意　義

1. 要求
   1) DB投資の回収
      ＤＢに多額投資して蓄積された膨大データの活用により資金回収したい。
      備考]医療分野ではカルテの電子化が課題であるが。
   2) データ資源の積極活用
      日常業務で付帯的に発生する大量データ、この社内最良の資産放置は許されない、何とか経営に役立てられないか。

2. 課題
   1) 巨大データの適正処理
      大量データを網羅的に全件分析できないか。
      注]サンプリング分析を想定の統計解析では精緻な分析は困難である。
   2) 複雑な問題の解明
      複雑な多変数問題を適正に処理できないか。
      注1]多変量解析では概して数個～数十個程度の変数に留まっている。
      注2]多変量解析では線形モデルがベースで非線形性が強ければ扱えない。

3. 意義
   従来手法で予見困難な知識も、データマイニングで洞察を実現する。
   備考]洞察(Insights)とは幾ら検索しても検出できない知識の把握を意味する。

主要機能

リンク分析　：有限事象間における全ての関連ルールを抽出する。
備考]予備解析や補足解析に使用されることもある。

クラスタリング：サンプルの層別により対象システムを解明する。
備考]予備解析に使用されることも多々ある。

モデリング　：対象システムの解明を行うと共に、推定モデルを生成し数値予測やクラス判別を行う。
備考]一般に、モデリングには付帯的にスコアリングも含められる。

逸脱の検出　：正常からの逸脱を検出する。(新しい研究分野)
備考]自動検出アルゴリズムの研究は継続されているが、中々実用化レベルに到らず、上記3機能により代替されている。

備考]上記機能を実現するデータマイニング基礎技術は、機械学習、統計学、ニューラルネットワーク技術、並びに数理技術に基づいている。即ち、決定木、BP、SOM、RBF、ファジー理論、遺伝的アルゴリズム、MBR、グラフ理論、ベイジアン理論、SVMなどを始め、種々の独自技術が用いられる。

リンク分析

1. 機能

1) アソシエーション分析

事象間の関連ルールを抽出する。

備考]多数のシンプルな関連ルールに特徴がある。

2) 時系列パターン分析

時間的な考慮に基づく事象の関連ルールを抽出する。

備考]時系列データの存在が前提になる。

3) 類似時系列パターン分析

事象間の挙動が類似パターンを採る領域を抽出する。

備考]nC2の全ての組合せを実行し、類似パターンと類似項目が検出される。

2. 手法

アソシエーション分析のアルゴリズムは幾つか提唱されているが、最近の主要な汎用ツールでは、Aprioriアルゴリズムが全面採用されている。

注1]時系列パターン分析のアルゴリズムはA prioriの類似手法である。

注2]Intelligent Minerの類似時系列パターン分析では、独自のアルゴリズミックな手法が採用されている。(IBM社は他社による使用の容認を宣言した)

クラスタリング

1. 機能

データの類似性に基づきサンプルを幾つかのグループに分割する。

2. 適性

1) 初期解析

全データ・セットを特性の類似する幾つかのサブデータ・セットへ分割し、各サブセットは各々次段における詳細解析で使用される。

備考]グループ化により、処理量を低減する目的で使用することもある。

2)本番解析

教師データが用意されない場合の問題解明に適用される。

備考]殆どのクラスタリング手法は教師なし学習である。

3. 手法

高精度の分割結果が期待できる非階層型クラスタリング法が採用される。

備考]階層型クラスタリングは所要計算時間は短いが、データマイニング・ツールでは採用を敬遠されている。

モデリング

1. 機能

履歴データや参照データを用いて予測モデルを生成しスコアリングを行う

1) 予測モデル
数値やカテゴリーの複数項目から、特定数値項目の予測モデルを生成する。

2) 判別モデル
数値やカテゴリーの複数項目から、特定カテゴリー項目の判別モデルを生成する。

3) 解析機能
モデル生成の過程で付帯的に解析機能も併せて提供される。

2. 適用

1) 予測モデル
①解析(特性の把握)、②予測モデルの生成、③予測(スコアリング)

2) 判別モデル
①解析(特性の把握)、②判別モデルの生成、③判別(スコアリング)

3. 手法

1) 予測モデル
①回帰木、②ニューラルネットワーク(BP)、③RBF、. . .

2) 判別モデル
①決定木、②ニューラルネットワーク(BP)、③RBF、④MBR、SVM、. . .

