

図.3 a は、第 1 主成分 (PC1) と第 2 主成分 (PC2) を 2 次元にプロットした図であり、同 b はそれに第 3 主成分 (PC3) を加えた 3 次元プロットの図である。

これらの結果より、ジェノタイプデータの数値特性としては、2 群はある程度分かれて分布しているものの、その境界は曖昧であり、やや単一の集団と化していると考えられる。また、数値特性を表す尺度である主成分についても、用いる主成分の数を増やした所で境界の不明瞭さを解消するのは困難であると思われる。

次に、「自己組織化写像 (Self Organization Map、SOM)」という手法を試した。これは、入力されたデー

タ自らの特徴と互いの相互作用のみによって、任意の次元へと写像させる人工ニューラルネットワークの 1 つである。例えば、本プロジェクトで扱うジェノタイプデータの様な複数の SNPs からなる多次元データを、各検体の特徴を反映した形で 2 次元空間に写像する事が可能である。つまりこの手法を用いれば、PCA では捉えきれなかった各 SNPs 間の相互作用を反映させつつ、緑内障群と健常対照群を効果的かつ簡易的に分類する事が期待される。

SOM の出力は離散的なデータ空間となり、主に格子状に“ユニット”と呼ばれる点を配置し、各ユニットが入力データの特徴を学習する人工ニューラルネットワークと直結された形で実装される。つまり、ユニット数が学習の特性に影響するために注意する必要がある、複数のパターンを試して総合的に検証した方が良いと考えられる。実際に、ジェノタイプデータに対してユニット数を複数試しながら解析を行った結果例を図.4 に示す。

これら 4 つの結果は、ユニットを正方向格子状に配置して、一辺の数を増加

させながら解析結果がどう変化するかという様子を見たものである。全体的にユニット数が増加しても、左側に赤で示された緑内障群、右側に青で示された健常対照群が寄る傾向があるものの、互いに上手く分離できていない検体が多い事も解る。また、SOM では上手く分離できるデータの場合、基本的に似た特性のデータ同士が集まる島状の固まりが複数できるが、このデータでは全体的に外縁周辺に連なって分布している。これは、SOM の解析能力では、緑内障群と健常対照群を明確に分離できていないために起きたものと考えられる。ただし、SOM は分類精度が良くは無いものの、PCA の様に比較的単一に近いまとまりにはなっておらず、両群間の差に関する何らかの特徴を捉えつつある事が示唆される。

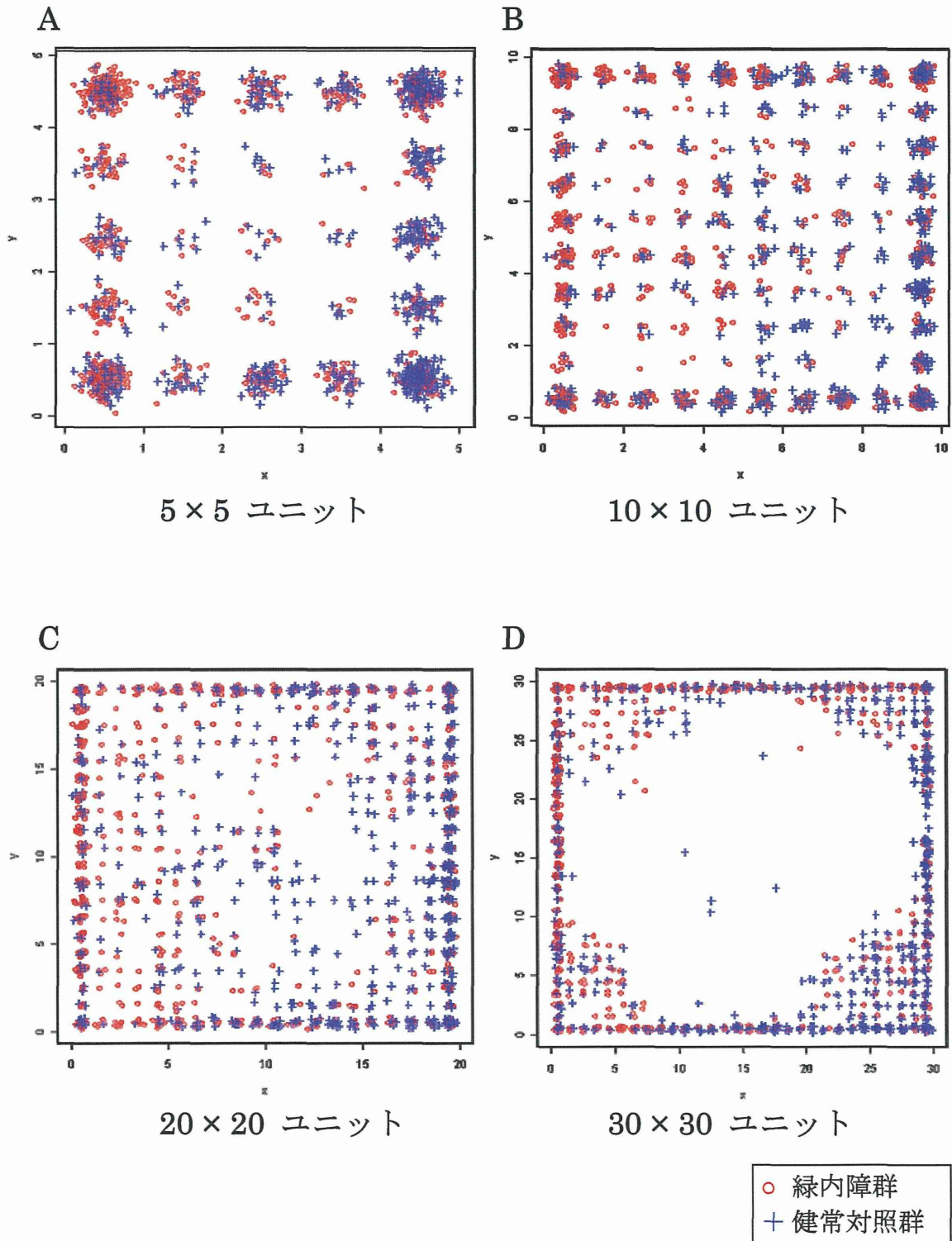
これら PCA・SOM による結果を踏まえ、単純な統計解析手法による2群の判別よりも、情報工学的・人工知能的手法により、積極的に2群を分類する手法を用いた方が良いと考えられる。特に、先のデータ変換法の検討時における LDA 結果（表.3）より、

GWAS 集団のみの学習による Custom 集団の診断結果が7割近くに達している事から、これを上回る結果を模索する事を目標とする。

また、GWAS 集団と Custom 集団は、実験に用いた SNP チップの製造元や形式が異なっている。つまり、この両者間の学習と診断で一定以上の効果を上げる事が出来れば、その様な差異によるバイアスに対してある程度堅固な性能を持つ手法と考えられる。これは、より複雑な環境要因の影響が懸念される医療現場で実際に運用される場合、各測定・検査の過程で様々なバイアスが混入し得る事を想定すると、重要視すべきことである。従って、これ以降の検討の基本方針は、GWAS 集団の特徴を学習した後、Custom 集団を診断した場合の精度を重点的に評価することとする。

先に述べた様に、表.3にてまとめた LDA の結果では、予測診断精度は7割程度であった。これは、対象となるデータが、線形の判別面によって2分される事を前提とした手法である。しかし、先の SOM 等の結果を見る限りは、線形よりも非線形の判別面の方が

図.4 SOM を用いたジェノタイプデータの解析結果



ジェノタイプデータの解析手法に沿っている可能性がある。そこで、非線形判別分析手法を試してみた。

具体的には、非線形判別分析の代表的な手法である Mahalanobis 距離を用いた解析を行った。これは、判別したい各群の群内分散と、それらの間の群間分散の両方を加味して各群の中心を規定し、各検体のデータについて中心からの距離を算出する方法である。この距離の事を Mahalanobis 距離という。算出方法は、各データの行列変換から求められ、本プロジェクトの場合は 165 SNPs のジェノタイプデータがこれに相当する。つまり、緑内障群と健常対照群の各特徴を最も表していると考えられる部分から、各検体がどの程度離れているのかを算出し、近い方の群をその検体の所属する群と考える。この時、群間分散・群内分散の算出に用いるデータを GWAS 集団のみに限れば、Custom 集団の各検体の距離を算出できるので、結果的に学習と予測診断という流れになる。

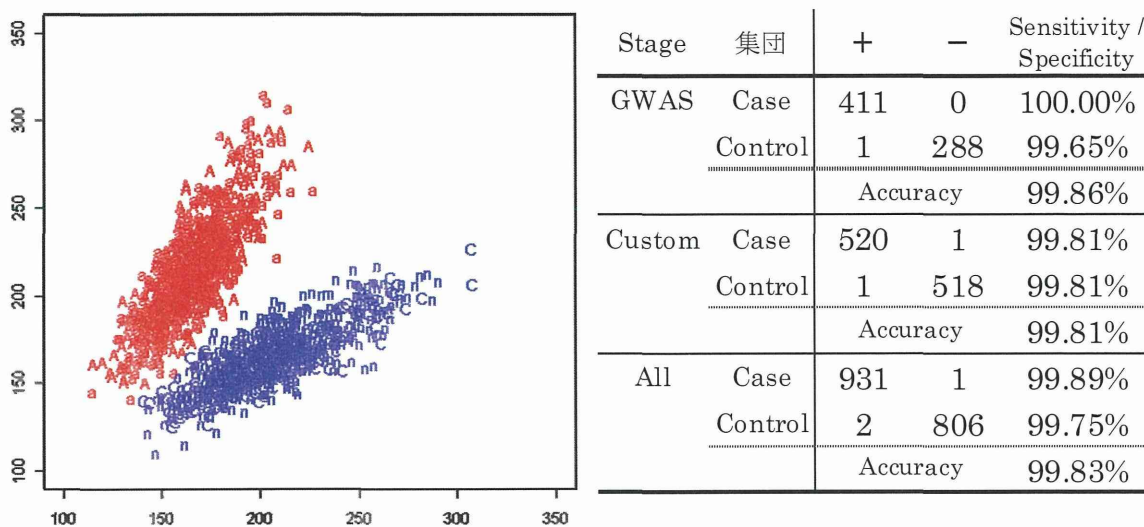
本プロジェクトのジェノタイプデータに対して、Mahalanobis 距離を算出した結果を図.5 に示す。まず、図中

