.449	2.745		CrossOver	2.000	1.400	1.500	353.000	0.278
2.000	3.000	0.000	373.000	0.278	3.17	88.45	5.76	0.551	1.106	0.456	2.742		CrossOver	2.000	2.500	0.000	373.000	0.547
4.000	3.400	0.000	373.000	0.278	2.70	89.21	5.65	0.478	1.115	0.040	2.701		CrossOver	2.000	2.500	0.000	373.000	0.556
2.000	3.000	0.000	373.000	0.366	3.29	87.52	5.98	0.551	1.094	0.626	2.660		CrossOver	2.000	1.500	1.000	373.000	0.278
2.000	2.500	0.000	353.000	0.278	2.56	74.76	5.03	0.509	0.935	0.004	2.601		CrossOver	2.000	2.500	0.000	373.000	0.278

Evaluationのカラム（N列）の後ろには空のカラム（O列）があるのでソートをかけても、P列以降は変更されない。そこで、二行目には評価関数が一番大きいものがある。その値は2.796であることがわかる。L/Hは0.5, Pre/80は1.128, (DL-2.5)^2は0.005である。この実験条件はエリート（一番優れている）なので、これは保存する。具体的には、Q2からU2には、C2からG2の値を入れる。その際には、=$C$2のようにセルの絶対位置を指定する。3行目以降は、交叉(Cross Over), 突然変異(Mutation)の操作を加える。本来はルーレットを回して、交叉、突然変異をおこなう行を決めるのであるが、ここでは固定行を使う。Q3とQ4のセルを見てみよう。

Q3=$C$2, Q4=$C$29となっている。

つまり、三行目には一番評価関数の高い二行目が代入され、四行目には29行目が代入されている。29というのは適当に選んだ数字だ。上位にあるもの同士を交叉すると、多様性が確保できなくなり、ローカルミニマムに収束してしまう事が多いので、適当に悪い遺伝子（行の下の方）と交叉させてやる。ここで、R列までは2行目、29行目を絶対位置指定で持ってくるが、S-U列では2行目、29行目を入れ替える。入れ替える位置によってオレンジ色でマークしてある。5列ある所のどこで入れかえるかは、本来は乱数で決めるのだが、固定位置にしてある。Q-U列では2行目、29行目の遺伝子が入り混じった遺伝子が作成されていることになる。上の画面には表示されていないが、27行目以降には突然変異の操作が示されている。

Q27=$C$2+($C$53-$C$2)*RAND()*0.3

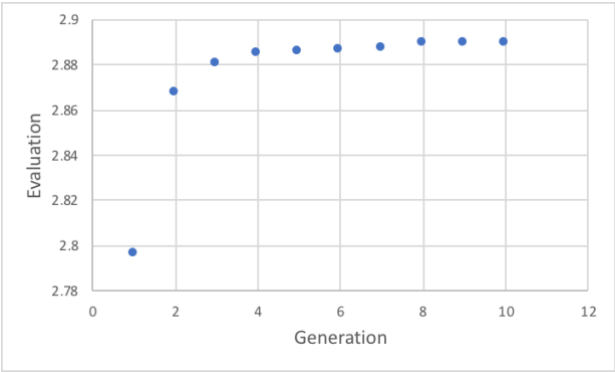
Q28=$C$2+($C$52-$C$2)*RAND()*0.3

$C$2はエリートの値になっている。$C$53はC列の最大値、$C$52は最小値が入っている。

($C$53-$C$2)*RAND()*0.3はC列の最大値からエリート値を引いたものに0.000-1.000の乱数をかけ、さらに0.3を掛けた値分だけエリート値を大きい方へ動かすことに相当する。逆にQ28では、小さい方に動かすことに相当する。どの列を突然変異させるかは、本来は乱数で決めるが、交叉と同様、固定位置で突然変異を行う。