逸脱の検出

1. 機能

標準からの逸脱の検出(Deviation Detection)を行う。

2. 適用

不当事象の発見
例えば、不適切治療を検出する。

3. 手法

実用的な専用手法は未だ提唱されていない。
現状では、クラスタリング、リンク分析、或いはモデリングが代替適用されている。

4.適用事例

1) リンク分析　:異常な組み合せの関連パターンの検出事例が挙げられる。
2) クラスタリング :標準クラスターから掛け離れた異常の検出事例が挙げられる。
3) モデリング　:製品グレードを低下させてる操作条件の検出事例が挙げられる。

備考]大規模データの高速処理が可能なツールか否かにより、微細兆候の検出限界に格差が生じる。

適　用

データマイニングは広範な問題に適用できるが、大枠では解析と予測の2つの対象に区分される。

1. 解析
   実システムの機能について解析する。

   補足]探索的なデータの解析により、実システムの機能を解明する。

2. 予測
   実システムをモデル化し予測や判別を行う。

   補足]アルゴリズムによっては、モデルの生成過程で解析も併せて可能になる。

注]データの質と量、並びにアルゴリズムにより予測結果が左右される。

適合問題

1. 巨大データの全件処理
   業務で日々蓄積された巨大データを本格活用する。
   備考]サンプリング分析ではなく全件分析を想定する。

2. 定式化の困難な問題
   解析対象が複雑過ぎて定式化が困難な問題に適用できる。
   備考]定式化を必要とせず、必要量の適正データが有ればよい。

3. 複雑な問題
   単純な線形モデリングでは解明困難なケースで威力を発揮する。
   備考]高性能で強力な探索能力を持つデータマイニング・ツールが必要である。

4. 兆候の把握
   従来手法では不可能であった微細な兆候の検出が可能である。
   備考]何処まで深く把握できるかはツールとスキルに依存する。

5. 短期の解決
   短期解決が重視される問題に有効で現実的な意義は大きい。
   備考]現実の世界では時間との勝負が結果を左右するケースは多い。

長所と短所

−長所

−①予見困難な知識の発見を期待できる。
−②大量データに適性がある。
−③定式化困難な問題を処理できる。（定式化、仕様分析、開発は不要）
−④寧ろ複雑な問題に適性がある。(大量の良質データが必要)
−⑤異種データ・タイプの混在を容易に許容する。

−短所

①単純な問題には必ずしも適性とは云えない。
−②小量データに対して必ずしも適性とは云えない。
−③傾向把握には必ずしも適性とは云い難い。
−④ラフ・データに対して適性とは云い難い。（汎化性が劣る意味ではない）
−⑤解析結果が一意的に求まらないケースもある。（解が単一とは限らない）
注] 試行毎に結果が異なっても信頼できないと云った短絡的な解釈は避けたい。

統計解析との共通点と相違点

1. 共通点

①データ解析を行う。
②知識の獲得を試みる。
③実システムの解明や予測を試みる。
④種々のアプリケーションで活用される。

2. 相違点

1) 統計解析

①大局的な傾向把握
②仮説検証型アプローチ
注]統計解析の歴史において、プリンストン大学のテュキーは、従来の仮説検証型ではデータ解析に役に立たないと指摘、新しい探索的なデータ解析を提唱した。(1960年代)
補足]データは加工しない生データを用いる思想の下に、グラフ表示して視覚的な探索により、仮設の生成や予備解析に使用され出した。(データマイニングではデータ・チェックに活用)

2) データマイニング

①局所的な詳細解明
②知識発見型アプローチ

統計解析との大局的な相違点

1. 統計解析

1)能動的な収集データ
仮説を検証するため計画的にデータ収集する。

2)仮説ベース
解析者が経験や洞察に基づき予め仮説を立てる。

3)検定ベース
仮説が統計的に有意であることを収集データで検定する。

2. データマイニング

1)受動的な収集データ
通常業務で膨大なデータが収集済みと想定している。
注]データの蓄積を行う際に、データマイニング手法では活用可能なデータ項目でも、不要視され削除される場合も多く、意図的データ収集が望まれることもある。

2)探索ベース
仮説を必要とせずデータに埋もれている有用な知識を網羅的に探索する。

3)検証ベース
巨大データから抽出され情報は、データ自ずからが物語る知識である。
注]学習が妥当に終了したか否か、データ自体やデータ加工の過程に問題は無かったか、新規データによる検証が望まれる。