A より、全検体を用いた学習とテスト結果では、GWAS・Custom 共にほぼ 100% の予測診断率を達成している。特に図を見れば、赤で示した緑内障群と青で示した健常対照群とが、ほぼ対角線上で二分されている事が解る。この結果より、ジェノタイプデータの特徴は Mahalanobis 距離で判別可能なものである事が解る。しかし一方で、現実的な診断を模した予測診断実験結果である B では、学習検体の判別は 100% 出来ているのに対して、テストデータでは Specificity がほぼ 0% に近づいている。これは、学習が非常に緑内障群の特徴を捉える方向で進んでしまったため、本来健常対照群である検体のほとんどを陽性と誤判定してしまった事による。以上をまとめると、A より非線形判別分析による精度向上の可能性は高まったものの、一方で Mahalanobis 距離を用いた方法はジェノタイプデータの解析に合致しない事が解った。

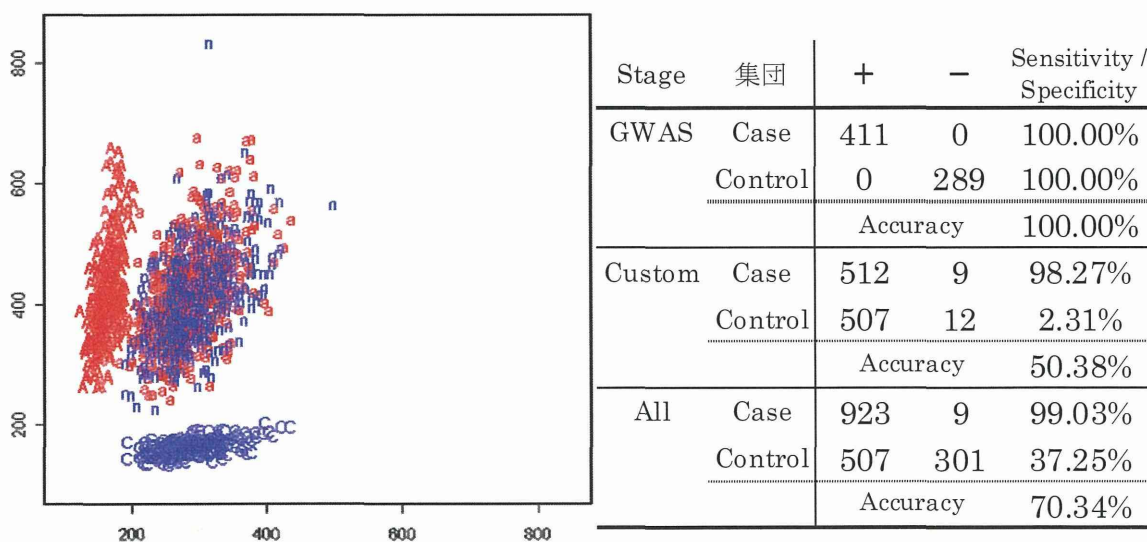
次に別の非線形判別的なアプローチを可能とする「サポートベクターマシン (Support Vector Machine、SVM)」を試した。この手法は、ジェ

図.5 Mahalanobis を用いた、ジェノタイプデータの解析結果

A 全検体を学習に用いた場合



B GWAS 集団のみを学習に用いた場合



- 縦軸は、健常対照群中心からの Mahalanobis 距離
- 横軸は、緑内障群中心からの Mahalanobis 距離
- 図中の各 Plot の記号が意味する内容は以下の通り

A	GWAS 緑内障群	C	GWAS 健常対照群
a	Custom 緑内障群	n	Custom 健常対照群

ノタイプデータの様な複雑かつ多次元なデータを、カーネル関数による数値変換を行う事で、単純かつ最も効果的な判別が可能な数値空間に射影する方法である。カーネル関数については、あらゆるデータ特性に対して各々最適に合致する様に、様々な種類が現在も世界中で考案されている。また、カーネル関数で射影された後の値に対する判別に対しては、2群を最も良く分離できる判別境界面を、その付近のデータから十分な距離（マージン）を取る形で作成する様に学習を行う。もし十分なマージンが無ければ、未知のデータに対する評価が、わずかな差異を過大に考慮する様になってしまうために、結果的に誤判定に繋がる事になってしまう。

SVM では、予測精度を向上させるために、各検体に対して前述のマージンを考慮した判別境界面からの距離を用いたスコアを算出する。本プロジェクトの様な2群の判別の場合、片方の群が判別境界面から+1、もう片方が-1になる様に、スコアは算出される。この様子を描いたものが、図.6である。これは、実際にジェノタイプデータを

用いた SVM 解析の結果、出力されたスコアをプロットしたものである。この図より、学習に用いた GWAS 集団では、SVM のスコアが Case では-1、Control では+1 付近に分布する様に学習されている様子が解かる。そして、この学習した SVM を用いた予測診断結果は、概ね Case が0より下、Control が0より上に来る様になっている。従って、各検体のスコアの正負を以って判別すれば、効果的な診断が可能と考えられる。

実際に診断した結果を表.4にまとめた。前述のカーネル関数に関しては、一般的に最も効果的と考えられている「動径基底関数（Radial Basis Function、RBF）」と、多項式を想定した「Polydot」の2つを試してみた。

何れのカーネル関数を用いた場合でも、表.3でまとめた LDA 結果よりも高い予測診断精度を発揮している事が解る。従って、SVM はジェノタイプデータを用いた症例診断に対して、有用な解析手法の1つであると考えられる。

また、少し異なる非線形判別として「決定木（Decision Tree、DT）」とい

図.6 ジェノタイプデータに対する SVM のスコア分布を用いた概念図

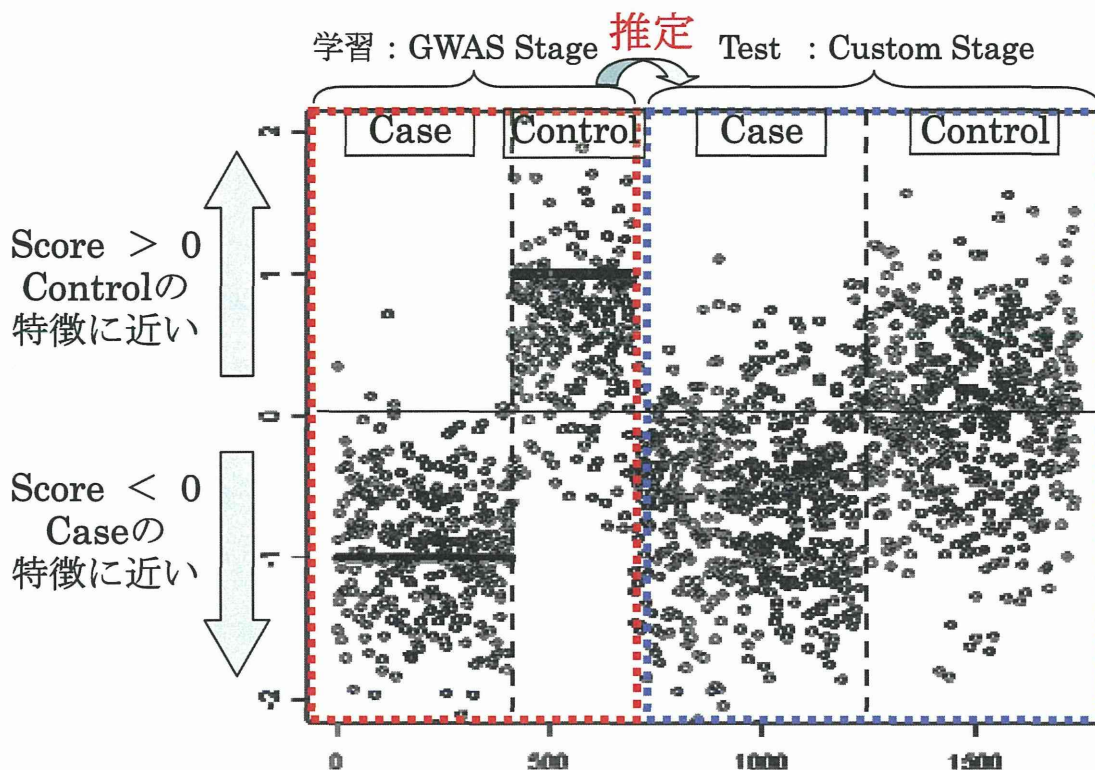


表.4 SVM の Kernel 関数の違いによる解析結果

Stage	集団	All Samples Learning				1st Stage Samples Learning				
		+	-	Sensitivity / Specificity	Accuracy	+	-	Sensitivity / Specificity	Accuracy	
RBF	GWAS	Case	381	30	92.70%	91.57%	403	8	98.05%	94.29%
		Control	29	260	89.97%		32	257	88.93%	
	Custom	Case	472	49	90.60%	86.35%	445	76	85.41%	69.90%
		Control	93	426	82.08%		237	282	54.34%	
Polydot	All	Case	853	79	91.52%	88.45%	848	84	90.99%	79.71%
		Control	122	686	84.90%		269	539	66.71%	
	GWAS	Case	358	53	87.10%	86.86%	395	16	96.11%	93.71%
		Control	39	250	86.51%		28	261	90.31%	
	Custom	Case	420	101	80.61%	77.21%	407	114	78.12%	66.35%
		Control	136	383	73.80%		236	283	54.53%	
	All	Case	778	154	83.48%	81.09%	802	130	86.05%	77.36%
		Control	175	633	78.34%		264	544	67.33%	