ソートをかけた段階で、最初のプールの値（C-G列）は確定するので、次世代プールの値も確定される。そこで、次世代プール(Q-U列)にある48組の遺伝子をコピーし、C-G列に“値のみ”のペーストを行う。すると、H—Jの物性予測値はすぐに計算される。そこで、再び、Evaluationの値でソートをかける。評価関数の値は、2.796から2.867に改善される。（乱数を用いているので、ここから先は計算毎に異なるので同じ値にならなくても気にすることはない。）

L/Hは0.5から0.517, Pre/80は1.128から1.145, (DL-2.5)^2は0.005から0.002になる。実験条件的には、アンモニアの量と1/ksの値が変わっている。これが第1世代になる。そしてその時の「Q-U列をコピーしてC-G列に値のみのペーストをし、ソート」を繰り返す。



4 世代ぐらい繰り返すと大体収束してしまう。
10 世代目には、L/H は 0.5 から 0.571, Pre/80 は 1.128 から 1.035, (DL-2.5)^2 は 0.005 から 0.037 になる。世代交代を繰り返すと、上位の実験条件は非常に似たものになってくる。(10 世代目ではトップの 5 つが同じになる。)そこで、解の多様性を確保するためには、交叉をするペアの一つは行番号の大きなものを選ぶ方が良い結果を与える。評価関数の係数を変化させたり(この例では、L/H が大きいものが優先されている)、交叉、突然変異の配分などを変えて、GA をどう設計すれば自分の欲しい答えが見つかるか、試してみよう。粒径の評価値を(DL-2.5)^2 から ABS((DL-2.5)に変えて評価してみよう。

エクセルのマクロを設定できるなら、ボタンを作って、ボタンが押されたら、コピー、値をペースト、ソートまでの自動化をしてみよう。ボタンが押されるたびに世代を交代できるようになるので、計算は非常に楽になる。
GA の基本形はこのエクセルでやったことが全てなので、目的変数、評価関数を書き換えるだけで他の事象に応用できるので、色々試してみよう。

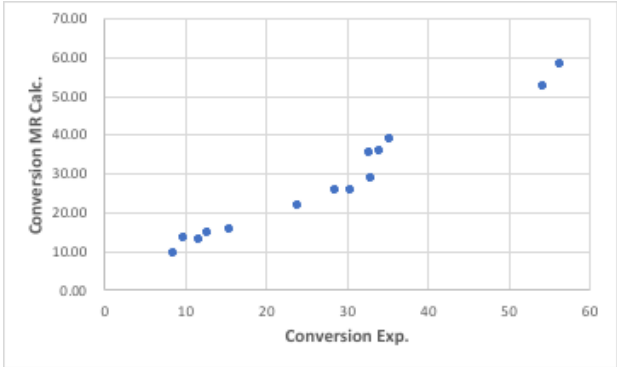
GA の例その 2

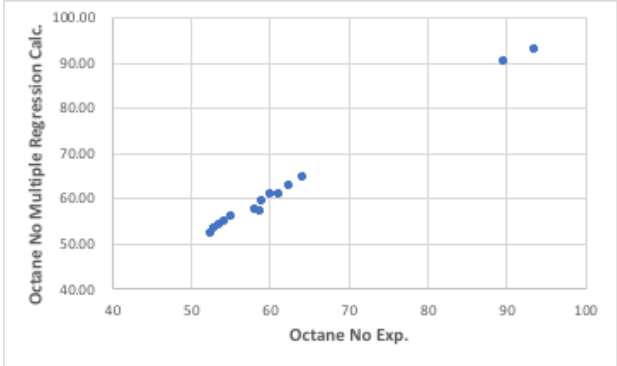
次はゼオライトによるナフサのリフォーミングを検討してみよう。実験条件 (Temp., LHSV, Feed Rate, P, Time) を変えた時の転嫁率(Conversion), オクタン価が論文に記載されている。Pak. J. Chem. 4(4): 167-184, 2014
ナフサはオクタン価が低いので高機能ガソリンを作る上で重要なプロセスだ。
授業では、転嫁率とオクタン価の予測式を作成した。