統計解析とデータマイニング

1. 統計解析

予め立てた1つの仮説をデータにより検証する。

例えば、スーパーマーケットでの顧客購買行動について、予め立てた仮説「パンを購入する顧客は牛乳を購入する」が有意か否か、サンプリング・データを用いて検定し、有意確率≦5%ならば、該当仮説は有意である。

2. データマイニング

着目事象間の全知識を自動的にデータから一括発見する。
例えば、スーパーマーケットでの「顧客の購買ルール」を発見せよと云う課題に対し、フル・データを用い指定範囲の全購買ルールが検出する。
ルール1:「ワサビを購入する顧客の85%はお刺身を購入する」
ルール2:「ビールを購入する顧客の55%はスルメを購入する」
ルール3:「パンを購入する顧客の40%は牛乳を購入する」
・・・
ルールn:「スルメを購入する顧客の20%はビールを購入する」
注]上記事例は、確信度20%以上のルールを検出する。

補足]統計解析の検定に対して、アソシエーションを対応比較したものである。

多変量解析とデータマイニング

1. 多変量解析

対象システム について線形モデリングを行う。

独立変量xと従属変数yの関係について、例えば、1つの特定関数を指定し、

$$y = a + bX$$

上式の最適な係数aとbをデータにより決定する。

2. データマイニング

対象システムについて非線形モデリングを行う。

独立変量xと従属変数yの関係について、例えば、ガウス関数を基底関数とし、

$$y = \sum_{i=1}^{n} wi * \exp\{(x-c)**2/2\sigma**2\}$$

上式の最適な重みw、基底関数の中心c、標準偏差σをデータにより決定する。

備考]前者は近似的な定式化（特定の1つの線形関数）であり、後者は厳密な定式化（任意の非線形関数）を実現する。

サマリー

- 発見型
  仮説の設定不要
- 意外性
  予期し難い知識の発見
- 完全制
  全ての有用な未知ルールを発見
- 自動的
  自動的な知識を検出
- 高速性
  巨大データの許容
- 網羅性
  総当たり的な探索
- データ・タイプ
  数値とカテゴリー値の混在
- 非線形モデル
  精緻な分析
- 定式化
  自動的な定式化
- 客観性
  解析者の主観に依存しない
- 容易性
  解析手法の専門知識は必須でない
- 明解性
  結果の解釈が容易
- 応答性
  結果の応答が迅速である
- 拡張性
  並列処理への対応

備考] Intelligent Minerはデータマイニングが目標とする上記特性の全てを実現している。

データマイニングの理論

各種機能を実現する基礎技術として、多数の技法が提唱されている。
特に、数理的な評価の高い汎用データマイニング・ツールIntelligent Minerに採用の基礎技術について説明する。

27

リンク分析

ー有限事象間の関連性に関して分析ー

○アソシエーション

○時系列パターン分析

○類似時系列パターン分析

汎用データマイニング・ツールIntelligent Minerに組み込まれている機能で、最も重要なアソシエーションに絞り言及する。
ここに、アソシエーションに対しては、幾つか手法が提唱されていたが、今日ではIBMが提唱したAprioriアルゴリズムが広く採用されている。