う手法も試した。DT は、与えられたデータに対して、各説明変数の値を何らかの基準により多段的に分岐させてゆき、最終的に分類を実現させる解析手法である。この時、複数の分岐を持った全体像が、まるで樹木の様な形をしている事からこの様な名前が付いている。理論的には、全ての説明変数について細かく分岐して行けば、非常に精度の高い分類が可能になるが、一方で少しの例外や分岐順序の違いにより、未知データに対する予測能力が低下する。従って DT には、何らかの評価係数、もしくはエントロピー等の算出を行い、出来る限り少ない分岐数で、より単純な木構造を構築しようとする学習機能が備わっている。

本プロジェクトに照らし合わせると、ジェノタイプデータを各 SNP の値により分岐させ、緑内障と健常者の予測診断を可能とする木構造を得る事が目的となる。また、DT では順序関係が明確になるため、SNP の関連する遺伝子に対し、生理活性の経路に関する知見を得られる可能性もある。

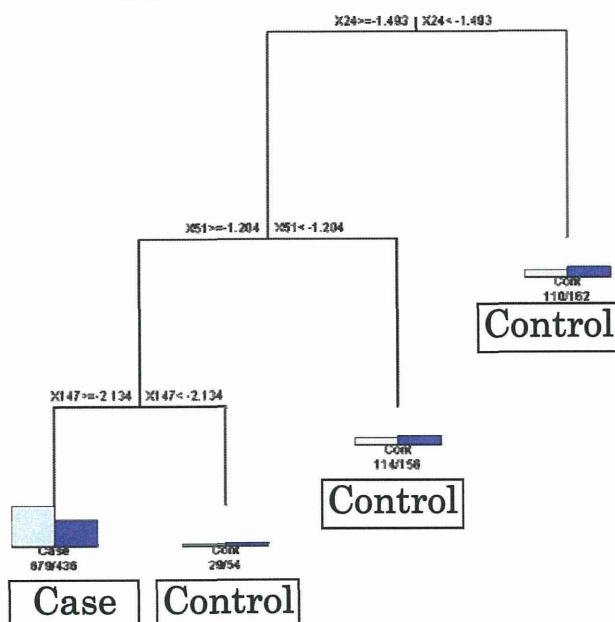
DT を用いて実際に診断した結果を、図.7 にまとめた。これまでの例と同様

に、全検体を用いて学習した結果 (A) と、GWAS 検体のみを用いて学習した結果 (B) の 2 種類を作成した。DT では、自動的に分岐のパターンが設定されるが、今回はいずれの DT でも、分岐点で右側の条件に該当する場合は Control、左側は Case という形で多段的に分類される木構造になった。これらの学習結果では、元データの 165 SNPs 中、DT で使用されるものが 2~3 SNPs と非常に少なくなっている。しかし、いずれも Case を陽性と判定する Sensitivity は高いものの、一方で Specificity は低くなる傾向があり、他の SNPs に比べて判別に対する寄与が大きい SNPs が少数あるものの、それら各々のみでは分類が不十分である事を示唆している。そのため、木構造を単純化すると、全体としても分類能力が低くなり、予測精度が悪くなると考えられる。

DT の結果を、165 SNPs 全てを使用した SVM 等の結果と合わせて考えてみると、疾患に関連する SNPs はもう少し多い可能性が考えられる。また DT は欠点として、一度分岐に使用した SNP は再度使用出来ない事、複数

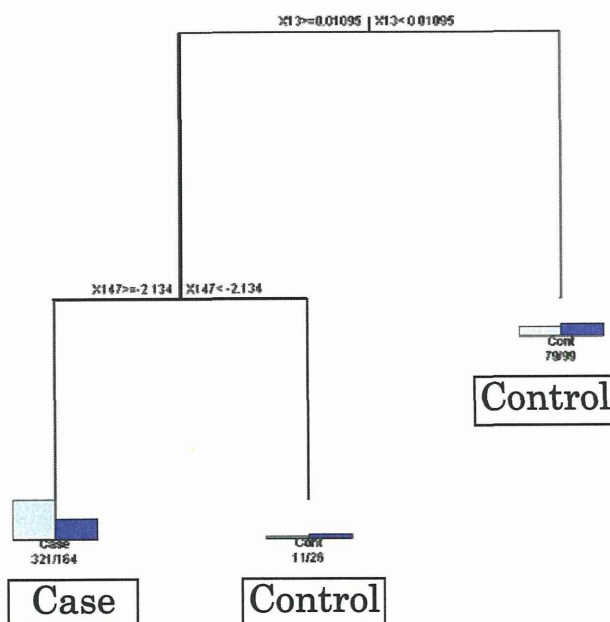
図.7 DTを用いたジェノタイプデータの解析結果

A 全検体を学習に用いた場合



Stage	集団	+	-	Sensitivity / Specificity
GWAS	Case	308	103	74.94%
	Control	147	142	49.13%
		Accuracy		64.29%
Custom	Case	371	150	71.21%
	Control	289	230	44.32%
		Accuracy		57.79%
All	Case	679	253	72.85%
	Control	436	372	46.04%
		Accuracy		60.40%

B GWAS集団のみを学習に用いた場合



Stage	集団	+	-	Sensitivity / Specificity
GWAS	Case	321	90	78.10%
	Control	164	125	43.25%
		Accuracy		63.71%
Custom	Case	371	150	71.21%
	Control	337	182	35.07%
		Accuracy		53.17%
All	Case	692	240	74.25%
	Control	501	307	38.00%
		Accuracy		57.41%

の経路が考えられるパターンの分類には弱い事が挙げられる。従って、実際の疾患には図. 7 に示す木構造が複数絡んでいる可能性もある。DT の使用は有用ではないものの、木構造を用いた解析手法全体が否定される訳ではなく、何らかの改良があれば有用になる可能性は十分にある。

最後に、「単純ベイズ分類器 (Naive Bayes Classifier、NBC)」を試した。これはこれまでの解析手法と異なり、確率論を用いた分類手法である。すなわち、手元にあるデータから各パターンの出現頻度を割り出す事で事前確率を算出し、ベイズ理論を用いる事で事後確率を算出し、各パターンの起こり得る確率を求める事で、分類を行うという手法である。本プロジェクトのデータで言えば、SNPs 毎のジェノタイプ頻度より、緑内障群と健常対照群とで異なる頻度があるので、診断対象のデータの持つジェノタイプデータが、最終的にどちらの傾向に近いかを確率によって算出する。実際に NBC で解析した結果を表. 5 にまとめた。

表. 5 NBC によるジェノタイプデータの解析結果

A 全検体を学習に用いた場合

Stage	集団	+	-	Sensitivity / Specificity
GWAS	Case	337	74	82.00%
	Control	38	251	86.85%
		Accuracy		84.00%
Custom	Case	379	142	72.74%
	Control	144	375	72.25%
		Accuracy		72.50%
All	Case	716	216	76.82%
	Control	182	626	77.48%
		Accuracy		77.13%

B GWAS 集団のみを学習に用いた場合

Stage	集団	+	-	Sensitivity / Specificity
GWAS	Case	361	50	87.83%
	Control	48	241	83.39%
		Accuracy		86.00%
Custom	Case	399	122	76.58%
	Control	211	308	59.34%
		Accuracy		67.98%
All	Case	760	172	81.55%
	Control	259	549	67.95%
		Accuracy		75.23%

GWAS 集団のみの学習から Custom 集団の予測診断を行った結果では、SVM に比べて、学習に用いた GWAS 集団自体の診断率は劣るものの、一方で Custom 集団への予測精度では上回っている。この結果は、LDA や SVM とは異なる傾向を有しつつも、

NBC もジェノタイプデータの予測診断には有用な手法の1つである事を示している。

さて、SVMの説明で用いた図.6を見ると、Caseの予測はほぼ同じ-1のスコア付近で分布しているのに対して、Controlの予測では+1ではなく0付近で分布してしまっている。これは、SVMの結果でSensitivityに比べてSpecificityが低い事(表.4を参照)を、別の角度で見たものになっており、SVMに限った事ではなく、LDA、DT、NBCでも同様の傾向がある。

原因の1つとしては、GWAS集団でCaseが411検体であるのに対して、Controlが289検体と両集団間の検体数に差がある事が考えられる。すなわち、Caseの方が多いために、その特徴を必要以上に強く捉える方向に学習が進んだ結果、健常対照群を陽性と誤判定する傾向が強まり、Specificityが下がったという可能性である。他の原因としては、使用しているSNPs自体に、Case寄りの判定を導きやすいものが含まれている可能性などが挙げられるが、いずれにせよ、何らかの方法でそうしたバイアスを取り除く

必要があると考えられる。

そこで、この様な好ましくない傾向に対処するため、CaseとControl両集団から同数ずつ検体を抽出し、両集団の数としての重みが等しくなる様にした解析を試す事にした。ただし、常に同じ検体を用いるとバイアスが生じるので、元の検体群から同数をランダム抽出した検体群を用いる事にする。

また、この様にランダム抽出した検体を用いて何らかの解析を繰り返し、各解析結果を最終的に統合する方法を「集団学習」と呼ぶ。通常は、ここまで述べたLDAやSVMを基礎的な解析として用い、それらの結果を統合して最終結果を出す事から、この様な方法をメタアナリシス手法という。

