サンプル・データ

	結婚	年齢	収入
顧客 A	Y	Y	L
顧客 B	Y	Y	M
顧客 C	Y	M	M
顧客 D	Y	Y	L
顧客 E	Y	M	M
顧客 F	N	M	M
顧客 G	N	E	H
顧客 H	N	E	H
顧客 I	N	M	M
顧客 J	N	E	E

結婚	年齢	収入

YES	NO	25-35	35-50	51-65	-25	26-40	>40
Y	N	Young	Middle	Experienced	Low	Mid	High

類似度マトリックス

	A	B	C	D	E	F	G	H	I	J
A	3	2	1	3	1	0	0	0	0	0
B	2	3	2	2	2	0	0	1	0	0
C	1	2	3	1	3	1	0	1	1	0
D	3	2	1	3	1	0	0	0	0	0
E	1	2	3	1	3	1	0	1	1	0
F	0	0	1	0	1	3	2	1	3	2
G	0	0	0	0	0	2	3	2	2	3
H	0	1	1	0	1	1	2	3	1	2
I	0	0	1	0	1	3	2	1	3	2
J	0	0	0	0	0	2	3	2	2	3

任意のレコード間の類似度は、一致している値の個数として計算される

類似度と相違度

クラスタリング例： {FGIJ} {ABC} {DE} {H}

クラスター内の類似度

合計 = 78

	F	G	I	J	A	B	C	D	E	H
F	3	2	3	2	0	0	0	1	1	1
G	2	3	2	3	0	0	0	0	0	2
I	3	2	3	2	0	0	0	1	1	1
J	2	3	2	3	0	0	0	0	0	2
A	0	0	0	0	3	2	3	1	1	0
B	0	0	0	0	2	3	2	2	2	1
C	0	0	0	0	3	2	3	1	1	0
D	1	0	1	0	1	2	1	3	3	1
E	1	0	1	0	1	2	1	3	3	1
H	1	2	1	2	0	1	0	1	1	3

クラスター間の相違度

合計 = 168

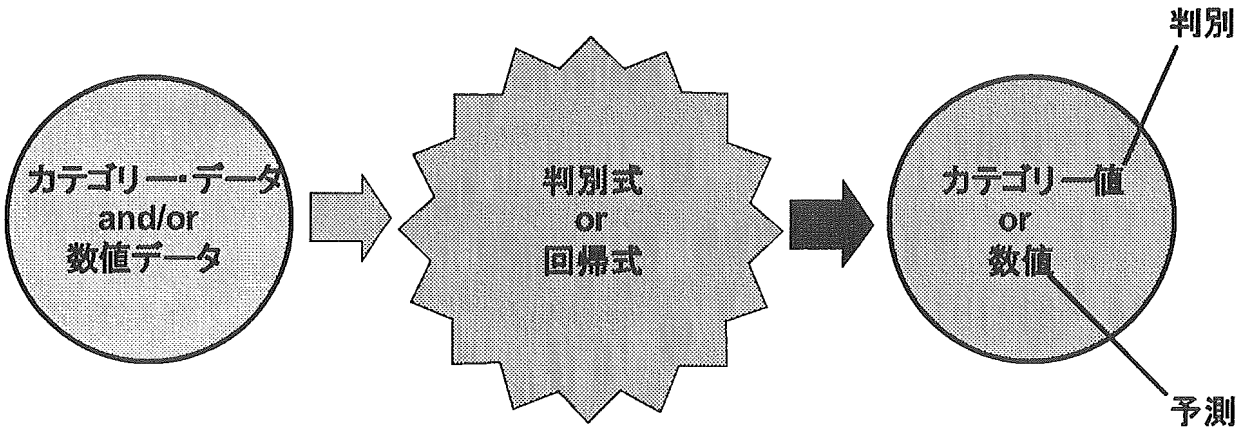
	F	G	I	J	A	B	C	D	E	H
F	0	1	0	1	3	3	3	2	2	2
G	1	0	1	0	3	3	3	3	3	1
I	0	1	0	1	3	3	3	3	3	2
J	1	0	1	0	3	3	3	3	3	1
A	3	3	3	3	0	1	0	2	2	3
B	3	3	3	3	1	0	1	1	1	2
C	3	3	3	3	0	1	0	2	2	3
D	2	3	2	3	2	1	2	0	0	2
E	2	3	2	3	2	1	2	0	0	2
H	2	1	2	1	3	2	3	2	2	0

合計の和 = 246

モデリング

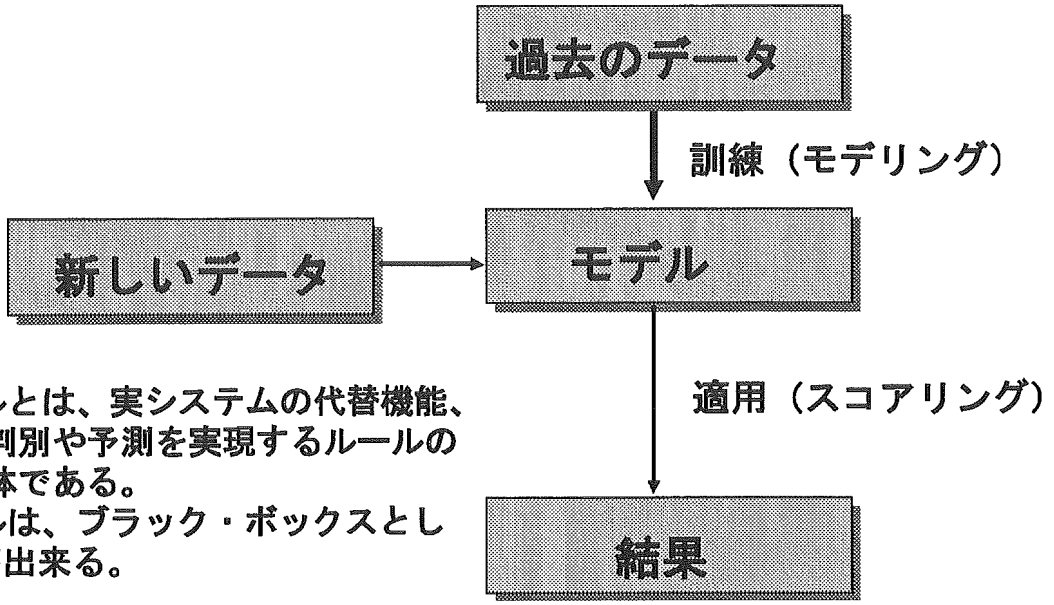
(1)判別(クラシフィケーション)

(2)予測(数値予測)



