マイクロアレイを用いた糖尿病患者の遺伝子データの解析

最終発表

6班　担当教員　佐藤（イリチュ）美佳
201120659　居城秀明
201120626　鈴木昭平
201120639　松原史浩

背景

近年「笑い」と健康との関連性が注目されている.

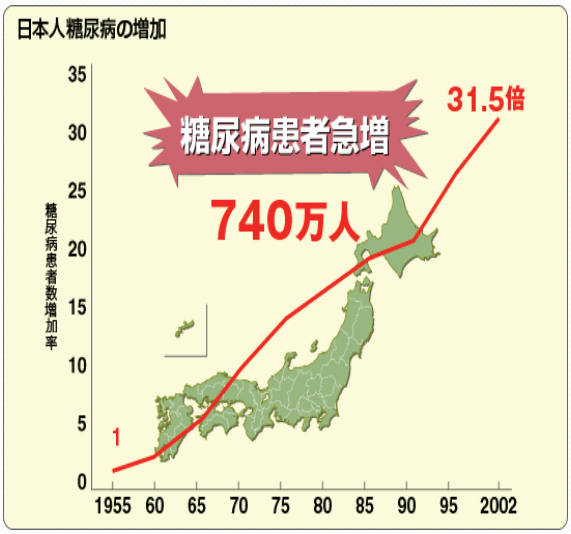
- 「笑い」がストレス解消となる
- 免疫力が高まり，がんが治る
- 痛みを和らげる
- 糖尿病の進展を抑える

糖尿病と「笑い」との関わりについて着目

糖尿病

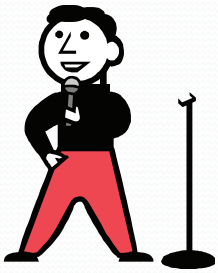
- 血糖値が異常に高い状態となる
- 他の病気との合併症をおこし日常生活に支障をきたす恐れ
- 糖尿病の分類
  - 1型
  - 2型　←大多数
  - 遺伝子発症型
  - その他疾患

- 東日本大震災時にも多数の糖尿病患者に対しインシュリン不足などの問題が生じていた。



糖尿病と遺伝子1

- 糖尿病患者に対し2種類の血糖値変化の実験
  - 食後に講義を聞き，血糖値変化を計測
  - 食後に漫才を鑑賞し，血糖値変化を計測



↓

漫才鑑賞後の血糖値上昇の平均値が大きく低下した！

糖尿病と遺伝子2

- 遺伝子解析の結果，「笑い」によって変化する遺伝子の存在が明らかとなった．
- 「笑い」によって血糖値を下げる作用を起こす遺伝子のスイッチがオンになったことがわかった．
- 「笑い」という生理現象がヒトの体に影響を及ぼす可能性を示唆する結果．

目的

- 糖尿病患者に「笑い」を与えた場合とそうでない場合に検出した遺伝子計測データ解析
  - マイクロアレイによる遺伝子発現データを利用
- 糖尿病患者への「笑い」への影響がどの遺伝子によって識別できるかを調査
  - マーカー遺伝子を同定し、識別関数を推定
- 識別関数の推定から、糖尿病患者の「笑い」による治療・効果を知るにあたり必要とされる遺伝子のデータの特定、効果の有無を判定