Run	OctaneNO	Temp	LHSV	FeedRate	P	Time	Conversion	OctaneNO	Conversion MR	OctaneNo MR
1	52.6	350	8	0.73	8	75	8.68	52.6	9.25	52.07
2	52.9	400	8	0.73	8	81	15.44	52.9	15.34	53.49
3	54.2	450	8	0.73	8	90	23.9	54.2	21.64	54.94
4	55.2	480	8	0.73	8	96	28.65	55.2	25.47	55.81
5	53.6	480	3.5	0.73	3.5	76	9.88	53.6	13.35	53.90
6	58.2	480	12	0.73	12	102	32.85	58.2	35.40	57.42
10	58.9	480	3.5	0.55	3.5	103	11.74	58.9	13.11	57.21
11	59.1	480	8	0.55	8	131	30.6	59.1	25.80	59.19
12	61.2	480	12	0.55	12	138	34.12	61.2	35.80	60.80
7	60.2	480	3.5	0.37	3.5	155	12.82	60.2	14.65	60.72
8	62.5	480	8	0.37	8	202	33.1	62.5	28.70	62.85
9	64.3	480	12	0.37	12	209	35.34	64.3	38.70	64.46
13	89.6	500	2	0.28	12	306	54.36	89.6	52.59	90.33
14	93.5	500	1.5	0.24	12	386	56.43	93.5	58.12	92.81

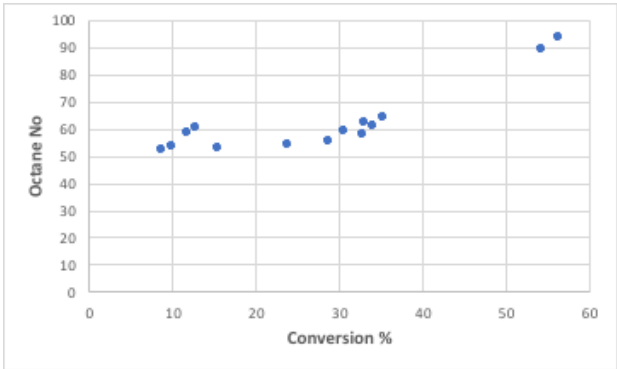
目的変数を入れ替えて重回帰計算する

目的変数(緑でマーク)は2つあるので入れ替えながら、重回帰計算を行い、計算値(青でマーク)を得た。





グラフ表示してみると、実験条件と実験結果は重回帰式で良好に記述できている事がわかる。



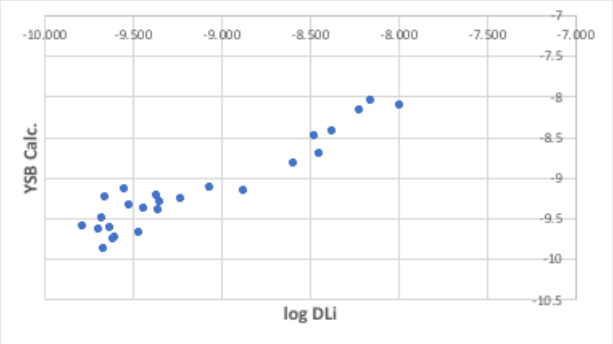
転嫁率とオクタン価には上図のような関係がある。
そこで、GA で求めたい答えは、重回帰式は正しいとして、転嫁率とオクタン価の両方が高くなるような実験条件を求めることになる。説明変数は 5 つなので、例題 1 で作った GA 計算のシートがそのまま流用できる。実験件数は 14 件なので、実験条件の範囲内で 48 件まで水増ししてあげる。目的変数の数は 2 つなので、1 列削除する。
さて、問題は評価関数だ。”転嫁率とオクタン価の両方が高くなる”を関数で表すにはどうしたら良いだろうか?
自分で考えてみよう。

GA の例その 3

この講座の No.5 データのクレンジングで Li イオンの

拡散係数を取り扱った。そこではニューラルネットワーク法で予測した拡散係数を GA でさらに拡散係数の高いものを探索した。拡散係数自体は組成に対して非線形であるので、ニューラルネットワーク法を用いないと正しく記述できない。ニューラルネットワーク法を用いた推算式は、GA で用いようとすると、プログラムを書けるか、マクロを書けるかしないと扱えないので例題としては使えない。そこで、原子種を絞って（O, Li, Al, Si, Nb）YSB を用いてクロスタームを導入した予測式を作成する。