非線形関数

モデリングとスコアリング



●モデルとは、実システムの代替機能、即ち判別や予測を実現するルールの集合体である。
●モデルは、ブラック・ボックスとして使用出来る。

クラスシフィケーション

①線形判別（伝統的なクラス判別法）

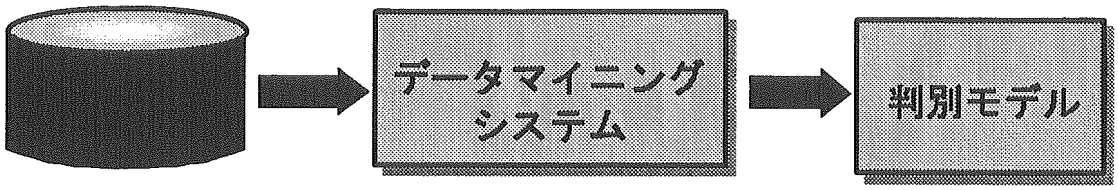
②決定木（Decision Tree）

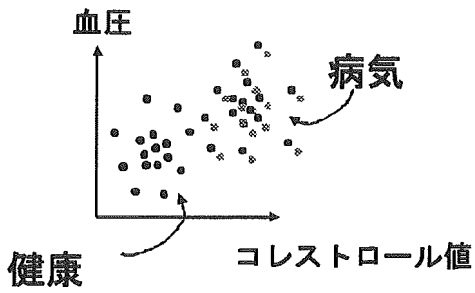
③ニューラル・ネットワーク（Back-Propagation）

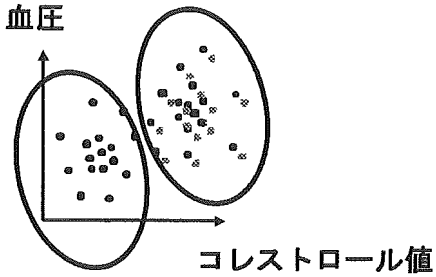
データマイニングに用いられるクラシフィケーション手法として、代表的な手法に決定木とニューラル・ネットワークを挙げることが出来る。、クラシフィケーションを習得するため、先ず簡単な線形判別分析から言及する。
なお、MBR（Memory Based Reasoning）や新しいSVM（Support Vector Machine）などは、Intelligent Minerなど代表的な 汎用データマイニング・ツールには採用されておらず省略する。