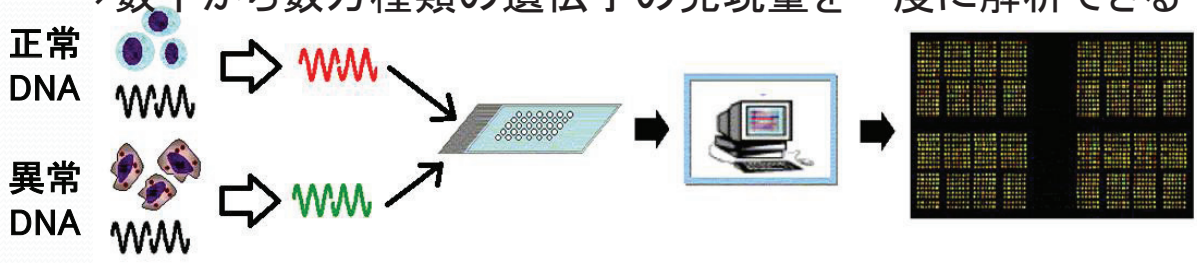
関連研究

著書：KNOWLEDGE DISCOVERY IN BIOINFORMATICS
出版社：WILLEY-INTERSCIENCE, 2007

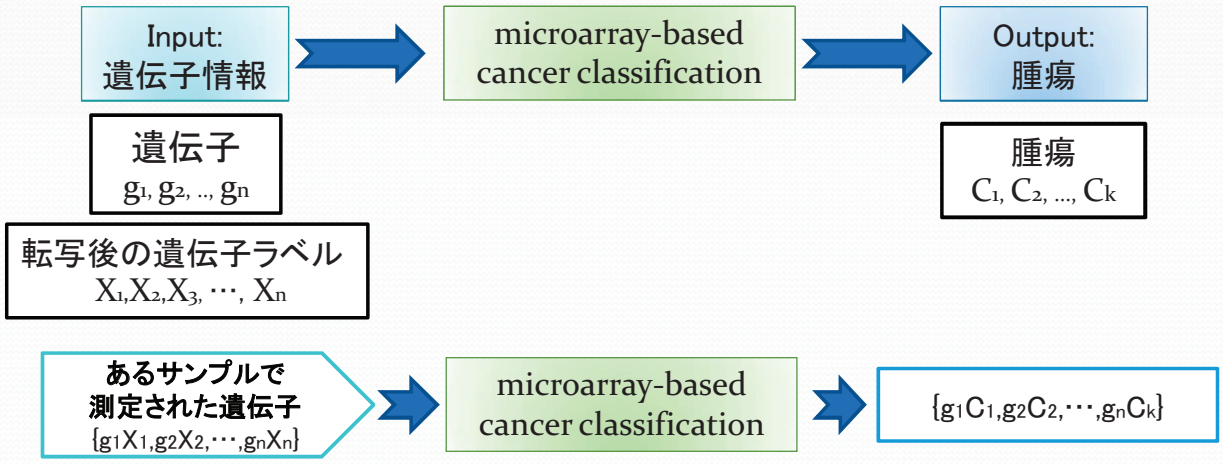
ガンのクラス分けを行うためにマイクロアレイデータから
マーカージンを識別する

DNA microarray technology