28

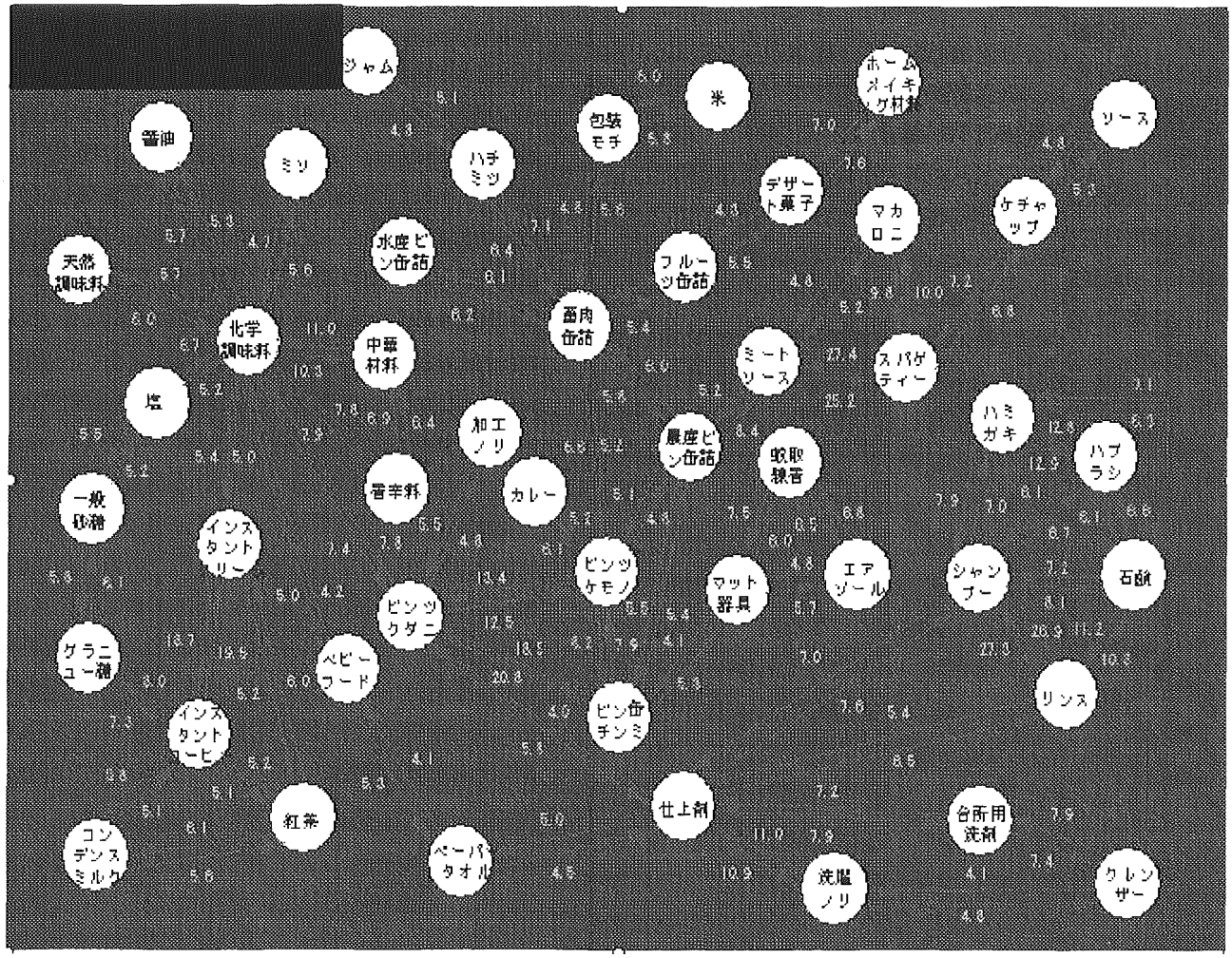
アソシエーション

相関関係ルールは2つのステップで抽出される。

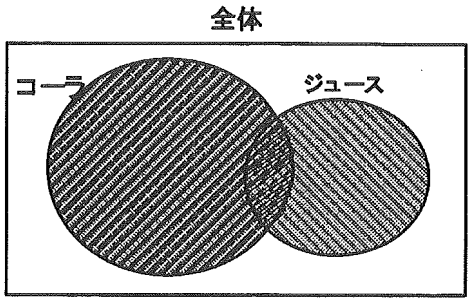
ステップ1： 最小支持度を満足するアイテム集合を全て取り出す。
ここに、アイテム集合はラージアイテムと呼ばれる。

ステップ2： ラージアイテム集合から最小確信度を満たす相関ルールを導き出す。

備考]データマイニングの特徴を最も良く表わしている代表的な機能の1つであるが、相関関係分析と翻訳されている。
多変量解析で共通処理される相関係数に基く相関分析と比べ、表現的には極めて紛らわしいが、相関分析は各変数が取る数値データから相互の関係を評価するのに対し、相関関係分析は特定の事象間の関連性を評価し、事象間に存在する相関ルールを抽出する。
種々のアルゴリズムが提唱されているが、現在はAprioriアルゴリズム(IBM／アルマデン基礎研究所で提唱し特許取得済み)が広く採用されている。