備考]クラシフィケーション(Classification)、クラス判別、クラス分類、分類などとも呼ばれる。

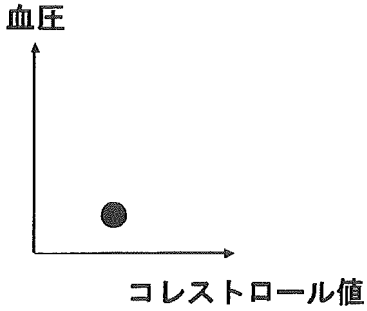
モデリング・フェーズ

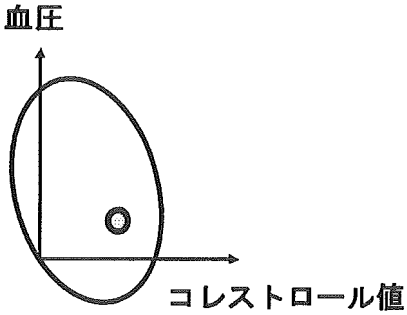






スコアリング・フェーズ



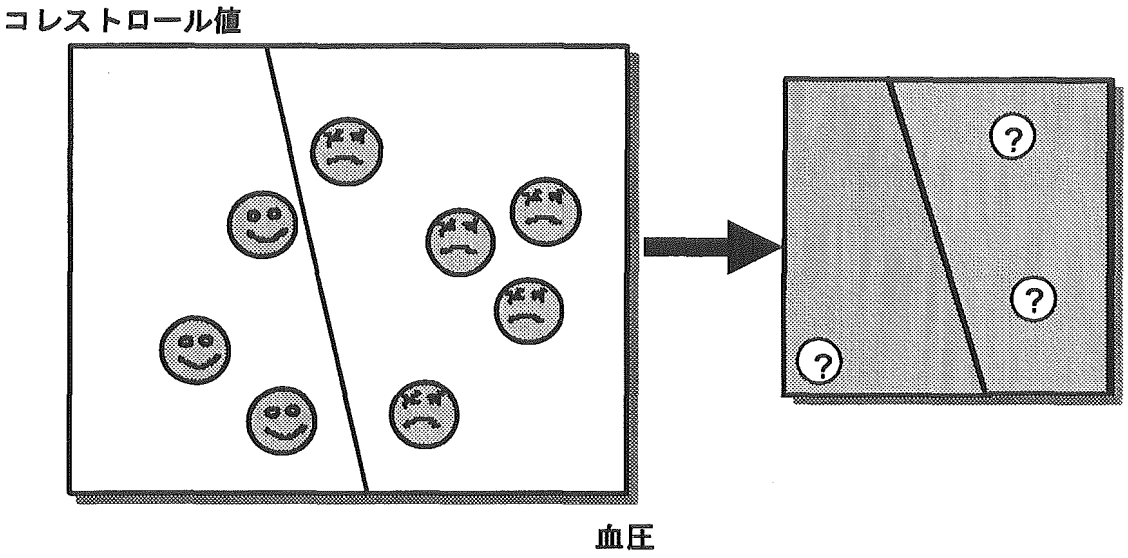


健康

備考]スコアリングとは、ここでは判別の予測を行うことを意味する。

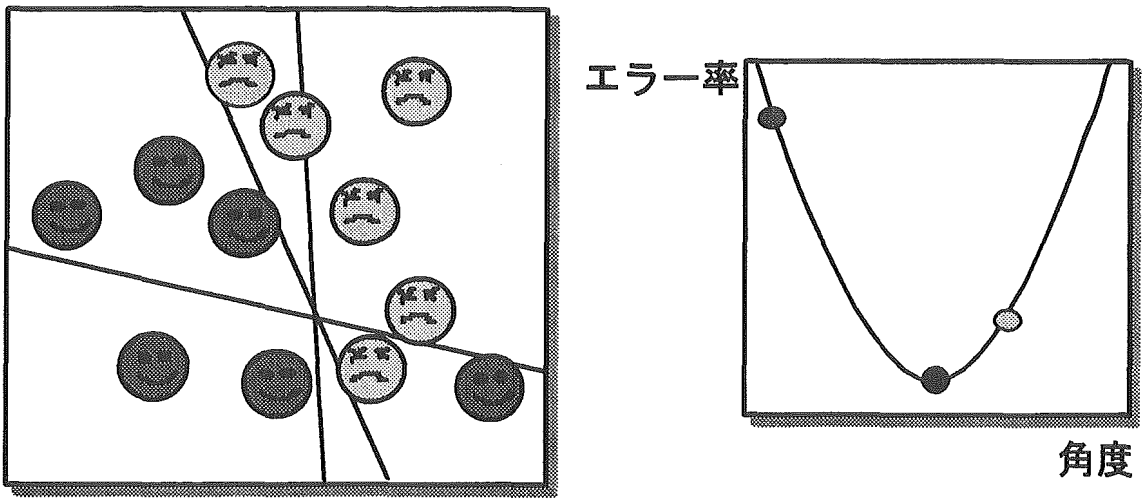
線形判別

●基本的な線形判別、マハラノビス判別、ロジステック判別などは空間の二分割に留まる。（後者の２つは軽い非線形）



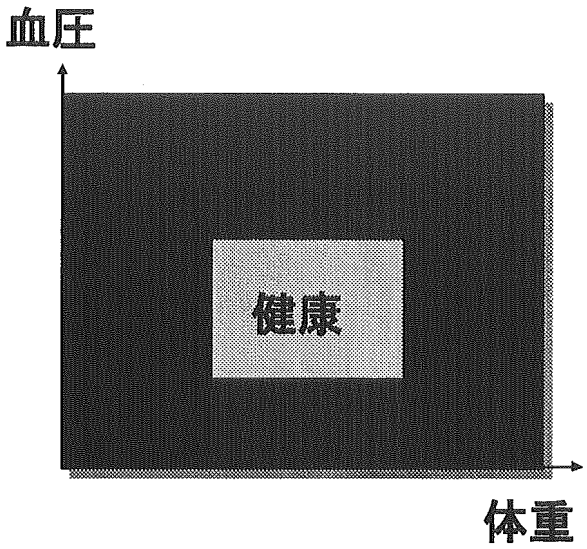
判別関数の求め方

●初期モデルから始めて、判別のエラー率を調べ、モデルを調整し再度判別エラーを評価する操作を繰り返す。



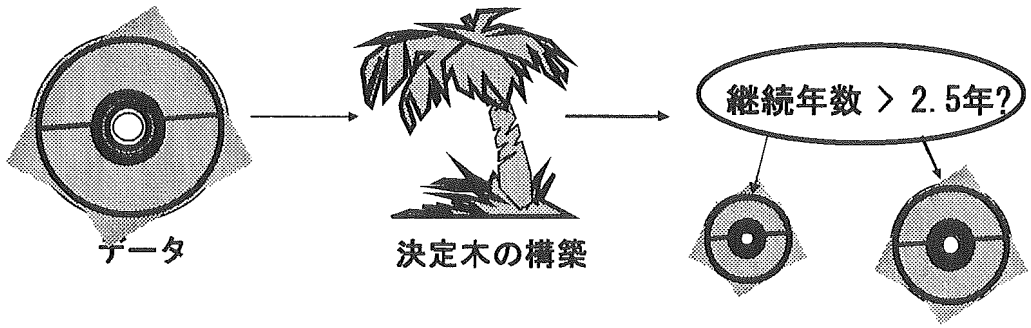
線形モデルの弱点

●簡単な問題なら、線形モデルを使えるが。
●複雑な問題では、判別率を上げるため、色々努力を要するし、解決できないケースもある。
●決定木やニューラル・ネットでは、このような問題に患わされる事はない。



決定木

●トップダウン・アプローチ、再帰的分割

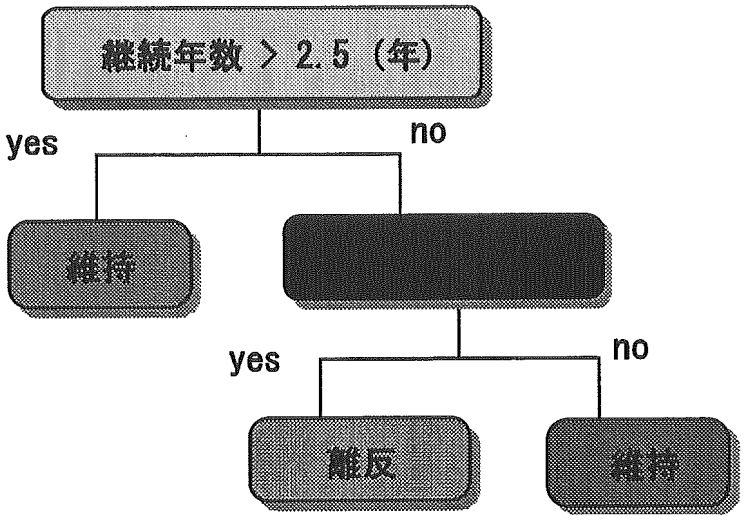


●構築された決定木からルールを抽出でき内容解釈を容易化する。
●構築された決定木モデルを未知のデータに適用し判別を行う。

備考]最初に登場した決定木アルゴリズム(1963)は、統計手法をベースにしたAIDアルゴリズムで、そこから発展したCHAIDを組み込んだ統計パッケージもある。
なお、データマイニング・ツールには、一般に機械学習をベースにした新しいアプローチの決定木アルゴリズムが採用されている。

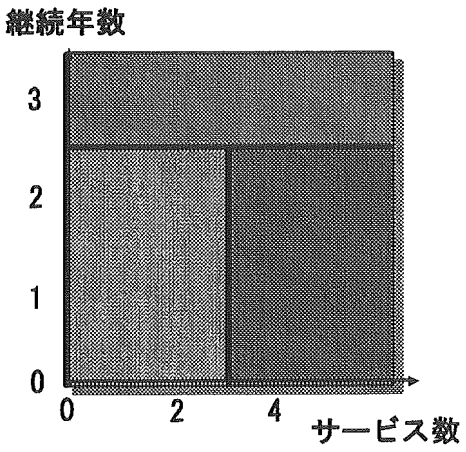
決定木による判別

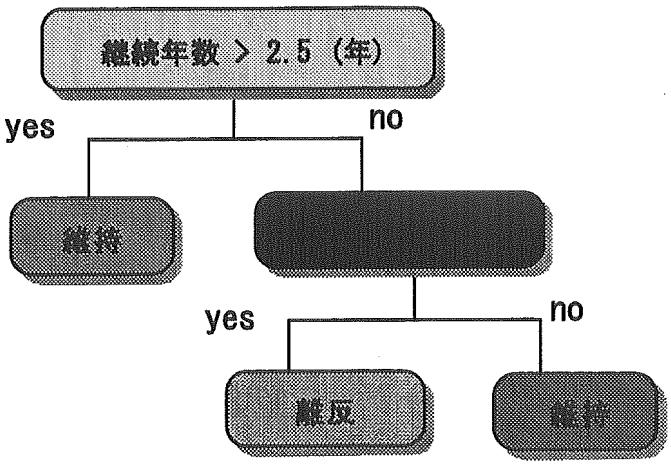
●決定木モデルは、複数の判別式の繋りであり、適切な結果のクラスに辿り着いている。
●例： 顧客維持データ