実際に、「バギング (Bootstrap Aggregating, Bagging)」と呼ばれる手法を試した。Baggingは、ブートストラップ法を用いたランダム抽出後の各解析結果を一つずつ記憶し、最終的にそれらの結果を多数決する事で、解の統合を図っている。例えば、ある検体に対して、101回のランダム抽出集団による基礎的解析手法を用いた

学習とテストを繰り返した時、陽性判定が 70 回、陰性判定が 31 回ならば、陽性として最終結果を出している。またこの時、ランダム抽出回数とその後の判定回数が偶数の場合、半分に割れて判定不可能になる事があるため、必ず奇数になる様に回数を設定する。例えば、100 回ランダム抽出する場合、1 回増やして 101 回等に設定する。ここでは、基礎的な解析手法として、LDA と SVM を用いている。また SVM では、カーネル関数として「RBF」「Polydot」「anovadot」「vanilladot」「tanhdot」「laplacedot」「besseldot」の 7 つを試してみた。

Bagging を実際に行った結果を、表. 6 に示す。表. 6 の見方は、まず横方向に、学習用の GWAS 集団である Case 411 と Control 289 の 2 群より、100、150、200、250 検体ずつを非復元抽出で選んだ検体を用いた事を示す。また縦方向には、11、25、51、75、101 回繰り返してランダム抽出したという事を示す。例えば、「LDA」「150 vs 150」「51」の部分は、Case と Control から 150 検体ずつ抽出したデータで LDA を学習し、全検体を用いたテス

トを 51 回繰り返した後、それらの結果を合算したら全体で 79.4 % の予測診断率があった事を示している。

表. 6 の結果より、LDA や RBF を用いた SVM では、今までの手法では精度の低かった Specificity が 6 割台後半まで回復しており、Custom 集団に対する予測診断精度も 7 割台に回復している。この結果より、このような手法を用いる事で、ジェノタイプデータによる予測診断は、7 割台の推定ができる可能性がある事が解った。

この様に Bagging は有効であると考えられる一方、ランダム抽出時に何らかのバイアスが存在すれば、解析結果の信頼性が下がってしまう。そこで、もう一つ別のメタアナリシス法である「AdaBoost」を試した。

Bagging では、元の集団から多数繰返しランダム抽出を行う事で、Case と Control 間の検体数の違い、及び各ランダム抽出集団間の違いを吸収していた。これに対して AdaBoost は、ブートストラップ法を用いてランダム抽出される側の元集団に対して補正を加える事で、Case と Control 間の差を是正する手法である。

表.6 Bagging を用いたジェノタイプデータの解析結果

		100 vs 100				150 vs 150				200 vs 200				250 vs 250			
		+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc
LDA																	
11	G Case	358	53	87.1%		371	40	90.3%		366	45	89.1%		365	46	88.8%	
	G Cont	18	271	93.8%	89.9%	28	261	90.3%	90.3%	24	265	91.7%	90.1%	29	260	90.0%	89.3%
	C Case	333	188	63.9%		388	133	74.5%		395	126	75.8%		388	133	74.5%	
	C Cont	224	295	56.8%	60.4%	195	324	62.4%	68.5%	185	334	64.4%	70.1%	182	337	64.9%	69.7%
	All Case	691	241	74.1%		759	173	81.4%		761	171	81.7%		753	179	80.8%	
	All Cont	242	566	70.0%	72.2%	223	585	72.4%	77.2%	209	599	74.1%	78.2%	211	597	73.9%	77.6%
25	G Case	365	46	88.8%		377	34	91.7%		376	35	91.5%		369	42	89.8%	
	G Cont	14	275	95.2%	91.4%	19	270	93.4%	92.4%	22	267	92.4%	91.9%	23	266	92.0%	90.7%
	C Case	371	150	71.2%		394	127	75.6%		380	141	72.9%		395	126	75.8%	
	C Cont	202	317	61.1%	66.2%	193	326	62.8%	69.2%	182	337	64.9%	68.9%	182	337	64.9%	70.4%
	All Case	736	196	79.0%		771	161	82.7%		756	176	81.1%		764	168	82.0%	
	All Cont	216	592	73.3%	76.3%	212	596	73.8%	78.6%	204	604	74.8%	78.2%	205	603	74.6%	78.6%
51	G Case	390	21	94.9%		376	35	91.5%		377	34	91.7%		369	42	89.8%	
	G Cont	8	281	97.2%	95.9%	17	272	94.1%	92.6%	22	267	92.4%	92.0%	27	262	90.7%	90.1%
	C Case	384	137	73.7%		394	127	75.6%		396	125	76.0%		398	123	76.4%	
	C Cont	184	335	64.5%	69.1%	180	339	65.3%	70.5%	183	336	64.7%	70.4%	179	340	65.5%	71.0%
	All Case	774	158	83.0%		770	162	82.6%		773	159	82.9%		767	165	82.3%	
	All Cont	192	616	76.2%	79.9%	197	611	75.6%	79.4%	205	603	74.6%	79.1%	206	602	74.5%	78.7%
75	G Case	382	29	92.9%		380	31	92.5%		372	39	90.5%		365	46	88.8%	
	G Cont	5	284	98.3%	95.1%	15	274	94.8%	93.4%	19	270	93.4%	91.7%	26	263	91.0%	89.7%
	C Case	388	133	74.5%		387	134	74.3%		387	134	74.3%		395	126	75.8%	
	C Cont	178	341	65.7%	70.1%	179	340	65.5%	69.9%	183	336	64.7%	69.5%	178	341	65.7%	70.8%
	All Case	770	162	82.6%		767	165	82.3%		759	173	81.4%		760	172	81.5%	
	All Cont	183	625	77.4%	80.2%	194	614	76.0%	79.4%	202	606	75.0%	78.4%	204	604	74.8%	78.4%
101	G Case	391	20	95.1%		378	33	92.0%		376	35	91.5%		369	42	89.8%	
	G Cont	8	281	97.2%	96.0%	13	276	95.5%	93.4%	22	267	92.4%	91.9%	27	262	90.7%	90.1%
	C Case	389	132	74.7%		385	136	73.9%		390	131	74.9%		398	123	76.4%	
	C Cont	195	324	62.4%	68.6%	187	332	64.0%	68.9%	178	341	65.7%	70.3%	184	335	64.5%	70.5%
	All Case	780	152	83.7%		763	169	81.9%		766	166	82.2%		767	165	82.3%	
	All Cont	203	605	74.9%	79.6%	200	608	75.2%	78.8%	200	608	75.2%	79.0%	211	597	73.9%	78.4%
SVM(rbfdot)																	
11	G Case	361	50	87.8%		364	47	88.6%		376	35	91.5%		382	29	92.9%	
	G Cont	32	257	88.9%	88.3%	22	267	92.4%	90.1%	23	266	92.0%	91.7%	19	270	93.4%	93.1%
	C Case	386	135	74.1%		397	124	76.2%		388	133	74.5%		401	120	77.0%	
	C Cont	185	334	64.4%	69.2%	184	335	64.5%	70.4%	179	340	65.5%	70.0%	175	344	66.3%	71.6%
	All Case	747	185	80.2%		761	171	81.7%		764	168	82.0%		783	149	84.0%	
	All Cont	217	591	73.1%	76.9%	206	602	74.5%	78.3%	202	606	75.0%	78.7%	194	614	76.0%	80.3%
25	G Case	355	56	86.4%		372	39	90.5%		381	30	92.7%		386	25	93.9%	
	G Cont	30	259	89.6%	87.7%	24	265	91.7%	91.0%	20	269	93.1%	92.9%	17	272	94.1%	94.0%
	C Case	390	131	74.9%		405	116	77.7%		398	123	76.4%		400	121	76.8%	
	C Cont	172	347	66.9%	70.9%	172	347	66.9%	72.3%	176	343	66.1%	71.3%	171	348	67.1%	71.9%
	All Case	745	187	79.9%		777	155	83.4%		779	153	83.6%		786	146	84.3%	
	All Cont	202	606	75.0%	77.6%	196	612	75.7%	79.8%	196	612	75.7%	79.9%	188	620	76.7%	80.8%
51	G Case	364	47	88.6%		369	42	89.8%		382	29	92.9%		388	23	94.4%	
	G Cont	29	260	90.0%	89.1%	27	262	90.7%	90.1%	23	266	92.0%	92.6%	19	270	93.4%	94.0%
	C Case	395	126	75.8%		401	120	77.0%		399	122	76.6%		397	124	76.2%	
	C Cont	172	347	66.9%	71.3%	174	345	66.5%	71.7%	176	343	66.1%	71.3%	176	343	66.1%	71.2%
	All Case	759	173	81.4%		770	162	82.6%		781	151	83.8%		785	147	84.2%	
	All Cont	201	607	75.1%	78.5%	201	607	75.1%	79.1%	199	609	75.4%	79.9%	195	613	75.9%	80.3%
75	G Case	364	47	88.6%		372	39	90.5%		382	29	92.9%		388	23	94.4%	
	G Cont	30	259	89.6%	89.0%	26	263	91.0%	90.7%	19	270	93.4%	93.1%	16	273	94.5%	94.4%
	C Case	403	118	77.4%		401	120	77.0%		400	121	76.8%		398	123	76.4%	
	C Cont	172	347	66.9%	72.1%	176	343	66.1%	71.5%	172	347	66.9%	71.8%	172	347	66.9%	71.6%
	All Case	767	165	82.3%		773	159	82.9%		782	150	83.9%		786	146	84.3%	
	All Cont	202	606	75.0%	78.9%	202	606	75.0%	79.3%	191	617	76.4%	80.4%	188	620	76.7%	80.8%
101	G Case	364	47	88.6%		369	42	89.8%		377	34	91.7%		390	21	94.9%	
	G Cont	30	259	89.6%	89.0%	26	263	91.0%	90.3%	20	269	93.1%	92.3%	17	272	94.1%	94.6%
	C Case	404	117	77.5%		400	121	76.8%		401	120	77.0%		397	124	76.2%	
	C Cont	179	340	65.5%	71.5%	172	347	66.9%	71.8%	177	342	65.9%	71.4%	171	348	67.1%	71.6%
	All Case	768	164	82.4%		769	163	82.5%		778	154	83.5%		787	145	84.4%	
	All Cont	209	599	74.1%	78.6%	198	610	75.5%	79.3%	197	611	75.6%	79.8%	188	620	76.7%	80.9%