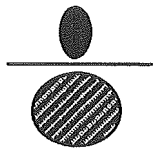


関連性の指標

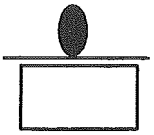


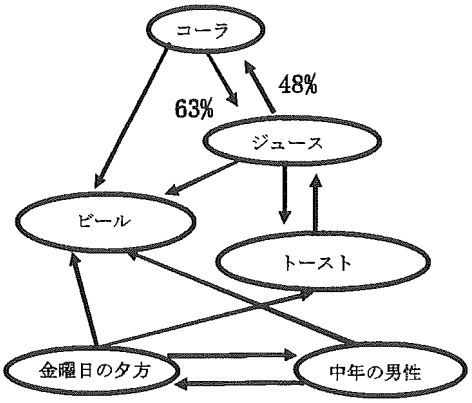
ルール：コーラ(ボディ)→ ジュース(ヘッド)

Confidence
確信度(コーラ→ジュース) ＝

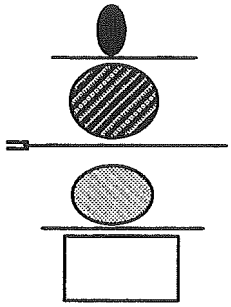


Support
サポート値(コーラ＋ジュース)＝





Lift
リフト値(コーラ→ジュース) ＝



注] リフト値により始めてルールの有用性を確認できる。

アソシエーションのサンプル・ケース

5人の顧客の買物状況

保理	前田	數田	横山	敷根
A B D	B C E	A B C D E	B D	A D E

ショッピング・バスケットの中身は、. . .

入力パラメータ

Support
Confidence

AとBの関係

Support	Confidence	
A&B 2 / 全顧客 5 = 0.4	A&B 2 / A 3 = 0.67	A&B 2 / B 4 = 0.5

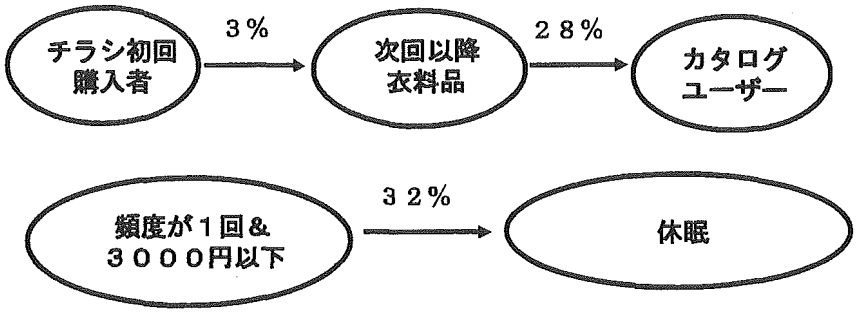
AとCの関係

Support	Confidence	
A&C 1 / 全顧客 5 = 0.2	A&C 1 / A 3 = 0.33	A&C 1 / C 2 = 0.5

時系列パターン分析

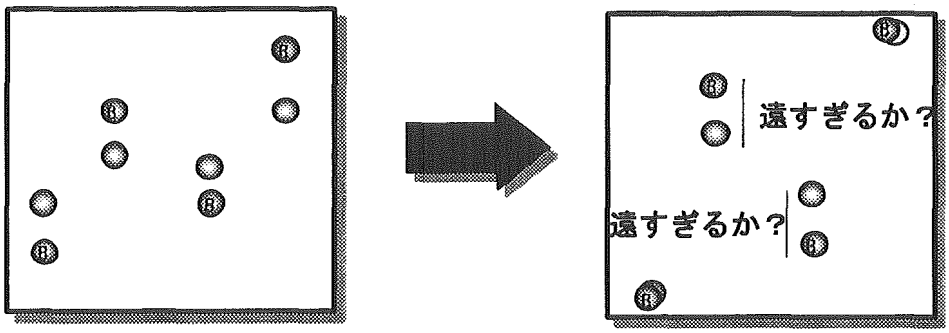
時系列的な前後関係に着目したアソシエーション

同一データ項目が時系列的に取るパターンを見つける。
与えられたトランザクションのデータベースにおいて、特定の期間で、
１つのトランザクションの中で、どの項目どうしがお互いに関連して
いるかを発見する。