決定木によるサンプル

●各質問によって空間を分割して行き、グループが分割されるにつれて各グループは純化されていく。（帰納法）

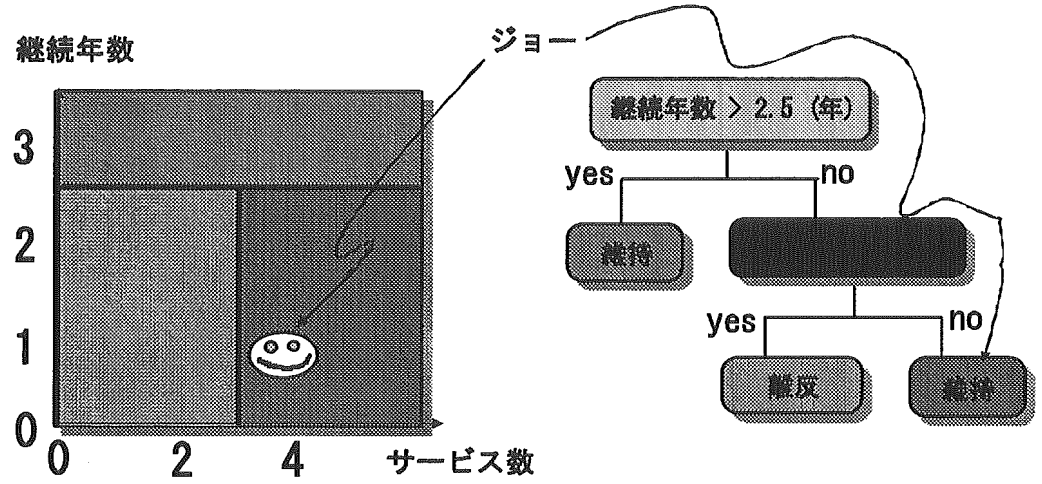




●各末葉は顧客の空間の領域を表している。

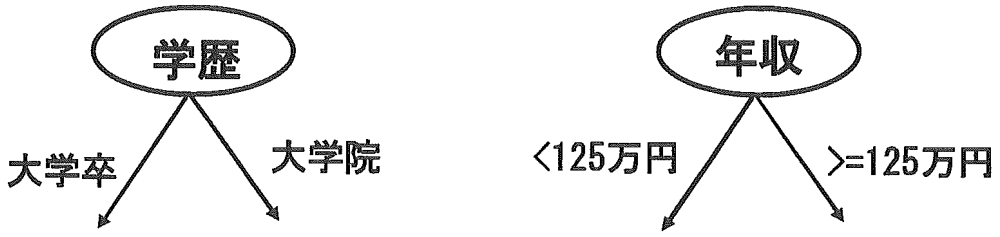
決定木モデルにより新規データのクラス判別

●ジョーは、顧客になってから一年経ち、その間に四回のサービスを受けている。
●決定木は、ジョーが離反しないと予測した。



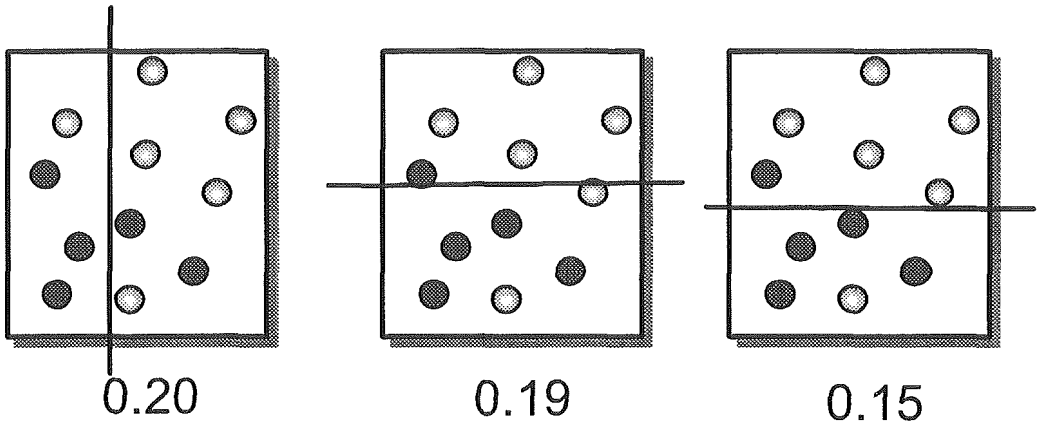
全体空間の分割

●全体の戦略は“分割して統治せよ”
●各々のステップでデータを二つ以上に分割する。
  –分割の候補は？
  –最適にするための評価の仕方は？
●候補：
  –カテゴリー値は、特徴的な値の集まりがあるか調べる。
  –連続値は、しきい値で分ける。



分割候補の評価

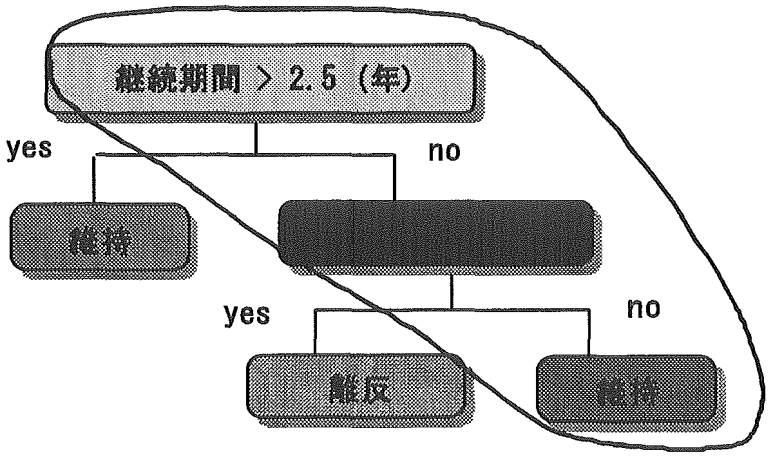
●目的： 出来る限り違いの出る分割にする。
●分割した各領域を不純度(GINI)で評価する。
●不純度の低い分割を採用する。



GINI＝p(1－p)

決定木モデルの読み方

●決定木の枝葉は、入力データ全体を相互排他的に分割するルールの集まりです。



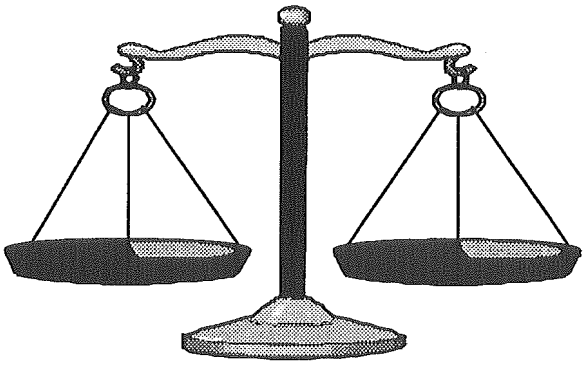
●継続年数 < 2.5年 かつ サービス数 >= 3 の場合、顧客は維持される。