		100 vs 100				150 vs 150				200 vs 200				250 vs 250			
		+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc
SVM(polydot)																	
11	G Case	368	43	89.5%		373	38	90.8%		392	19	95.4%		396	15	96.4%	
	G Cont	22	267	92.4%	90.7%	9	280	96.9%	93.3%	1	288	99.7%	97.1%	4	285	98.6%	97.3%
	C Case	377	144	72.4%		392	129	75.2%		376	145	72.2%		382	139	73.3%	
	C Cont	185	334	64.4%	68.4%	192	327	63.0%	69.1%	197	322	62.0%	67.1%	198	321	61.8%	67.6%
	All Case	745	187	79.9%		765	167	82.1%		768	164	82.4%		778	154	83.5%	
	All Cont	207	601	74.4%	77.4%	201	607	75.1%	78.9%	198	610	75.5%	79.2%	202	606	75.0%	79.5%
25	G Case	358	53	87.1%		390	21	94.9%		394	17	95.9%		402	9	97.8%	
	G Cont	19	270	93.4%	89.7%	6	283	97.9%	96.1%	0	289	100.0%	97.6%	4	285	98.6%	98.1%
	C Case	387	134	74.3%		392	129	75.2%		385	136	73.9%		382	139	73.3%	
	C Cont	186	333	64.2%	69.2%	170	349	67.2%	71.3%	175	344	66.3%	70.1%	196	323	62.2%	67.8%
	All Case	745	187	79.9%		782	150	83.9%		779	153	83.6%		784	148	84.1%	
	All Cont	205	603	74.6%	77.5%	176	632	78.2%	81.3%	175	633	78.3%	81.1%	200	608	75.2%	80.0%
51	G Case	375	36	91.2%		384	27	93.4%		404	7	98.3%		404	7	98.3%	
	G Cont	21	268	92.7%	91.9%	5	284	98.3%	95.4%	0	289	100.0%	99.0%	5	284	98.3%	98.3%
	C Case	398	123	76.4%		393	128	75.4%		384	137	73.7%		388	133	74.5%	
	C Cont	177	342	65.9%	71.2%	195	324	62.4%	68.9%	188	331	63.8%	68.8%	195	324	62.4%	68.5%
	All Case	773	159	82.9%		777	155	83.4%		788	144	84.5%		792	140	85.0%	
	All Cont	198	610	75.5%	79.5%	200	608	75.2%	79.6%	188	620	76.7%	80.9%	200	608	75.2%	80.5%
75	G Case	367	44	89.3%		391	20	95.1%		402	9	97.8%		402	9	97.8%	
	G Cont	16	273	94.5%	91.4%	1	288	99.7%	97.0%	0	289	100.0%	98.7%	5	284	98.3%	98.0%
	C Case	391	130	75.0%		392	129	75.2%		389	132	74.7%		385	136	73.9%	
	C Cont	174	345	66.5%	70.8%	178	341	65.7%	70.5%	176	343	66.1%	70.4%	191	328	63.2%	68.6%
	All Case	758	174	81.3%		783	149	84.0%		791	141	84.9%		787	145	84.4%	
	All Cont	190	618	76.5%	79.1%	179	629	77.8%	81.1%	176	632	78.2%	81.8%	196	612	75.7%	80.4%
101	G Case	375	36	91.2%		391	20	95.1%		405	6	98.5%		404	7	98.3%	
	G Cont	15	274	94.8%	92.7%	3	286	99.0%	96.7%	0	289	100.0%	99.1%	4	285	98.6%	98.4%
	C Case	393	128	75.4%		389	132	74.7%		391	130	75.0%		384	137	73.7%	
	C Cont	172	347	66.9%	71.2%	177	342	65.9%	70.3%	184	335	64.5%	69.8%	189	330	63.6%	68.7%
	All Case	768	164	82.4%		780	152	83.7%		796	136	85.4%		788	144	84.5%	
	All Cont	187	621	76.9%	79.8%	180	628	77.7%	80.9%	184	624	77.2%	81.6%	193	615	76.1%	80.6%
SVM(anovadot)																	
11	G Case	359	52	87.3%		367	44	89.3%		386	25	93.9%		401	10	97.6%	
	G Cont	29	260	90.0%	88.4%	9	280	96.9%	92.4%	2	287	99.3%	96.1%	0	289	100.0%	98.6%
	C Case	354	167	67.9%		385	136	73.9%		376	145	72.2%		366	155	70.2%	
	C Cont	197	322	62.0%	65.0%	191	328	63.2%	68.6%	191	328	63.2%	67.7%	196	323	62.2%	66.3%
	All Case	713	219	76.5%		752	180	80.7%		762	170	81.8%		767	165	82.3%	
	All Cont	226	582	72.0%	74.4%	200	608	75.2%	78.2%	193	615	76.1%	79.1%	196	612	75.7%	79.3%
25	G Case	366	45	89.1%		382	29	92.9%		392	19	95.4%		407	4	99.0%	
	G Cont	15	274	94.8%	91.4%	6	283	97.9%	95.0%	0	289	100.0%	97.3%	0	289	100.0%	99.4%
	C Case	374	147	71.8%		368	153	70.6%		373	148	71.6%		371	150	71.2%	
	C Cont	185	334	64.4%	68.1%	179	340	65.5%	68.1%	193	326	62.8%	67.2%	199	320	61.7%	66.4%
	All Case	740	192	79.4%		750	182	80.5%		765	167	82.1%		778	154	83.5%	
	All Cont	200	608	75.2%	77.5%	185	623	77.1%	78.9%	193	615	76.1%	79.3%	199	609	75.4%	79.7%
51	G Case	367	44	89.3%		389	22	94.6%		402	9	97.8%		410	1	99.8%	
	G Cont	18	271	93.8%	91.1%	6	283	97.9%	96.0%	0	289	100.0%	98.7%	0	289	100.0%	99.9%
	C Case	369	152	70.8%		375	146	72.0%		375	146	72.0%		369	152	70.8%	
	C Cont	184	335	64.5%	67.7%	192	327	63.0%	67.5%	200	319	61.5%	66.7%	181	338	65.1%	68.0%
	All Case	736	196	79.0%		764	168	82.0%		777	155	83.4%		779	153	83.6%	
	All Cont	202	606	75.0%	77.1%	198	610	75.5%	79.0%	200	608	75.2%	79.6%	181	627	77.6%	80.8%
75	G Case	371	40	90.3%		388	23	94.4%		404	7	98.3%		411	0	100.0%	
	G Cont	14	275	95.2%	92.3%	1	288	99.7%	96.6%	0	289	100.0%	99.0%	0	289	100.0%	100.0%
	C Case	381	140	73.1%		380	141	72.9%		362	159	69.5%		375	146	72.0%	
	C Cont	190	329	63.4%	68.3%	185	334	64.4%	68.7%	181	338	65.1%	67.3%	193	326	62.8%	67.4%
	All Case	752	180	80.7%		768	164	82.4%		766	166	82.2%		786	146	84.3%	
	All Cont	204	604	74.8%	77.9%	186	622	77.0%	79.9%	181	627	77.6%	80.1%	193	615	76.1%	80.5%
101	G Case	371	40	90.3%		388	23	94.4%		404	7	98.3%		409	2	99.5%	
	G Cont	15	274	94.8%	92.1%	1	288	99.7%	96.6%	0	289	100.0%	99.0%	0	289	100.0%	99.7%
	C Case	374	147	71.8%		370	151	71.0%		368	153	70.6%		372	149	71.4%	
	C Cont	180	339	65.3%	68.6%	187	332	64.0%	67.5%	192	327	63.0%	66.8%	194	325	62.6%	67.0%
	All Case	745	187	79.9%		758	174	81.3%		772	160	82.8%		781	151	83.8%	
	All Cont	195	613	75.9%	78.0%	188	620	76.7%	79.2%	192	616	76.2%	79.8%	194	614	76.0%	80.2%