	A	B	C	D	E	F	G	H	I
1	ICSD No.	Log D_{Li}	O	Li	Al	Si	Nb		Calc.
2	32594	-7.993	57.97	13.04	14.49	14.49	0.00		-8.1032876
3	24896	-8.157	59.14	11.30	13.01	16.56	0.00		-8.0513565
4	16229	-8.222	40.00	50.00	10.00	0.00	0.00		-8.1679928
5	81249	-8.469	44.44	44.44	0.00	11.11	0.00		-8.4825023
6	109053	-8.446	50.00	37.50	0.00	0.00	12.50		-8.7081772
7	35252	-8.372	47.37	42.11	0.00	0.00	10.53		-8.4271648
8	25752	-8.587	46.67	40.00	0.00	13.33	0.00		-8.8205752
9	84449	-9.066	57.17	14.24	14.29	14.29	0.00		-9.1239145



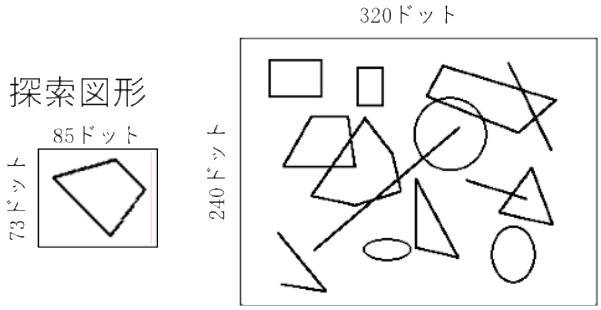
すると、上図のように拡散係数が高い領域では、そこそこ精度の高い相関式が得られる。
説明変数は 5 つであるので、前の GA のエクセルがそのまま使える。評価関数は目的変数が Log DLi の 1 つで、大きいほど良いので設定は簡単だ。

この場合、問題になるのは、5 つの原子種の%を全部たすと 100%になるように調整しなければならない事だ。
交叉や突然変異させた後にトータルが 100 になるように変換する。(この場合は、酸素欠陥の問題は端折ってある)
具体的には、交叉や突然変異を行った結果の入っている列のさらに後ろの列に、100%で規格化した列を作り、そちらをコピーして、C-G 列に値のみのペーストを行う。
もう 1 つの問題は、5 種類全部を使ってはいけない事だ。さらに、Nb を使った場合には、Al, Si は必ず 0%にしなければならない。Li, O 以外に 1 つ以上は選択しなければいけない。
この条件を満たすように改造して、シートを完成させよう。

雑記

遺伝的アルゴリズム(GA)の汎用シミュレータは無い。それは評価関数や制限事項がケース・バイ・ケースになってしまうからだ。しかし、それを自分でなんとかできるのであれば、GA の動作原理を簡単に学ぶ事ができる。実際にはプログラムを書ければさらに高度な事が可能である。今回の例のように一つシートを作れば、使い回しができるのと同じで、プログラムに関しても一つ雛形を作れば後の改造は容易だ。ぜひチャレンジしてほしい。

自分が GA のプログラムを初めて書いたのは 2000 年頃だ。化学用のプログラムでは無い。(出典が何だったか、色々探しているが見つからない。きちんと引用文献をつけたいのだが。。。　誰かご存知の方がおられたらお知らせ願いたい。)
GA で解いた最初の問題は次の画像認識の問題だ。320*240 の領域に色々な図形が書かれている。コンピュータの認識できるのは、各ドットが 0（白）か 1（黒）かだけだ。その中に 85*73 の探索図形があるかどうか、その位置はどこかを回答させる問題だ。人間であれば簡単にどこにその図形があるか認識できるであろう。