類似時系列パターン分析

赤の挙動と青の挙動に関する類似度を調べることを目的とし、
窓の中で最大値と最小値が同じになるようスケーリングする。

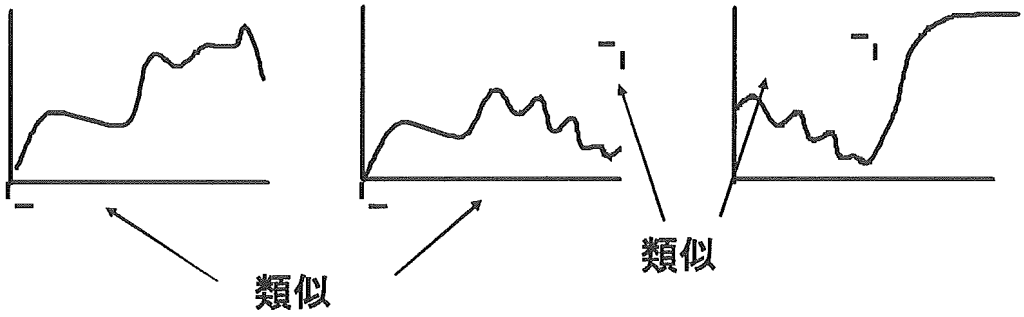


どの点の組合わせも「遠すぎ」ないなら窓は一致と見なす。

備考]現在のところ、類似時系列パターン分析の機能を提供する汎用データクイニング・
ツールは稀である。

類似度の評価指標

時系列パターンから類似したパターン・セットを発見する。



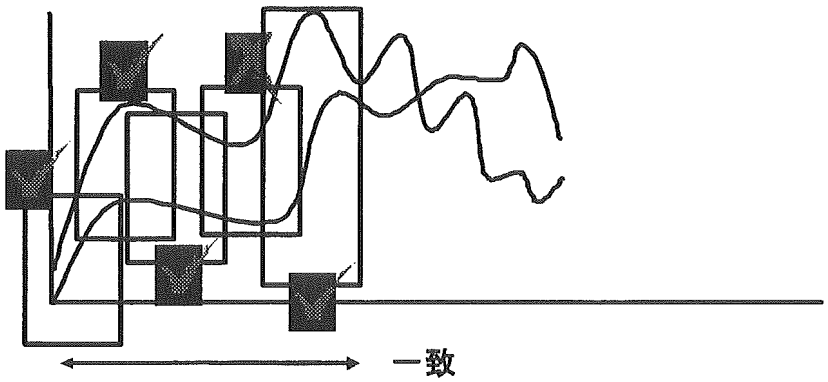
４つの評価指標

Epsilon	:類似判定（相対許容誤差）
Gap	:許容範囲外ながら非類似としない最大幅（連続点の数）
Window Size	:類似比較の対象とする幅（ウィンドウ幅）
Matching Length	:非類似と認定する最小幅

注］全体の長さで一致部分の長さを割った値：比率

比較領域の窓ずらし

類似した時系列を発見するためには、
一連の類似した窓を発見する



パラメータとして連続して一致する窓の数を指定する。
別のパラメータとしてノイズの許容度を指定し、
若干の不一致のために情報を見落とす事がないようにする。

クラスタリング

①K-means法（古典的な非階層型クラスター分割法）

②コホーネン・フューチャー・マップ法

③デモグラフィック・クラスタリング法

備考］データマイニングに用いられるクラスタリング手法として、Intelligent Minerに採用されている手法について述べる。最初に、クラスタリングの基本的な考え方を習得するため、基本的で簡単なK-means法について説明する。
注］K－means法は性能面などで劣り、データマイニングには殆ど採用されない。

K-Ｍｅａｎｓ法

●古典的には代表的な非階層型のクラスタリング手法である。
●ユーザーがクラスターの個数Ｋを予め与える。
●各クラスターの中心点の初期値は自動決定される。
●各クラスターの中心点と構成要素は逐次的に更新される。