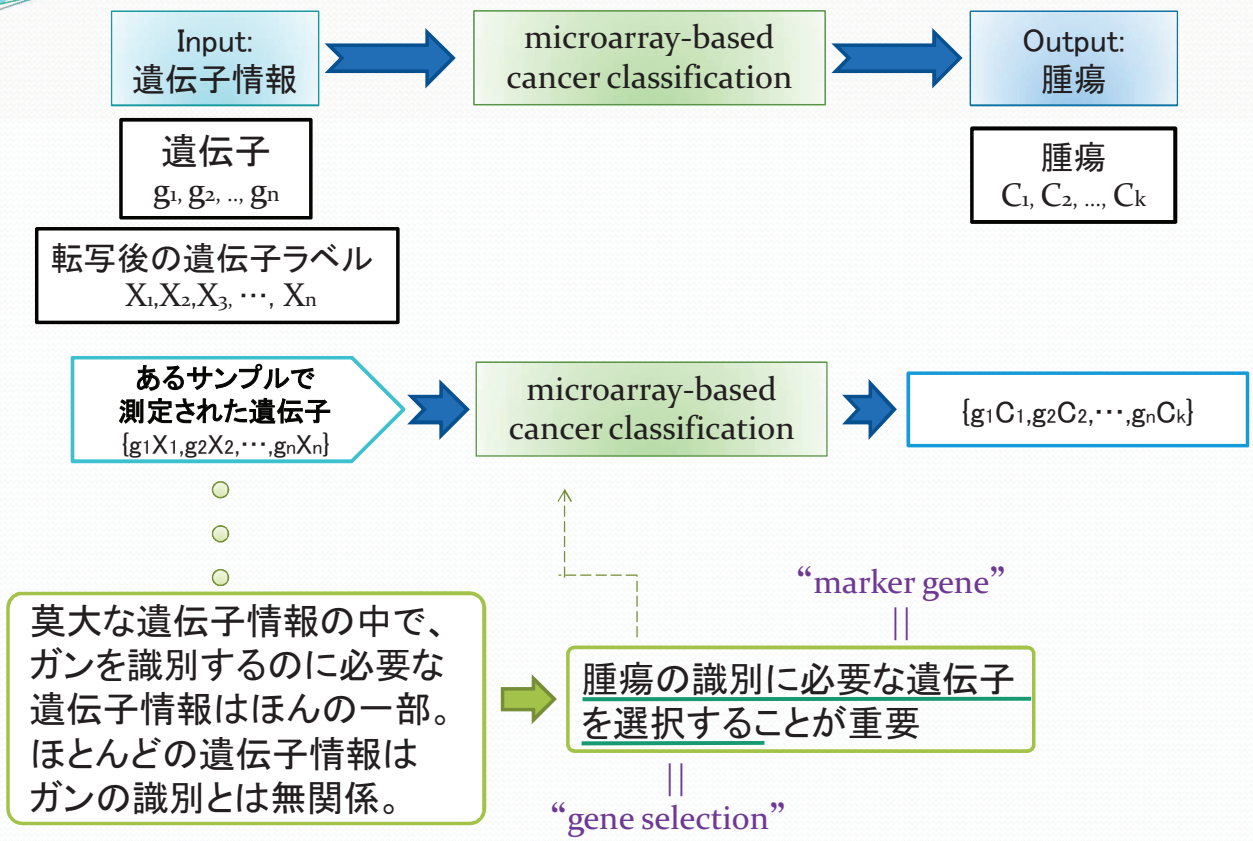
- ガン研究は医学分野において最も重要な研究の一つ。腫瘍を正しく分類することは、ガン診断と薬の開発において非常に重要。
  ⇒遺伝子情報解析によるアプローチが提案されている。
- DNAマイクロアレイ技術の登場により、遺伝子の識別が以前よりも効果的にできるようになった。
  - DNAマイクロアレイ技術
    ⇒数千から数万種類の遺伝子の発現量を一度に解析できる