最近の Google のディープ・ラーニングによる画像認識の走りのような技術開発だ。

簡単のため、図形の拡大、縮小は無視できるとする。
探索図形の点を θ 回転し，(Xm、Ym）に平行移動する計算は、

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \mathrm{Cos}\theta & -\mathrm{Sin}\theta \\ \mathrm{Sin}\theta & \mathrm{Cos}\theta \end{pmatrix} \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} + \begin{pmatrix} X_m \\ Y_m \end{pmatrix}$$

で表す事ができる。探索空間は、320(横)*240(縦)*360(回転)で 27,648,000 になる。これを全て計算して一致度が一番高いものを答えとするのは大変な事なのはわかるだろう。遺伝子型を[θ, Xm, Ym]として、乱数で作った遺伝子を 1000 個ばらまいて、一致度が高いものほどルーレットのポケットを大きくして遺伝的アルゴリズムで探索させると直ぐに答えは見つかる。図として保存された構造式やグラフをコンピュータに認識させる事ができるので必死

になってGAを勉強した。勉強が進むにつれて、画像認識よりは利用価値の高い化学の分野が多いことに気がついて、画像認識自体は余りやらなくなったが、先に述べたようにプログラム自体は使い回しが効くので、未だに変数名などはこのころの物を使い回している。

遺伝的アルゴリズムを使うか、ディープ・ラーニングを使うかは別として、「コンピュータは画像を 0,1 でしか認識できない」というのは非常な驚きと言える。子供にも簡単にできるような図形認識をさせるのに、膨大な計算をするか、膨大なタグ付けされた学習データを用意して学習させるしかない。しかも、ちょっとでも変化させると（デフォルメさせると）人間と違って応用は全く効かない。
拡大、縮小などは行列計算に簡単に組み込める。結局、コンピュータのできる認識は、0,1 と行列計算のような論理計算だけということになる。
最近、人工知能(AI)が非常に流行りだが、脳科学者や数学者は、機械には知能は無いと言い切っている。あるのは、統計と確率、論理計算と検索だけで、これは知能では無い。

このAIブームには限界が見えているし、直ぐに廃れるだろう。でも廃れるというのも間違いで、熟成期に入ると考えた方が良い。自分がGAをゴリゴリやっていた2000年頃は、第２次AIブームの終わりの頃でそのあとは下火になった。その頃から画像認識やディープ・ラーニングの基礎が熟成し始め、昨今のインターネット上のビッグデータが利用できるようになって一気に花開いたのが、現在の第３次AIブームだろう。今ここで撒かれたタネが熟成期にどう育って行くのか楽しみなところだ。少なくとも人間の子供程度の認識力は持って欲しいものだ。
人間の脳が図形を見たときには、図形の認識をする脳部位、記憶を司る脳部位、推論をする脳部位など様々な Neural Network が同時に活性化される。**Activate Networks** がキーになると自分は思っている。

では、その間、化学系の研究者は何もしないでただ待っているのか？
今回の例のように、論理的な評価関数を主観的に作る。化学者にしか理解できない制限事項の論理式を作る。それができれば、MAGICIAN、手品師のように問題解決できるようになるだろう。

Pirika マテリアル・ゲノム のページ
以下 PDF

第１回 イントロダクション 2018.8.23
第２回 データ収集と昔ながらのやり方 2018.8.24
第０回 物性推算と逆設計と呼んでいた時の話 2000.8.28
なんと 18 年前！
第3a回 ポリマー設計と３つのMI（その１） 2018.9.3
第3b回 ポリマー設計と３つのMI（その２） 2018.9.3
第 4a 回 MI に適した簡単なデータベースの利用法 2018.9.4
第4b回 複雑なポリマーのデータベース化 2018.9.7
　プレゼン用：MI を使う時のデータベース構築法 2018.9.11
　プレゼン用：複雑なポリマーの設計と DB 2018.9.15
第５回 データのクレンジング 2018.8.28
第6a回 ニューラルネットワーク法の初歩 2018.9.25
第 6b 回 ニューラルネットワーク法を使った Drug Design 2018.9.22
第７回 遺伝的アルゴリズム(GA)を理解しよう