		100 vs 100				150 vs 150				200 vs 200				250 vs 250			
		+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc
SVM(vanilladot)																	
11	G Case	368	43	89.5%		373	38	90.8%		392	19	95.4%		396	15	96.4%	
	G Cont	22	267	92.4%	90.7%	9	280	96.9%	93.3%	1	288	99.7%	97.1%	4	285	98.6%	97.3%
	C Case	377	144	72.4%		392	129	75.2%		376	145	72.2%		382	139	73.3%	
	C Cont	185	334	64.4%	68.4%	192	327	63.0%	69.1%	197	322	62.0%	67.1%	198	321	61.8%	67.6%
	All Case	745	187	79.9%		765	167	82.1%		768	164	82.4%		778	154	83.5%	
	All Cont	207	601	74.4%	77.4%	201	607	75.1%	78.9%	198	610	75.5%	79.2%	202	606	75.0%	79.5%
25	G Case	358	53	87.1%		390	21	94.9%		394	17	95.9%		402	9	97.8%	
	G Cont	19	270	93.4%	89.7%	6	283	97.9%	96.1%	0	289	100.0%	97.6%	4	285	98.6%	98.1%
	C Case	387	134	74.3%		392	129	75.2%		385	136	73.9%		382	139	73.3%	
	C Cont	186	333	64.2%	69.2%	170	349	67.2%	71.3%	175	344	66.3%	70.1%	196	323	62.2%	67.8%
	All Case	745	187	79.9%		782	150	83.9%		779	153	83.6%		784	148	84.1%	
	All Cont	205	603	74.6%	77.5%	176	632	78.2%	81.3%	175	633	78.3%	81.1%	200	608	75.2%	80.0%
51	G Case	375	36	91.2%		384	27	93.4%		404	7	98.3%		404	7	98.3%	
	G Cont	21	268	92.7%	91.9%	5	284	98.3%	95.4%	0	289	100.0%	99.0%	5	284	98.3%	98.3%
	C Case	398	123	76.4%		393	128	75.4%		384	137	73.7%		388	133	74.5%	
	C Cont	177	342	65.9%	71.2%	195	324	62.4%	68.9%	188	331	63.8%	68.8%	195	324	62.4%	68.5%
	All Case	773	159	82.9%		777	155	83.4%		788	144	84.5%		792	140	85.0%	
	All Cont	198	610	75.5%	79.5%	200	608	75.2%	79.6%	188	620	76.7%	80.9%	200	608	75.2%	80.5%
75	G Case	367	44	89.3%		391	20	95.1%		402	9	97.8%		402	9	97.8%	
	G Cont	16	273	94.5%	91.4%	1	288	99.7%	97.0%	0	289	100.0%	98.7%	5	284	98.3%	98.0%
	C Case	391	130	75.0%		392	129	75.2%		389	132	74.7%		385	136	73.9%	
	C Cont	174	345	66.5%	70.8%	178	341	65.7%	70.5%	176	343	66.1%	70.4%	191	328	63.2%	68.6%
	All Case	758	174	81.3%		783	149	84.0%		791	141	84.9%		787	145	84.4%	
	All Cont	190	618	76.5%	79.1%	179	629	77.8%	81.1%	176	632	78.2%	81.8%	196	612	75.7%	80.4%
101	G Case	375	36	91.2%		391	20	95.1%		405	6	98.5%		404	7	98.3%	
	G Cont	15	274	94.8%	92.7%	3	286	99.0%	96.7%	0	289	100.0%	99.1%	4	285	98.6%	98.4%
	C Case	393	128	75.4%		389	132	74.7%		391	130	75.0%		384	137	73.7%	
	C Cont	172	347	66.9%	71.2%	177	342	65.9%	70.3%	184	335	64.5%	69.8%	189	330	63.6%	68.7%
	All Case	768	164	82.4%		780	152	83.7%		796	136	85.4%		788	144	84.5%	
	All Cont	187	621	76.9%	79.8%	180	628	77.7%	80.9%	184	624	77.2%	81.6%	193	615	76.1%	80.6%
SVM(tanhdot)																	
11	G Case	282	129	68.6%		280	131	68.1%		271	140	65.9%		258	153	62.8%	
	G Cont	85	204	70.6%	69.4%	107	182	63.0%	66.0%	107	182	63.0%	64.7%	116	173	59.9%	61.6%
	C Case	368	153	70.6%		372	149	71.4%		364	157	69.9%		379	142	72.7%	
	C Cont	193	326	62.8%	66.7%	201	318	61.3%	66.3%	191	328	63.2%	66.5%	197	322	62.0%	67.4%
	All Case	650	282	69.7%		652	280	70.0%		635	297	68.1%		637	295	68.3%	
	All Cont	278	530	65.6%	67.8%	308	500	61.9%	66.2%	298	510	63.1%	65.8%	313	495	61.3%	65.1%
25	G Case	294	117	71.5%		274	137	66.7%		268	143	65.2%		253	158	61.6%	
	G Cont	96	193	66.8%	69.6%	99	190	65.7%	66.3%	108	181	62.6%	64.1%	117	172	59.5%	60.7%
	C Case	381	140	73.1%		380	141	72.9%		379	142	72.7%		378	143	72.6%	
	C Cont	182	337	64.9%	69.0%	184	335	64.5%	68.8%	196	323	62.2%	67.5%	193	326	62.8%	67.7%
	All Case	675	257	72.4%		654	278	70.2%		647	285	69.4%		631	301	67.7%	
	All Cont	278	530	65.6%	69.3%	283	525	65.0%	67.8%	304	504	62.4%	66.1%	310	498	61.6%	64.9%
51	G Case	281	130	68.4%		279	132	67.9%		262	149	63.7%		255	156	62.0%	
	G Cont	92	197	68.2%	68.3%	99	190	65.7%	67.0%	108	181	62.6%	63.3%	118	171	59.2%	60.9%
	C Case	366	155	70.2%		376	145	72.2%		371	150	71.2%		365	156	70.1%	
	C Cont	174	345	66.5%	68.4%	187	332	64.0%	68.1%	185	334	64.4%	67.8%	179	340	65.5%	67.8%
	All Case	647	285	69.4%		655	277	70.3%		633	299	67.9%		620	312	66.5%	
	All Cont	266	542	67.1%	68.3%	286	522	64.6%	67.6%	293	515	63.7%	66.0%	297	511	63.2%	65.0%
75	G Case	290	121	70.6%		281	130	68.4%		265	146	64.5%		258	153	62.8%	
	G Cont	92	197	68.2%	69.6%	98	191	66.1%	67.4%	107	182	63.0%	63.9%	122	167	57.8%	60.7%
	C Case	380	141	72.9%		381	140	73.1%		372	149	71.4%		367	154	70.4%	
	C Cont	184	335	64.5%	68.8%	187	332	64.0%	68.6%	181	338	65.1%	68.3%	187	332	64.0%	67.2%
	All Case	670	262	71.9%		662	270	71.0%		637	295	68.3%		625	307	67.1%	
	All Cont	276	532	65.8%	69.1%	285	523	64.7%	68.1%	288	520	64.4%	66.5%	309	499	61.8%	64.6%
101	G Case	289	122	70.3%		281	130	68.4%		267	144	65.0%		248	163	60.3%	
	G Cont	88	201	69.6%	70.0%	102	187	64.7%	66.9%	112	177	61.2%	63.4%	122	167	57.8%	59.3%
	C Case	380	141	72.9%		377	144	72.4%		374	147	71.8%		366	155	70.2%	
	C Cont	174	345	66.5%	69.7%	178	341	65.7%	69.0%	178	341	65.7%	68.8%	177	342	65.9%	68.1%
	All Case	669	263	71.8%		658	274	70.6%		641	291	68.8%		614	318	65.9%	
	All Cont	262	546	67.6%	69.8%	280	528	65.3%	68.2%	290	518	64.1%	66.6%	299	509	63.0%	64.5%