マイクロアレイ技術に基づいたガンの分類法



マイクロアレイ技術に基づいたガンの分類法



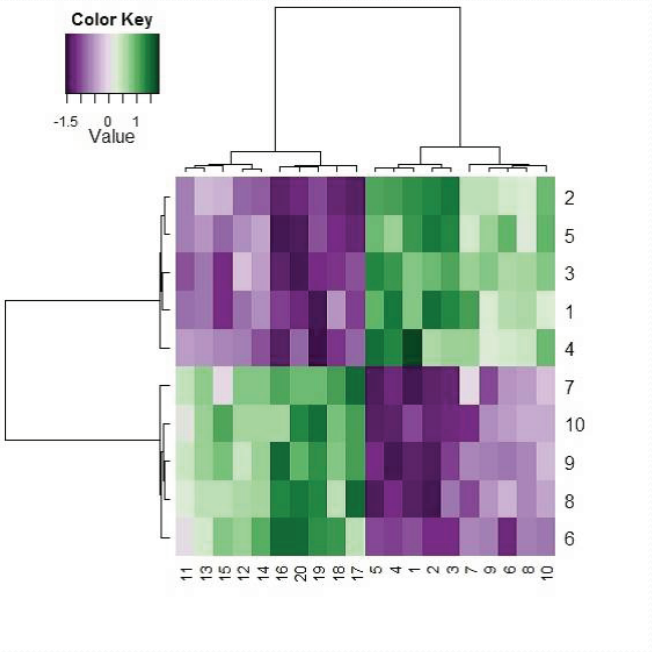
分析手法について

分析手法

- ヒートマップ
- 階層的クラスタリング
- 非階層的クラスタリング
- サポートベクターマシン（SVM）

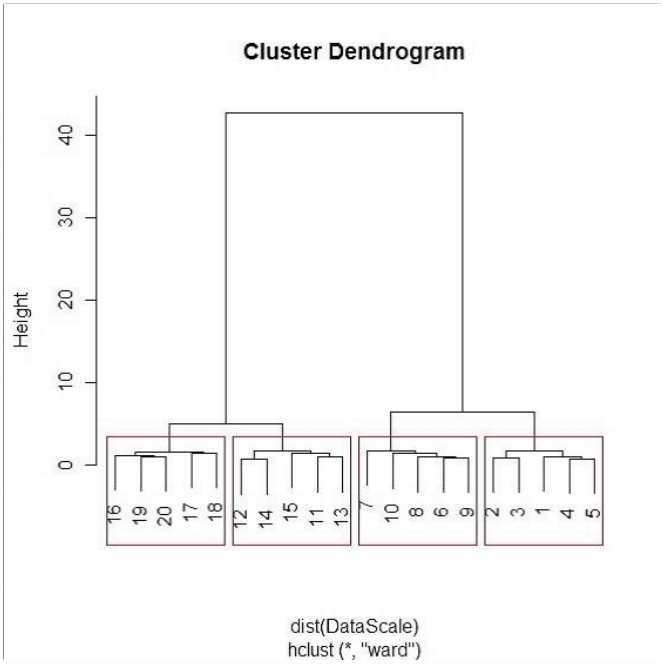
ヒートマップ

- 個体と変数それぞれに対し階層的クラスタリングを適用し、類似度を色の濃淡で表現している。
- 個体と変数を同時に可視化できるため、全体像をつかみやすい。



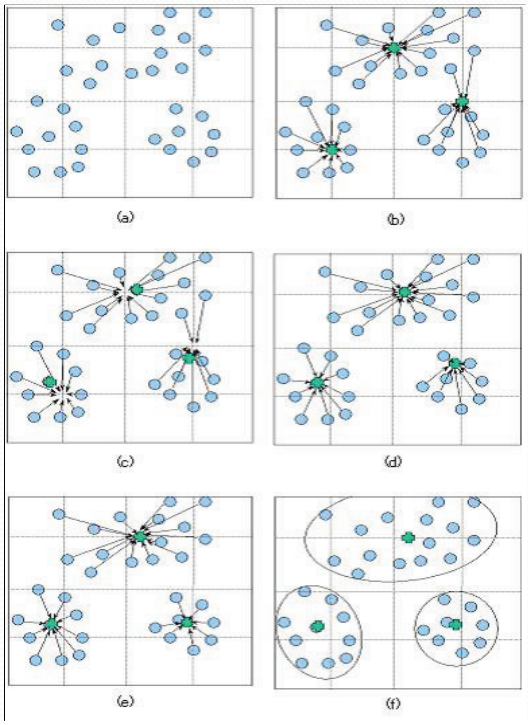
階層的クラスタリング

- 階層的クラスター分析とは、個体間の類似度あるいは非類似度(距離)に基づいて、最も似ている個体から順次に集めてクラスターを作っていく方法である。
- クラスターが作られていく様子を樹形図で示すことができる。



非階層的クラスタリング

- 大規模データのクラスター分析には、非階層的クラスター法が多用されている。
- クラスター数を指定し、クラスター中心の更新とクラスター割り当ての更新を繰り返し行う。
- 目的関数の交互最適化によって実現する。



Support Vector Machine

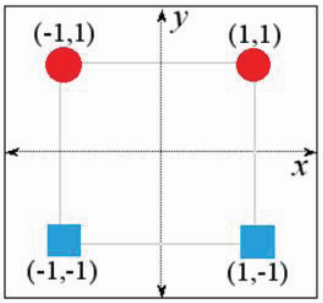
- 多くの手法の中でも認識性能が一番優れた学習モデルの一つ
- 精度が良い分、計算量が多く速度が遅い