決定木の枝刈り（1）

●一般に、決定木はトレーニング・データに完全フィットできる。
●決定木を枝刈りすることによって、未知データに対する予測の精度を向上できる。
●枝刈り方法：最小記述長（MDL:Minimum Description Length）
ルールとデータの記述長を最小にするために枝刈り(枝を葉を置換える)を繰返し行う。
利点:未知データの予測に良い適合をする本質的パターンに適合し、ノイズに患わされない小さな決定木読みやすさ、理解のしやすさ

決定木の枝刈り（２）

訓練の正確性



予測での正確性，モデルの明確性

●全てのデータマイニング手法は、訓練データでの正確性とモデルの汎化性とのトレードオフの間に置かれる。
●Occam の剃刀： 単純なモデルが最良である。

決定木モデルの評価

●テスト・データの分類結果の正誤表を見ると、正答率と誤答率が分かる。
- 正解率 + 誤差率 = 100%
- 誤差　１４５ (145/1322=11%)

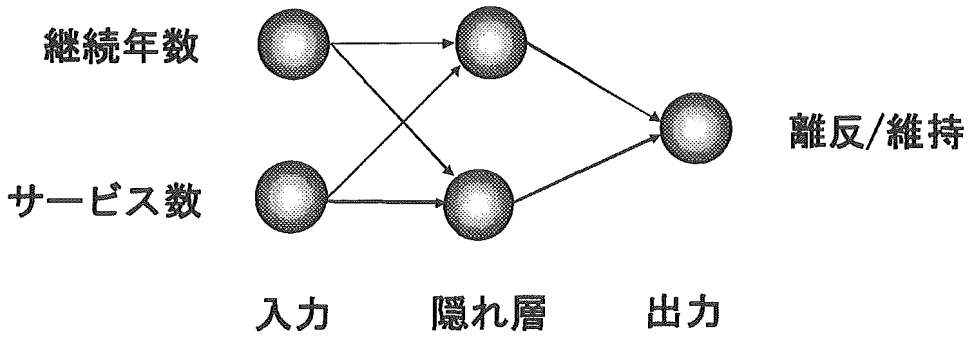
決定木による正誤表

		離反	維持
実績	離反	562	132
	維持	13	615

誤答　正答

パックプロパゲーション法

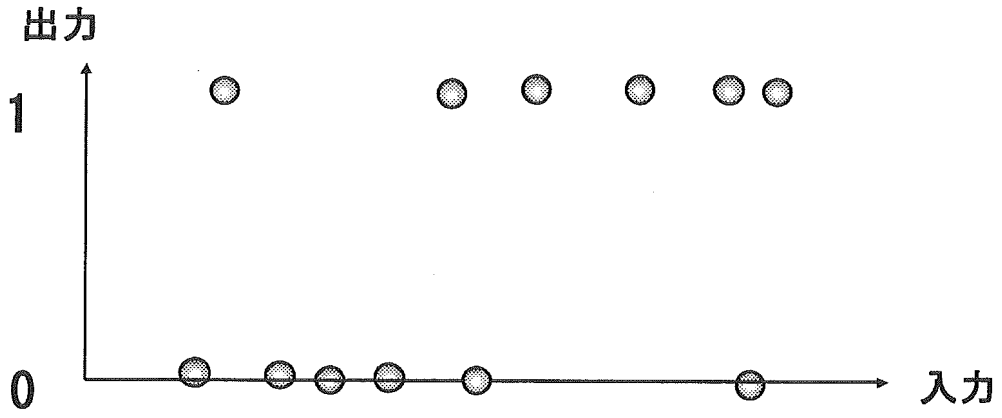
●モデルは「ニューロン」の集まりであり、各々のニューロンは各々特定の強さを持って接続されている。



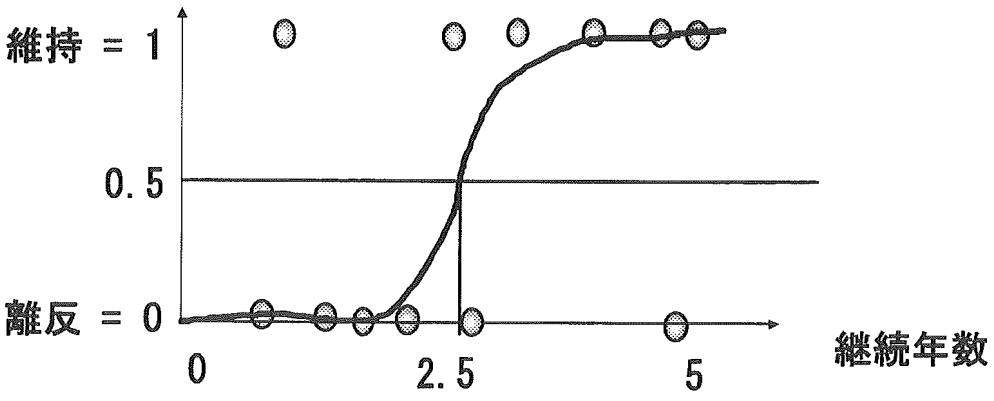
備考]BP(バップロパゲーション)法は、ニューラルネットワークを用いる最も代表的なモデル化手法で、一般にニュラルネットワーク技法と云えば、BP法を表している。また、隠れ層がないニューラルネットワークは、統計的手法のロジスティック回帰と等価である。