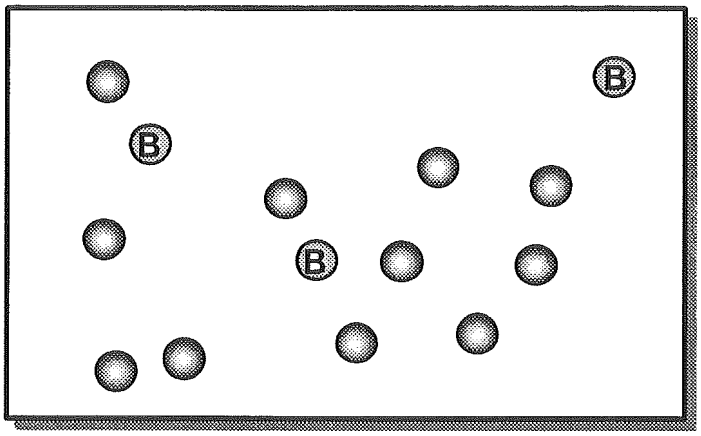
補足］大規模問題の場合、性能の劣るＫ－ｍｅａｎｓ法では、主成分分析や因子分析を用いて、予め変数の数を低減して適用する。

備考］K－Means法ではクラスター数を指定するが、クラスタリング手法の中には、クラスター数を指定せず、最適数が自動決定される手法もある。

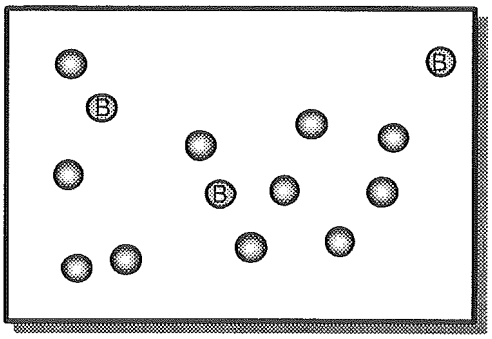
説明用サンプル

●データ
桃色の点（１１個）はサンプルで、２変数の座標値を表す。
青は、初期設定した各クラスター（Ｋ＝３個）の中心点を表す。

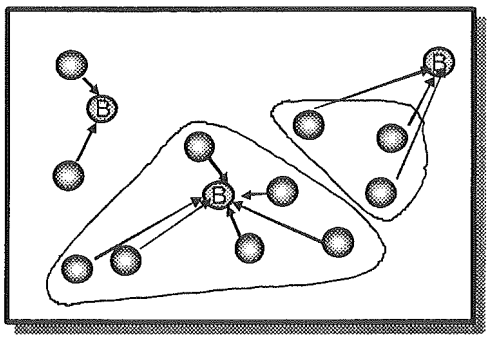
●アルゴリズム（繰返し：②と③）
①最初にクラスター数を決め、各クラスターの中心点（初期値）を設定する。
②各サンプルについて、最も近いクラスターの中心点を探し所属要素とする。
③各クラスターの所属要素を用いて中心点を求める。



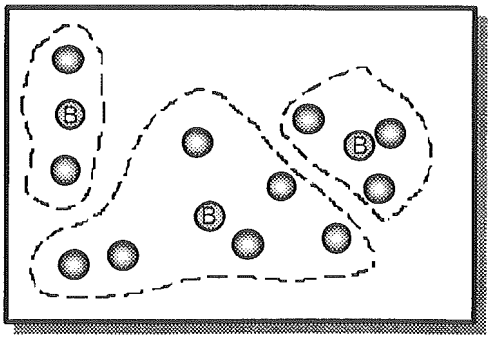
Ｋ－ｍｅａｎｓ法の処理手順(1)



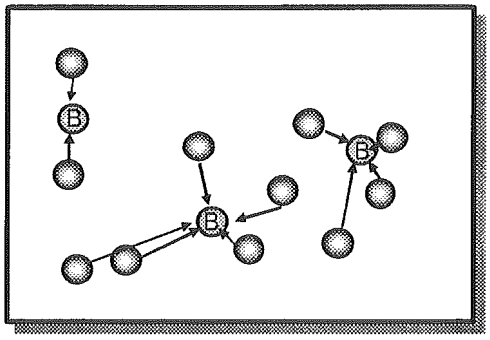
(1)サンプルとクラスター(初期点)