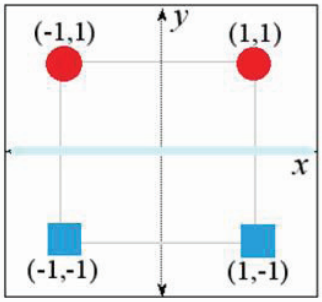
- 優れた認識性能を発揮するのは
  - 線形分離不可能な場合はカーネルを使って高次元での線形分離
  - マージンが最大になるように線形分離

  を行っているため

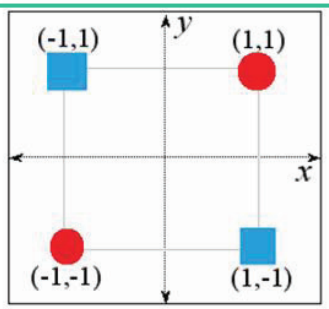
Support Vector Machineとは・・・

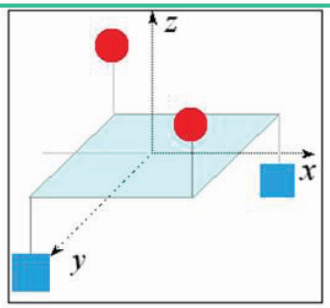
- カーネルを用いた高次元での線形分離





直線 y=0





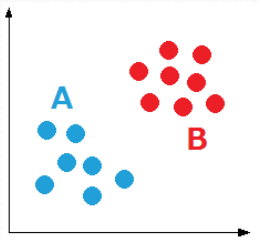
平面 z=0

境界線を一本の直線で引くはできない

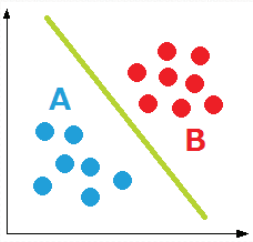
新しい変数z=xyを導入し3次元空間へ写像

Support Vector Machineとは・・・

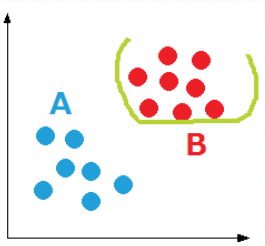
- マージンが最大になるように線形分離



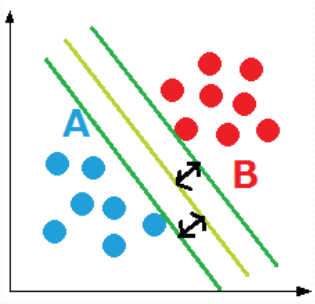
AとBを分離したいときにどのように線を引くか



このように線を引く人が多いが



なぜこのように引かないのか
⇒クラスAのデータの周りの領域がクラスBのよりも大きくなりすぎ



SVMでは・・
データの中で最も他のクラスと近い点を基準として、そのユークリッド距離が最も大きくなるような位置に識別境界を設定

- 他のクラスと最も近い点＝サポートベクトル
- 距離が最大になるようにする＝マージン最大化

作業手順と結果

結果は主にPre_valueとvalueについて紹介する

対象データの内容

- ID_REF:遺伝子番号
- VALUE:対数比(PRE_VALUEのlog2)
- CH1_MEAN:Cy5の信号強度
- CH2_MEAN:Cy3の信号強度
- PRE_VALUE:信号強度の比

- CH2_Mマイクロアレイ(二色法)による糖尿病患者の遺伝子発現データの一部
  - 22575遺伝子×4種の計測値×21サンプル
  - 出典:NCBI(National Center for Biotechnology Information)
    - URL:http://www.ncbi.nlm.nih.gov/