自分がニューラル・ネットになってみる

●ニューラル・ネットは、博士論文が書けるくらい難しいが、使うだけなら簡単である。
●まずは試してみよう： 連続曲線を使ってデータをフィットさせる。



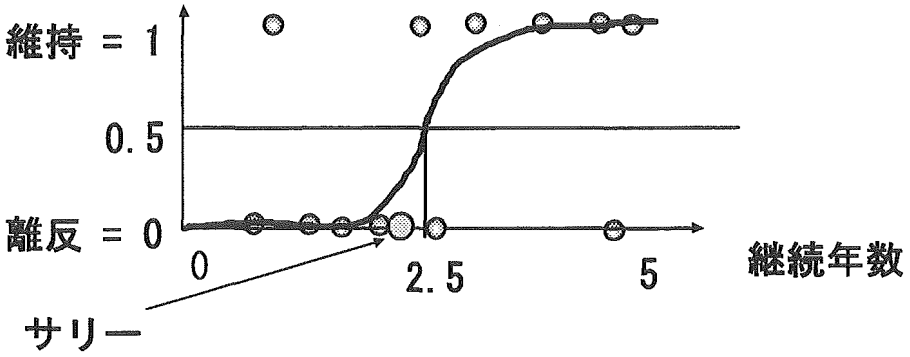
シグモイド関数



曲線を引くことによって、あなたは小さなデータマイニングを行った！（しかし、まだまだ先がある...）

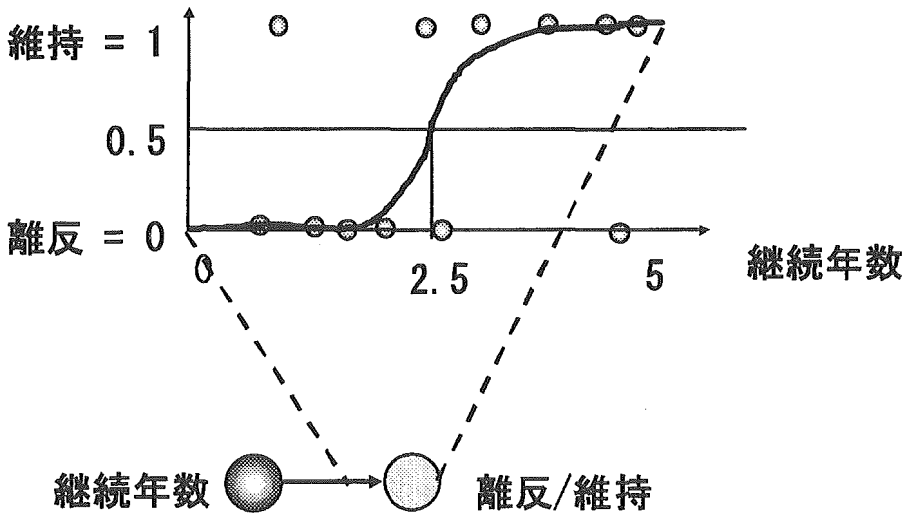
備考]ロジスティック関数、S字型関数とも呼ばれ、$f(x)=1/(1+\exp(-\lambda x+\theta))$ で表される。

モデルとして曲線を採用



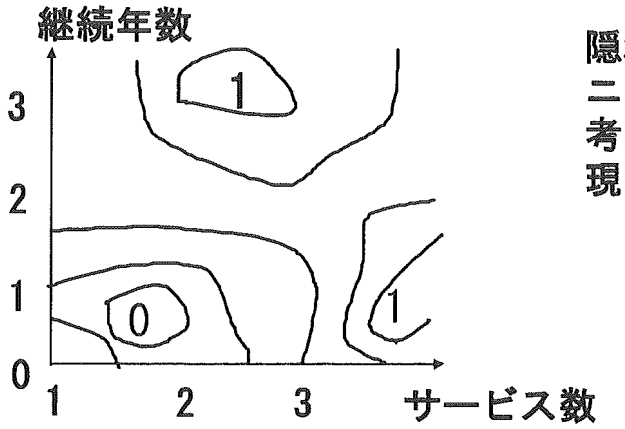
サリーちゃんは顧客として2年3ヶ月継続している。
彼女は何処に分類されるか?
継続年数=2.5年（2年6ヵ月）での値は0.5である。
選択肢としては、彼女は「離反」か「不明」

単一のニューロン

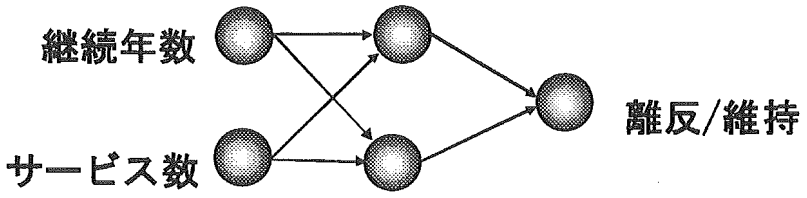


重みとニューロンが曲線の位置, 向き
および傾きを決める

複数のニューロン



隠れ層ニューロンと出力層ニューロンをN次元空間で考えると、複雑な曲面を表現できる。



ニューラル・ネット学習の追加情報

●学習の停止?

-自動/半自動停止には、幾つかの方法があるが、それぞれ欠点を持っている。

例えば、

- Intelligent Miner の方法は：
  - 精度の目標値： 正解率の目標値を与える
  - 誤差限界 ： 誤差率の最大値を与える
- 正解率 + 誤差率 + 不明率 = 100%

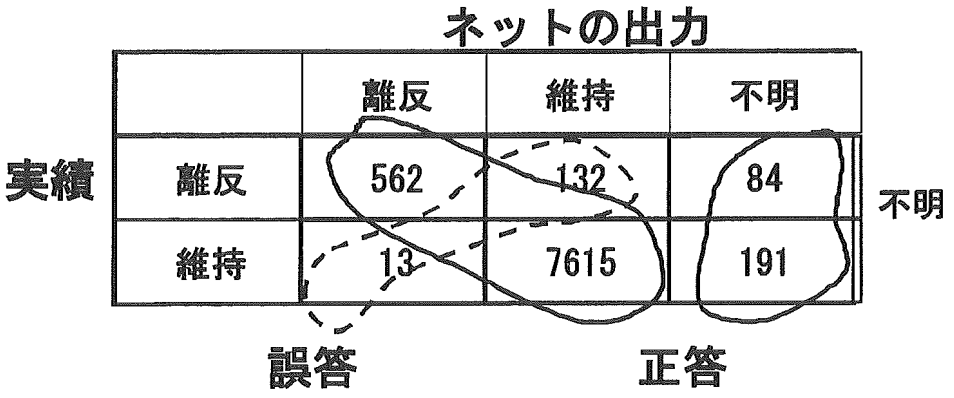
ニューラル・ネットへの入力

●カテゴリー入力は：
- 「N者択一」にコード化する
  - (例えば、"赤", "青", "緑"をそれぞれ"001", "010", "100"にする。)

●数値入力は：
- 正規化してもよい。そうすれば、元データの絶対値に依存せず良い結果になる場合が多い。
  (例えば、時間単位が日でも年でも結果に影響しない)

ニューラル・ネットによる結果

●テスト・データの分類結果の正誤表を見ると、正答率と誤答率が分かる



モデルの解釈

●ニューラル・ネットは厳密なモデル化
を期待できる。
●ニューラル・ネットは解釈しにくい。

Intelligent Minerでは、
○ニューラル・ネットの適切な構造を決定する(隠れ層ニューロンの数,隠れ層の数)
○エージェントが、ユーザーに代わって最適な構造を決定するネットワークの学習
即ち、重みを調整するため、逆伝搬を使う(Sigmoid曲線の形状と位置を調整する)過剰学習を避けるためにサンプル・データの一部を検証のために残しておく出力にたいする入力変数の感度解析を行い、どの入力変数が重要かを教えてくれる。
なお、ニューラル・ネット・モデルよりも、この入力変数の重要度の方に価値を見いだすユーザーがいる。

予測

①ステップワイズ多項式回帰(多変量解析の一種)

②回帰木(Regression Tree)

③ニューラル・ネットワーク(Back-Propagation)