	100 vs 100				150 vs 150				200 vs 200				250 vs 250				
	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	
SVM(laplacedot)																	
11	G Case	350	61	85.2%		351	60	85.4%		345	66	83.9%		361	50	87.8%	
	G Cont	30	259	89.6%	87.0%	20	269	93.1%	88.6%	22	267	92.4%	87.4%	26	263	91.0%	89.1%
	C Case	368	153	70.6%		358	163	68.7%		339	182	65.1%		379	142	72.7%	
	C Cont	162	357	68.8%	69.7%	136	383	73.8%	71.3%	132	387	74.6%	69.8%	164	355	68.4%	70.6%
	All Case	718	214	77.0%		709	223	76.1%		684	248	73.4%		740	192	79.4%	
	All Cont	192	616	76.2%	76.7%	156	652	80.7%	78.2%	154	654	80.9%	76.9%	190	618	76.5%	78.0%
25	G Case	349	62	84.9%		360	51	87.6%		351	60	85.4%		360	51	87.6%	
	G Cont	29	260	90.0%	87.0%	24	265	91.7%	89.3%	21	268	92.7%	88.4%	23	266	92.0%	89.4%
	C Case	375	146	72.0%		374	147	71.8%		351	170	67.4%		362	159	69.5%	
	C Cont	155	364	70.1%	71.1%	153	366	70.5%	71.2%	135	384	74.0%	70.7%	147	372	71.7%	70.6%
	All Case	724	208	77.7%		734	198	78.8%		702	230	75.3%		722	210	77.5%	
	All Cont	184	624	77.2%	77.5%	177	631	78.1%	78.4%	156	652	80.7%	77.8%	170	638	79.0%	78.2%
51	G Case	355	56	86.4%		358	53	87.1%		368	43	89.5%		362	49	88.1%	
	G Cont	29	260	90.0%	87.9%	25	264	91.3%	88.9%	26	263	91.0%	90.1%	23	266	92.0%	89.7%
	C Case	373	148	71.6%		369	152	70.8%		380	141	72.9%		352	169	67.6%	
	C Cont	155	364	70.1%	70.9%	149	370	71.3%	71.1%	159	360	69.4%	71.2%	140	379	73.0%	70.3%
	All Case	728	204	78.1%		727	205	78.0%		748	184	80.3%		714	218	76.6%	
	All Cont	184	624	77.2%	77.7%	174	634	78.5%	78.2%	185	623	77.1%	78.8%	163	645	79.8%	78.1%
75	G Case	354	57	86.1%		359	52	87.3%		360	51	87.6%		362	49	88.1%	
	G Cont	27	262	90.7%	88.0%	24	265	91.7%	89.1%	24	265	91.7%	89.3%	22	267	92.4%	89.9%
	C Case	367	154	70.4%		368	153	70.6%		365	156	70.1%		364	157	69.9%	
	C Cont	147	372	71.7%	71.1%	148	371	71.5%	71.1%	149	370	71.3%	70.7%	143	376	72.4%	71.2%
	All Case	721	211	77.4%		727	205	78.0%		725	207	77.8%		726	206	77.9%	
	All Cont	174	634	78.5%	77.9%	172	636	78.7%	78.3%	173	635	78.6%	78.2%	165	643	79.6%	78.7%
101	G Case	359	52	87.3%		353	58	85.9%		361	50	87.8%		360	51	87.6%	
	G Cont	27	262	90.7%	88.7%	22	267	92.4%	88.6%	24	265	91.7%	89.4%	21	268	92.7%	89.7%
	C Case	372	149	71.4%		360	161	69.1%		367	154	70.4%		359	162	68.9%	
	C Cont	156	363	69.9%	70.7%	145	374	72.1%	70.6%	146	373	71.9%	71.2%	143	376	72.4%	70.7%
	All Case	731	201	78.4%		713	219	76.5%		728	204	78.1%		719	213	77.1%	
	All Cont	183	625	77.4%	77.9%	167	641	79.3%	77.8%	170	638	79.0%	78.5%	164	644	79.7%	78.3%
SVM(besseldot)																	
11	G Case	297	114	72.3%		306	105	74.5%		357	54	86.9%		381	30	92.7%	
	G Cont	93	196	67.8%	70.4%	33	256	88.6%	80.3%	10	279	96.5%	90.9%	3	286	99.0%	95.3%
	C Case	285	236	54.7%		260	261	49.9%		275	246	52.8%		264	257	50.7%	
	C Cont	313	206	39.7%	47.2%	299	220	42.4%	46.2%	297	222	42.8%	47.8%	304	215	41.4%	46.1%
	All Case	582	350	62.4%		566	366	60.7%		632	300	67.8%		645	287	69.2%	
	All Cont	406	402	49.8%	56.6%	332	476	58.9%	59.9%	307	501	62.0%	65.1%	307	501	62.0%	65.9%
25	G Case	299	112	72.7%		319	92	77.6%		358	53	87.1%		398	13	96.8%	
	G Cont	105	184	63.7%	69.0%	34	255	88.2%	82.0%	3	286	99.0%	92.0%	2	287	99.3%	97.9%
	C Case	279	242	53.6%		267	254	51.2%		273	248	52.4%		259	262	49.7%	
	C Cont	312	207	39.9%	46.7%	308	211	40.7%	46.0%	307	212	40.8%	46.6%	291	228	43.9%	46.8%
	All Case	578	354	62.0%		586	346	62.9%		631	301	67.7%		657	275	70.5%	
	All Cont	417	391	48.4%	55.7%	342	466	57.7%	60.5%	310	498	61.6%	64.9%	293	515	63.7%	67.4%
51	G Case	301	110	73.2%		321	90	78.1%		368	43	89.5%		402	9	97.8%	
	G Cont	89	200	69.2%	71.6%	33	256	88.6%	82.4%	0	289	100.0%	93.9%	2	287	99.3%	98.4%
	C Case	258	263	49.5%		265	256	50.9%		258	263	49.5%		260	261	49.9%	
	C Cont	299	220	42.4%	46.0%	294	225	43.4%	47.1%	297	222	42.8%	46.2%	294	225	43.4%	46.6%
	All Case	559	373	60.0%		586	346	62.9%		626	306	67.2%		662	270	71.0%	
	All Cont	388	420	52.0%	56.3%	327	481	59.5%	61.3%	297	511	63.2%	65.3%	296	512	63.4%	67.5%
75	G Case	307	104	74.7%		331	80	80.5%		375	36	91.2%		409	2	99.5%	
	G Cont	98	191	66.1%	71.1%	26	263	91.0%	84.9%	1	288	99.7%	94.7%	2	287	99.3%	99.4%
	C Case	262	259	50.3%		267	254	51.2%		263	258	50.5%		257	264	49.3%	
	C Cont	298	221	42.6%	46.4%	300	219	42.2%	46.7%	296	223	43.0%	46.7%	293	226	43.5%	46.4%
	All Case	569	363	61.1%		598	334	64.2%		638	294	68.5%		666	266	71.5%	
	All Cont	396	412	51.0%	56.4%	326	482	59.7%	62.1%	297	511	63.2%	66.0%	295	513	63.5%	67.8%
101	G Case	306	105	74.5%		328	83	79.8%		378	33	92.0%		406	5	98.8%	
	G Cont	100	189	65.4%	70.7%	26	263	91.0%	84.4%	0	289	100.0%	95.3%	2	287	99.3%	99.0%
	C Case	266	255	51.1%		263	258	50.5%		266	255	51.1%		256	265	49.1%	
	C Cont	304	215	41.4%	46.3%	301	218	42.0%	46.3%	299	220	42.4%	46.7%	295	224	43.2%	46.2%
	All Case	572	360	61.4%		591	341	63.4%		644	288	69.1%		662	270	71.0%	
	All Cont	404	404	50.0%	56.1%	327	481	59.5%	61.6%	299	509	63.0%	66.3%	297	511	63.2%	67.4%

※ 略称

G : GWAS, C : Custom, Cont : Control, Se/Sp : Sensitivity/Specificity, Acc : Accuracy

具体的には、例えば Case 集団内において、411 検体全てを使ったり、それらから単純にランダム抽出したりする解析手法では、多数の検体に含まれるパターンのみを学習して、少数の判別し難い検体を見落とす可能性があった。一方の AdaBoost では、疾患の特徴が非常に解りやすい検体や、逆に判定に苦慮する検体を重点的に抽出する事で、Case 集団の判別能力を高めようとしている。結果的に、全検体を用いなくても高い分類能力を得る事ができるため、Control との検体数の差も是正する事に繋がる。

Bagging と同様の LDA 及び SVM を用いた 8 種類の基礎解析手法を用いて、AdaBoost を行った結果を表. 7 に示す。Resampling 回数は、11 回、51 回、101 回の 3 パターンを試した。これらの結果より、LDA や RBF を用いた SVM は、学習に用いた GWAS 集団に対する判別は 100%に近いものの、Custom 集団に対する予測診断精度は Bagging ほど高くない事が解った。ただし、別のメタアナリシス手法を用いても、ある程度の結果を出している事から、サンプリングのバイアスはほぼ

無いと考えられる。

以上より、ジェノタイプデータ解析において有用と考えられる手法としては、RBF を用いた SVM が最も良く、次いで LDA や NBC があり、メタアナリシスも有効だと解った。

(3) 使用 SNPs の検討

ここまでの検討では、使用する SNPs は全て基本的な統計値等で絞り込んだ 165 SNPs を用いていた。しかし、実際に解析で有用な SNPs はもっと少ない可能性があり、余分な SNPs は偽陽性の原因になりうる。何故ならば、GWAS 集団で学習し、Custom 集団でテストするという流れでは、GWAS 集団のみで強く特徴が現れている SNPs のパターンを過学習している可能性があり、それが Custom 集団への予測診断精度を悪化させる要因になり得るからである。

しかし一方で、165 SNPs から一定の SNPs の組合せを選ぶにしても、その組合せ数は膨大になり、網羅的に逐一組合せを試すのは難しい。例えば、165 SNPs から 100 SNPs を選ぶ時の組合せ総数は約 7.05×10^{46} 通りにも