Data table

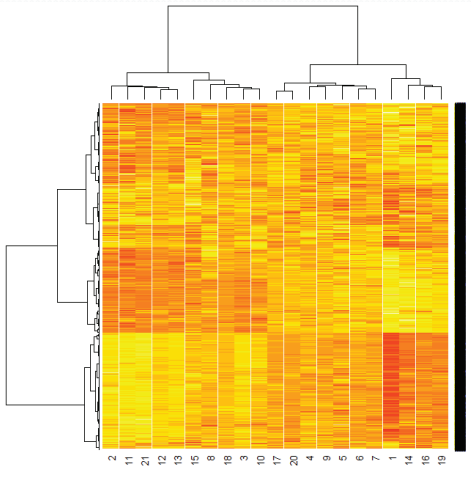
ID_REF	VALUE	CH1_MEAN	CH2_MEAN	PRE_VALUE
1				
2				
3	-0.7622343	18.51991503	31.4119	0.589582771
4	0.2419953	122.3525453	103.4583	1.182626674
5	0.1416103	131.806311	119.4833	1.103135844
6				
7				
8	-0.0992633	595.8232299	638.2613	0.93350988
9	-0.5299413	488.6305953	705.5193	0.692582889
10				
11	-0.0067943	38.24675826	38.4273	0.995301732
12	0.0109753	26.59313788	26.3916	1.00763644
13	0.8848023	16.06021686	8.6976	1.846511321
14				
15	0.1686863	518.0012934	460.8413	1.124034008
16	-0.8596113	189.941626	344.6583	0.551101268
17				
18	-0.3701953	370.1857865	478.4753	0.773677944
19				
20				

作業手順1（ヒートマップの作成）

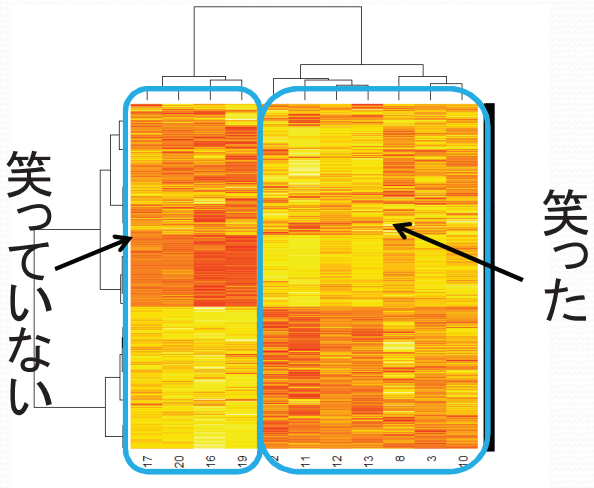
1. データの整形
   - （2万遺伝子×4種計測値）が21ファイル
     ⇒（2万遺伝子×21サンプル）を4ファイルに
   - 値の入っていない遺伝子を削除
2. 解析
   - ヒートマップの作成
3. データの再整形
   - 解析結果からサンプルに対して特徴抽出を行う
   - 特定のサンプルを省いて遺伝子データの再整形
4. 再解析
   - 特徴抽出したデータでヒートマップの作成

解析結果

- valueに対するヒートマップ出力結果
- 階層的クラスタリングの再計算のオプションはWard法
- 対象としたサンプル{2, 3, 8, 10, 11, 12、13, 16, 17, 19, 20}



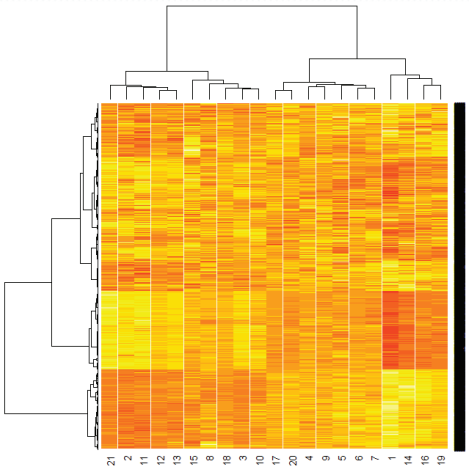
valueに対するヒートマップ



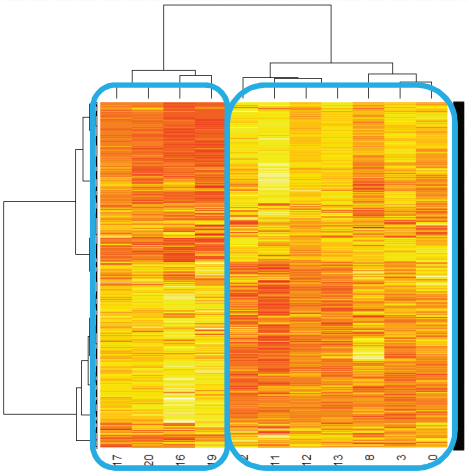
value内の特定のサンプルを
省いたヒートマップ

解析結果

- Per_valueに対するヒートマップ出力結果
- 階層的クラスタリングの再計算のオプションはWard法
- 対象としたサンプル {2, 3, 8, 10, 11, 12、13, 16, 17, 19, 20}