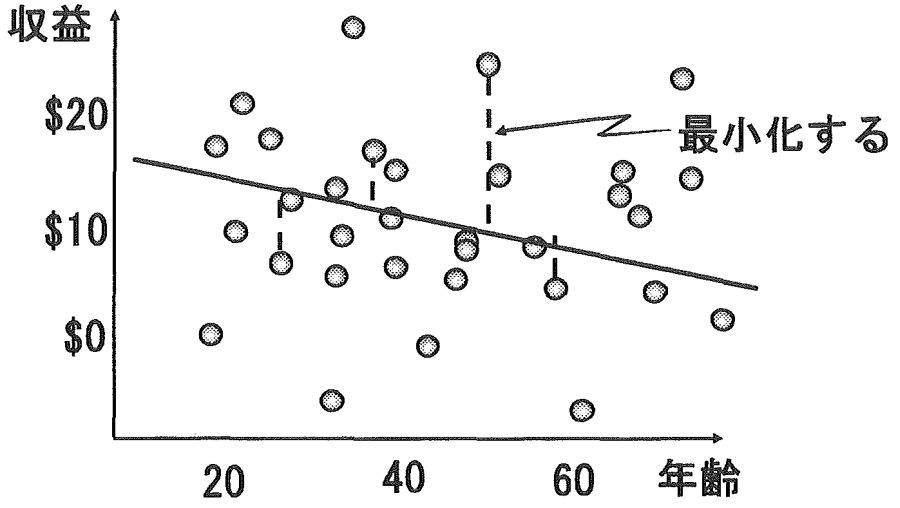
④ラジアル・ベーシス・ファンクション(Radial Basis Function)

ステップワイズ多項式回帰については特に述べないが、典型的な非線形回帰のRBFを習得するため、簡単な線形回帰と比較しながら述べる。
また、回帰木とニューラル・ネットワークは、判別の処で記述した内容と類似のため省略する。