表.7 AdaBoost を用いたジェノタイプデータの解析結果

		11 Resamplings				51 Resamplings				101 Resamplings			
		+	-	Se/Sp	Acc	+	-	Se/Sp	Acc	+	-	Se/Sp	Acc
LDA	G Case	405	6	98.5%		411	0	100.0%		411	0	100.0%	
	G Cont	3	286	99.0%	98.7%	0	289	100.0%	100.0%	0	289	100.0%	100.0%
	C Case	371	150	71.2%		374	147	71.8%		390	131	74.9%	
	C Cont	197	322	62.0%	66.6%	223	296	57.0%	64.4%	216	303	58.4%	66.6%
	All Case	776	156	83.3%		785	147	84.2%		801	131	85.9%	
	All Cont	200	608	75.2%	79.5%	223	585	72.4%	78.7%	216	592	73.3%	80.1%
SVM (rbfdot)	G Case	410	1	99.8%		411	0	100.0%		411	0	100.0%	
	G Cont	1	288	99.7%	99.7%	1	288	99.7%	99.9%	0	289	100.0%	100.0%
	C Case	417	104	80.0%		404	117	77.5%		406	115	77.9%	
	C Cont	219	300	57.8%	68.9%	220	299	57.6%	67.6%	229	290	55.9%	66.9%
	All Case	827	105	88.7%		815	117	87.4%		817	115	87.7%	
	All Cont	220	588	72.8%	81.3%	221	587	72.6%	80.6%	229	579	71.7%	80.2%
SVM (polydot)	G Case	411	0	100.0%		410	1	99.8%		411	0	100.0%	
	G Cont	1	288	99.7%	99.9%	0	289	100.0%	99.9%	0	289	100.0%	100.0%
	C Case	414	107	79.5%		412	109	79.1%		413	108	79.3%	
	C Cont	224	295	56.8%	68.2%	229	290	55.9%	67.5%	226	293	56.5%	67.9%
	All Case	825	107	88.5%		822	110	88.2%		824	108	88.4%	
	All Cont	225	583	72.2%	80.9%	229	579	71.7%	80.5%	226	582	72.0%	80.8%
SVM (anovadot)	G Case	408	3	99.3%		411	0	100.0%		410	1	99.8%	
	G Cont	1	288	99.7%	99.4%	0	289	100.0%	100.0%	1	288	99.7%	99.7%
	C Case	395	126	75.8%		386	135	74.1%		390	131	74.9%	
	C Cont	227	292	56.3%	66.1%	224	295	56.8%	65.5%	231	288	55.5%	65.2%
	All Case	803	129	86.2%		797	135	85.5%		800	132	85.8%	
	All Cont	228	580	71.8%	79.5%	224	584	72.3%	79.4%	232	576	71.3%	79.1%
SVM (vanilladot)	G Case	411	0	100.0%		410	1	99.8%		411	0	100.0%	
	G Cont	1	288	99.7%	99.9%	0	289	100.0%	99.9%	0	289	100.0%	100.0%
	C Case	414	107	79.5%		412	109	79.1%		413	108	79.3%	
	C Cont	224	295	56.8%	68.2%	229	290	55.9%	67.5%	226	293	56.5%	67.9%
	All Case	825	107	88.5%		822	110	88.2%		824	108	88.4%	
	All Cont	225	583	72.2%	80.9%	229	579	71.7%	80.5%	226	582	72.0%	80.8%
SVM (tanhdot)	G Case	297	114	72.3%		284	127	69.1%		277	134	67.4%	
	G Cont	137	152	52.6%	64.1%	132	157	54.3%	63.0%	131	158	54.7%	62.1%
	C Case	390	131	74.9%		397	124	76.2%		398	123	76.4%	
	C Cont	253	266	51.3%	63.1%	232	287	55.3%	65.8%	240	279	53.8%	65.1%
	All Case	687	245	73.7%		681	251	73.1%		675	257	72.4%	
	All Cont	390	418	51.7%	63.5%	364	444	55.0%	64.7%	371	437	54.1%	63.9%
SVM (laplacedot)	G Case	411	0	100.0%		411	0	100.0%		411	0	100.0%	
	G Cont	289	0	0.0%	58.7%	289	0	0.0%	58.7%	289	0	0.0%	58.7%
	C Case	521	0	100.0%		521	0	100.0%		521	0	100.0%	
	C Cont	519	0	0.0%	50.1%	519	0	0.0%	50.1%	519	0	0.0%	50.1%
	All Case	932	0	100.0%		932	0	100.0%		932	0	100.0%	
	All Cont	808	0	0.0%	53.6%	808	0	0.0%	53.6%	808	0	0.0%	53.6%
SVM (besseldot)	G Case	406	5	98.8%		411	0	100.0%		411	0	100.0%	
	G Cont	1	288	99.7%	99.1%	0	289	100.0%	100.0%	0	289	100.0%	100.0%
	C Case	331	190	63.5%		287	234	55.1%		296	225	56.8%	
	C Cont	371	148	28.5%	46.1%	356	163	31.4%	43.3%	350	169	32.6%	44.7%
	All Case	737	195	79.1%		698	234	74.9%		707	225	75.9%	
	All Cont	372	436	54.0%	67.4%	356	452	55.9%	66.1%	350	458	56.7%	67.0%

※ 略称

G : GWAS, C : Custom, Cont : Control, Se/Sp : Sensitivity/Specificity, Acc : Accuracy

上り、仮に1分間に10万回計算できるコンピュータを使用しても、36年以上かかる事になる。よって、網羅的な計算ではなく、何らかの手段で組合せ最適化を行う必要がある。

従って、ここでは進化的計算手法の一つである「分布推定アルゴリズム (Estimation of Distribution Algorithm, EDA)」を用いて、SNPsの組合せ最適化を検証する。進化的計算法とは、ダーウィンの進化論にヒントを得た人工知能的アプローチである。

概略を図.8Aに示し、以下EDAの動作を、図中のフローチャートの手順を追って説明する。EDAでは、まずランダムに、使用するSNPが1、使用しないSNPが0と表される合計165個の1と0からなる配列を作成する。この様な配列を「解候補」と呼び、一定数作成して解候補集団とする。これが、フローチャート中の「初期集団の生成」のプロセスに相当する。

次に、各解候補に対して使用するSNPsのみを用いて解析を行い、予測診断率を算出した後、降順に解候補を並べる。この時、上位ほど診断精度が

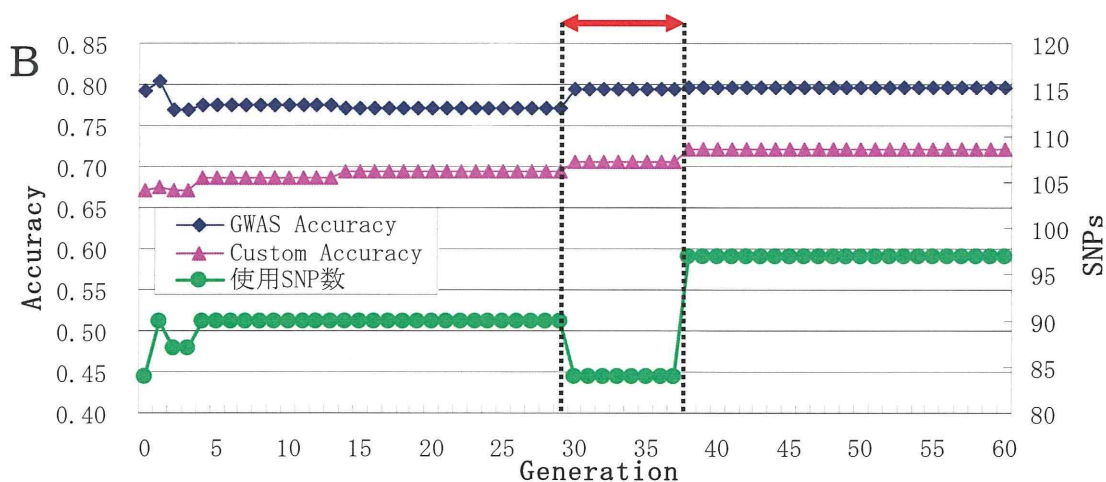
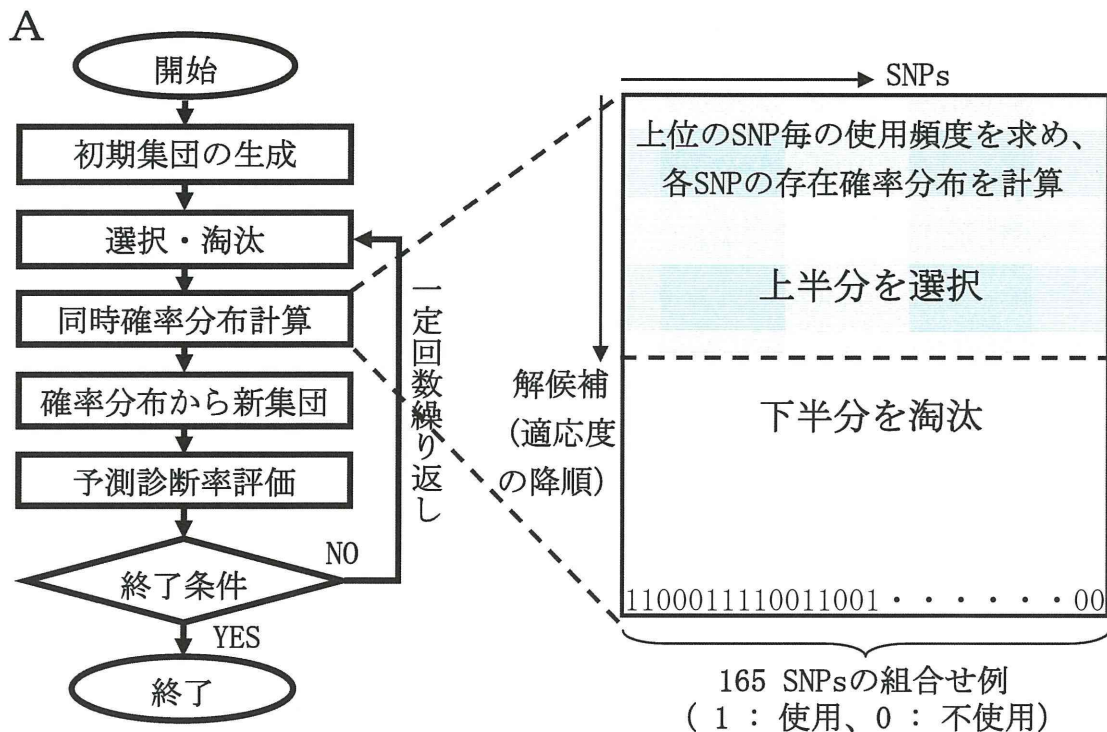
高いSNPsの組合せとなり、下位ほど診断に向かないSNPsとなる。従って、上位のみを残して(選択)、下位を解候補には不向きなものとして消す(淘汰)。これが進化論を模した「選択・淘汰」のプロセスであり、本プロジェクトでは解候補集団を二分して、上位のみを選択、残りを淘汰させている。

選択した解候補集団については、各SNPsで「1」となっている解候補が幾つあるかという頻度を数え上げる。そして選択集団総数の内、何回使用されているかという割合を算出する。この割合は、診断精度が比較的高い解候補中に該当SNPが存在する確率を表す。すなわち、この確率が高いSNP程、診断精度の向上に寄与しているものと考えられる。これが「同時確率分布計算」のプロセスになる。

ここで作られた確率分布を元に、先に淘汰された分の解候補を新規に作成する。これが「確率分布から新集団」のプロセスになる。例えば、あるSNPで0.9という確率が算出された場合、新集団には90%の解候補中にそのSNPが含まれている事になる。

そして再び、それらSNPsの組合せ

図.8 EDA を用いた SNPs 組合せ最適化



C

Stage	集団	+	-	Sensitivity / Specificity	Accuracy
GWAS	Case	327	84	79.56%	79.71%
	Control	58	231	79.93%	
Custom	Case	377	144	72.36%	72.02%
	Control	147	372	71.68%	
All	Case	704	228	75.54%	75.11%
	Control	205	603	74.63%	