Pre_valueに対するヒートマップ



Pre_value内の特定のサンプルを省いたヒートマップ

ヒートマップに対する考察

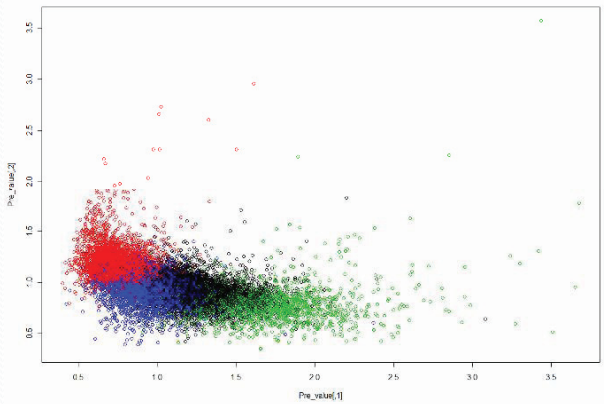
- ヒートマップの出力によって遺伝子とサンプルの全体像を見ることができた。
- サンプルに対する階層的クラスタリングに関して特定のサンプルを省くことで笑ったサンプルとそうでないサンプルをうまく分割することができた。
  - すべてのサンプルで笑うという生理的な要素が数値として明確に表れる保証はないため、うまく分割されるような処理を行うことで特徴抽出をしていることを意味する。
- 遺伝子に対する階層的クラスタリングに関して、データ数が膨大なため樹形図による考察ができず、また計算時間もかなりかかることがわかった。

作業手順2
（非階層的クラスタリングとSVM）

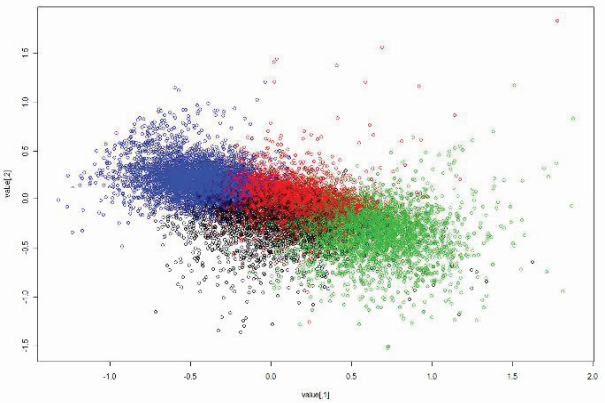
1. データの整形
   - ヒートマップと同様のデータ整形
2. 前処理
   - 非階層的クラスタリングの適用
3. データの再整形
   - クラスターごとに遺伝子データを分割し再整形
4. 再解析
   - 分割された遺伝子データに対しSVMを実行