備考]予測手法の代表的アプローチとして、回帰と比較の2つの手法が挙げられる。

線形回帰

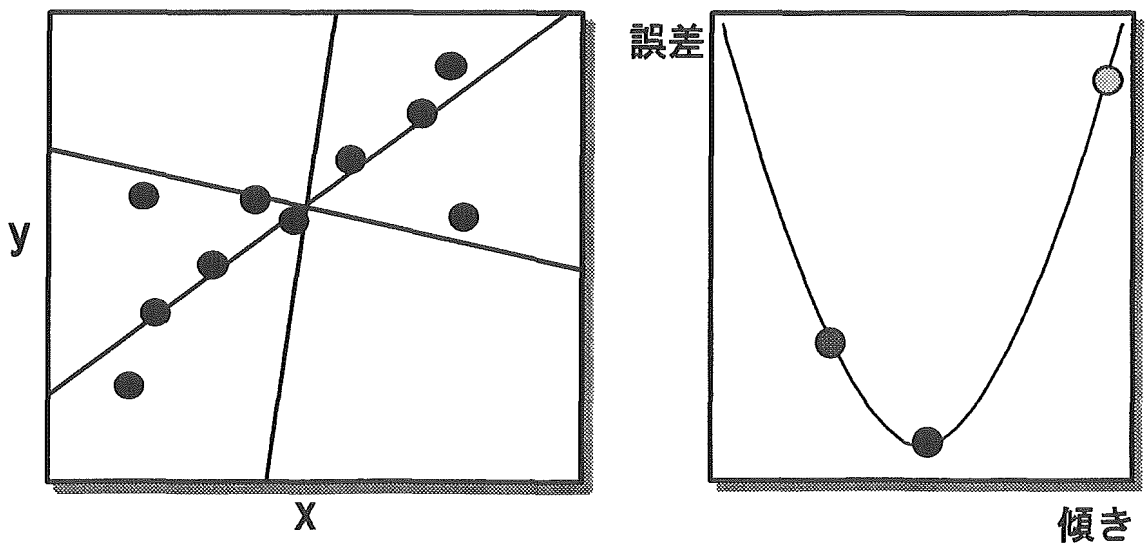
●全ての点に出来る限り近くなるような直線を見付ける。



備考]近似直線f(x)の係数は最小二乗法により求める。

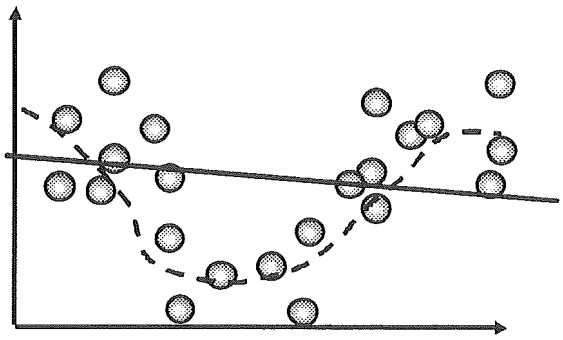
線形回帰式の求め方

●誤差の作る曲面の広域的な最小値が単一であれば、最適な線形回帰式は容易に求められる。

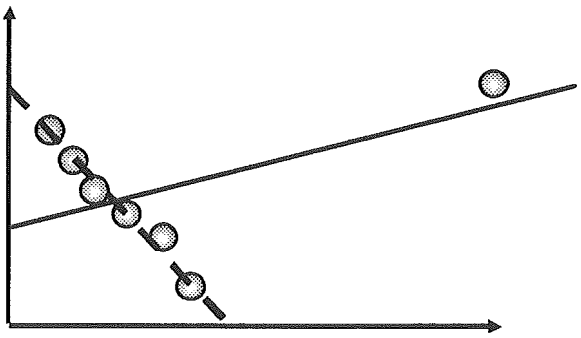


線形回帰の欠点

●線形のみしか扱えない。

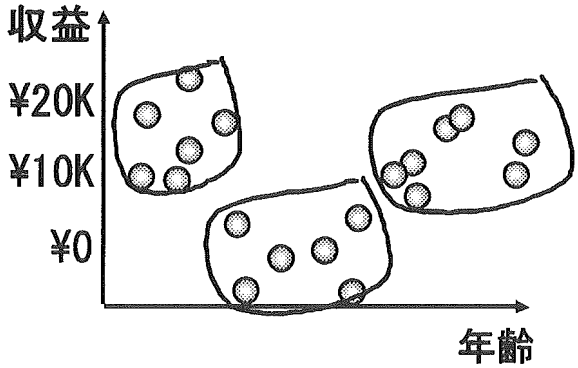


●わずかな異常値に大きな影響を受けやすい。

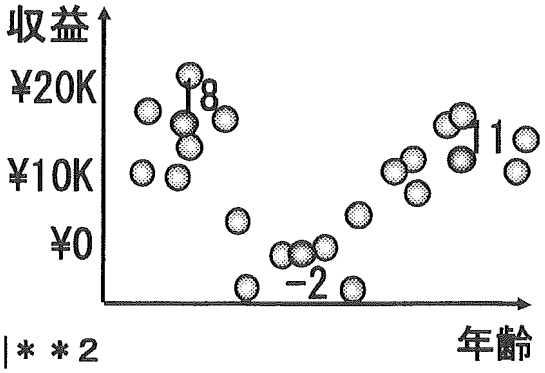


ラジアル・ベーシス・ファンクション法

●入力値の似通ったグループの作る領域を見つける。



●各領域に対してＲＢＦの中心を求める。この値がこのグループの平均出力となる。

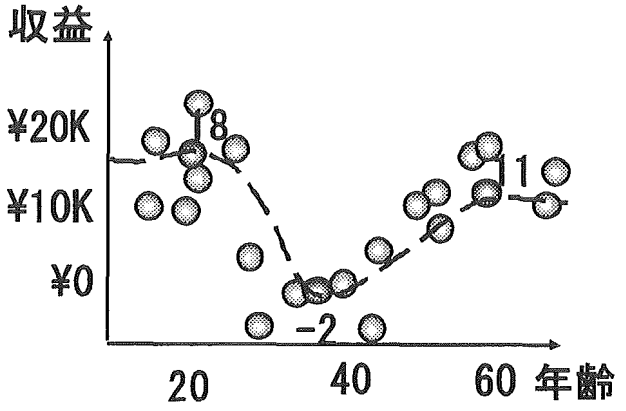


$$H[f] = \sum_{i=1}^{n} (y_i - f(x_i)) * *2 + \lambda ||Pf|| * *2$$

$$f(X) = \sum_{j=1}^{n} (Wj * \exp(-(X-Cj) * *2 / 2\sigma * *2))$$

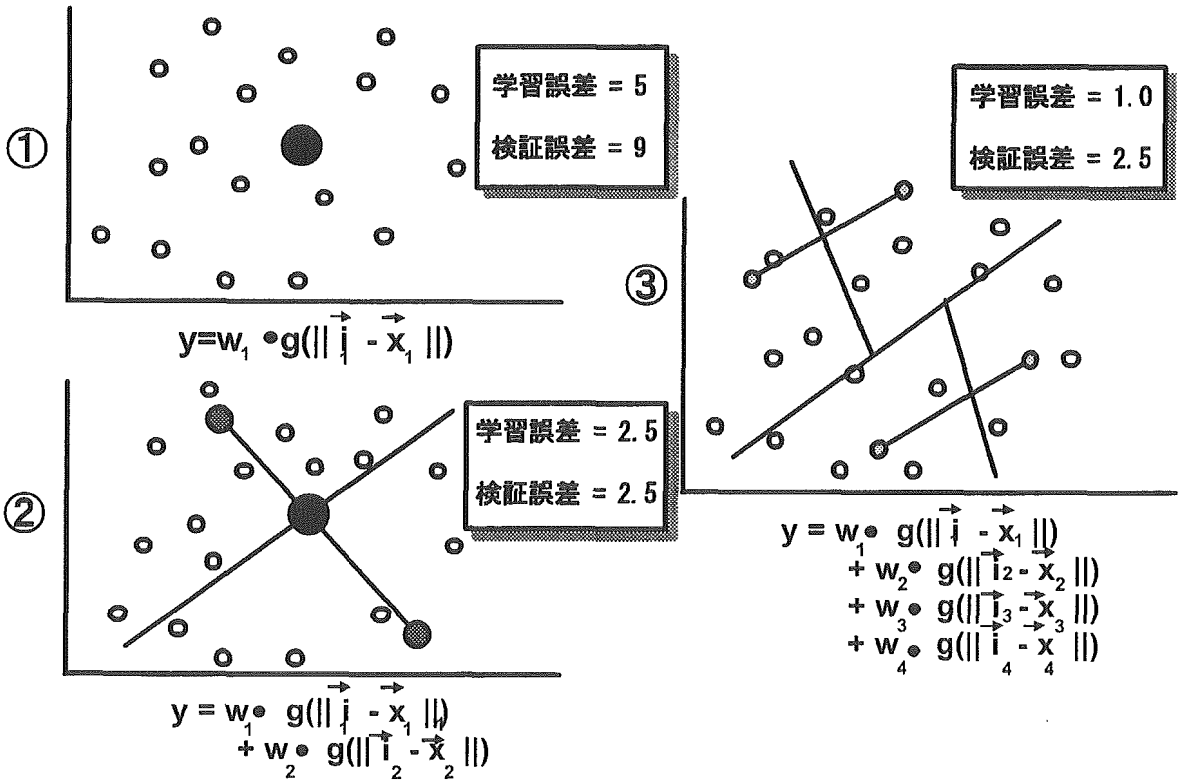
複数個のＲＢＦによる回帰曲線

●クラスター中心間の値を内挿するためＲＢＦ中心の加重平均をとる。



●各中心点から離れるにしたがって重みは急激に減少する。
クラスター中心に近いほど値は中心の予測値に近くなる。

ＲＢＦの原理



ＲＢＦ設計における配慮

●モデルの予測値と実際の値との差を最小化するだけでは最適とはいえない。ノイズの影響を考慮しなければならない。
●ノイズに対して過剰学習する恐れがある。

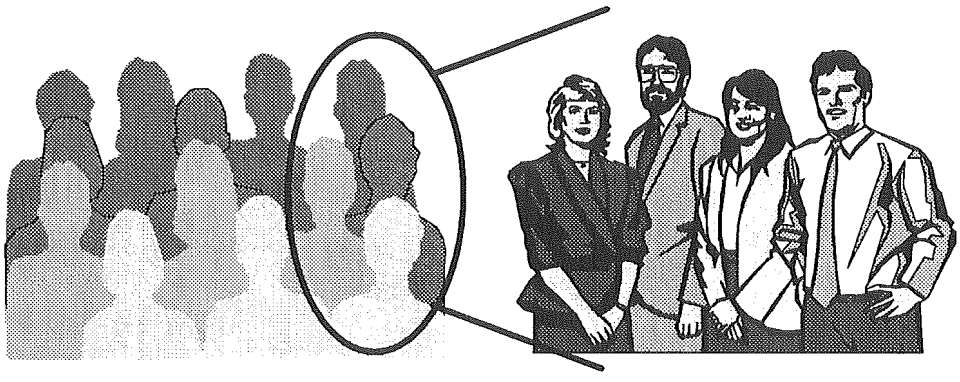
●RBFは1/3のデータをテスト用に残し、このデータを用いて予測精度を検証し繰返し、計算による改善が見られなくなった時に停止する。

●ユーザーは、繰返し計算回数の最小値と最大値を指定することも出来る。

ＲＢＦの方見

●ＲＢＦは目的変数の正確な予測が出来るようデータをグループ化する。
●更に同じグループのデータが持つ特性を明らかにする。

グループ Ａ



備考]RMSE(最小二乗誤差)が所定の許容値に収まるまで改善を繰り返す。