前処理結果

- 前処理としてPre_valueとvalueに対して非階層的クラスタリングを適用し、Rのplot関数によって出力
- クラスター数は4としている



Pre_valueに対する
非階層的クラスタリングの結果



Valueに対する
非階層的クラスタリングの結果

SVMによる判別結果

PreValue_1

pred	boring	laugh
boring	6	0
laugh	1	14

PreValue_2

pred	boring	laugh
boring	6	0
laugh	1	14

PreValue_3

pred	boring	laugh
boring	2	0
laugh	5	14

PreValue_4

pred	boring	laugh
boring	7	0
laugh	0	14

value_1

pred	boring	laugh
boring	7	0
laugh	0	14

value_2

pred	boring	laugh
boring	7	0
laugh	0	14

value_3

pred	boring	laugh
boring	4	0
laugh	3	14

value_4

pred	boring	laugh
boring	6	0
laugh	1	14

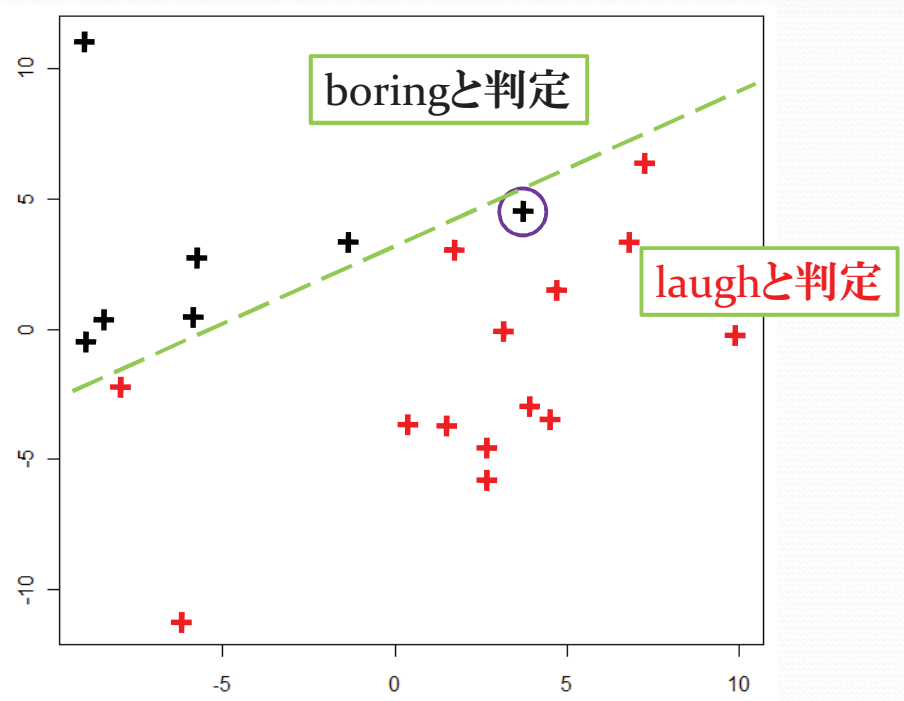
解析結果（SVMの結果をマッピング）

PreValue_1

pred	boring	laugh
boring	6	0
laugh	(1)	14

+ boring
+ laugh

PreValue_1を2次元に射影した図



SVM分析結果に対する考察

- 高次元小標本データの次元の縮約を行うために、非階層クラスタリングを用いた。得られたクラスターごとにSVMを適用した結果、良い判別結果が得られた
- 漫才を聞いた患者と退屈な講義を聞いた患者に基づく識別関数の同定ができた

  - 「笑い」が糖尿病患者の遺伝子発現に対し何らかの関与をしているのではないかという可能性が伺えた
  - 高次元小標本データの分析は一般に“次元の呪い”により、通常の多変量解析による分析ができないが、本グループ演習で実施した解析法により適切な識別が可能となった

まとめ

- 糖尿病の「笑い」による効用を解析
  - ヒートマップを用い，サンプルを抽出することで識別子同定を可能とした
  - 高次元少標本データの問題を解決
    - 非階層クラスタリングによって遺伝子(次元)分類し，次元を縮約
  - クラスタリングした遺伝子群に対しSVMを適用
    - 識別関数を同定する方法を提案．明確に判別が行われ，同定に成功した
- 糖尿病患者治療に対する笑いの効用を示唆する結果が得られた

今後の展望

- より多くのサンプルに対して実行
  - 提案法の妥当性検証の必要性
- 遺伝子の専門家の見地から調査
  - 明確な分類結果が得られたので…
    - 同じクラスターに属する遺伝子についてより詳細な特徴を見出せる可能性
    - 分類することのできた被験者に共通する性質が見出せる可能性
- 「笑い」には個人差がある
  - 個人差をどのように適切に分析に反映させるか