(2)クラスターの要素を決定

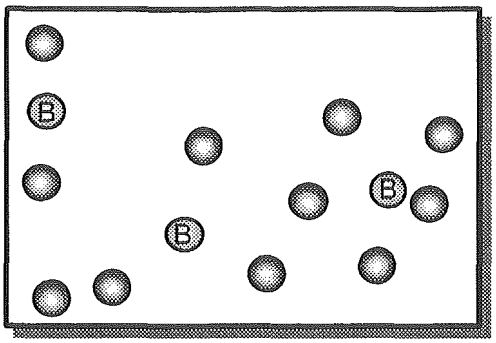


(3)クラスターの中心点を算出

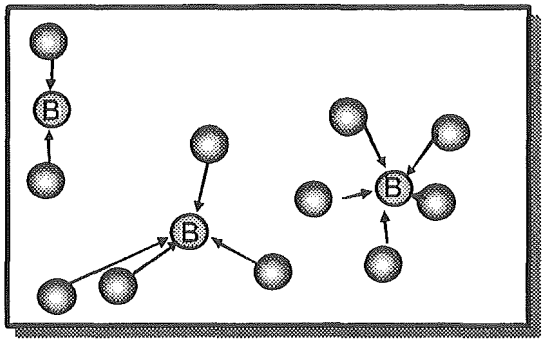


(4)クラスターの新要素を決定

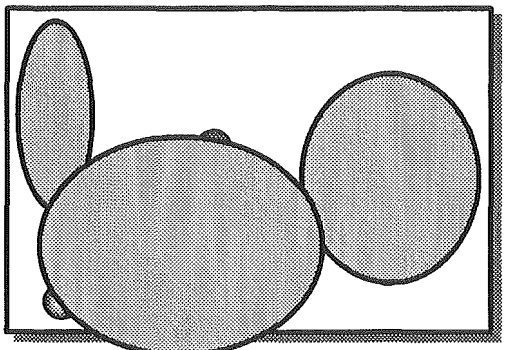
K－ｍｅａｎｓ法の処理手順(2)



(5)クラスターの中心点を更新



(6)クラスターの新要素を更新



(7)クラスターの中心点を最終更新

クラスターの中心点が全て前回とほぼ同じになった時点で繰返しを終了する。

K－Means法の欠点

●カテゴリー・データの扱いにはタイプ制約がある。
- 入力データ・タイプとして
  - 数値　　：　5.4, 4000, -11.434, . . .
  - カテゴリー：　"大きい", "白", . . .

注] カテゴリーは許容されない。

●類似度や関連性などを含め、距離尺度は必ずしも適切でない。

●ユーザーがクラスターの個数Kを与えなければならない。

●局所的なクラスターを求めるのは困難である。

●大規模問題では計算所要時間が膨大になる。

コホーネン・フューチャー・マップ法

○競合作用に基く学習方式を採る教師なし学習である。

○コホーネンの自己組織化特徴マップは、競合学習型のニューラル・ネットワークである。

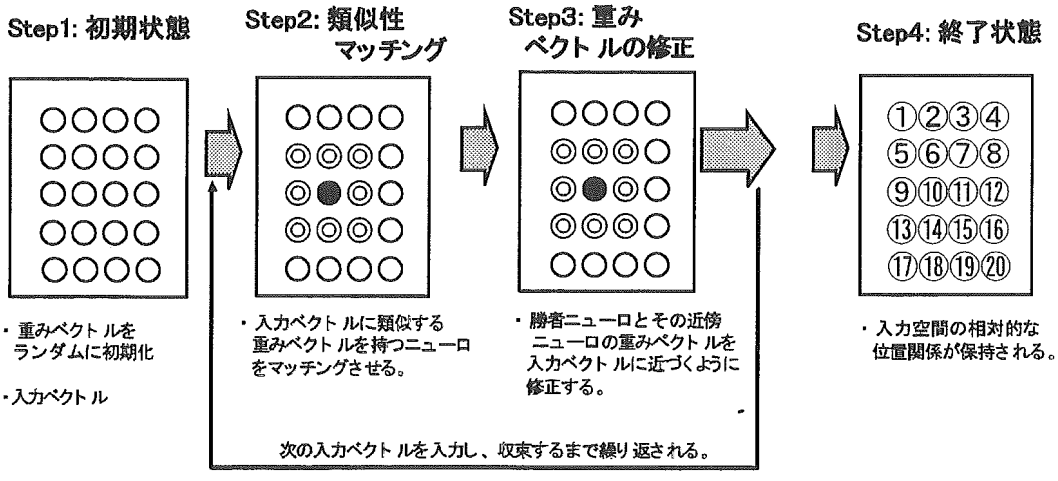
○多次元の入力データ空間における相互関係をニューロンの低次元空間(1次元、或いは2次元的な繋がりマップ)に写像することができ、有用な情報を容易に取り出せる。

○コホーネンの学習則では、1つの入力データに対して多数のニューロンが反応するが、最も強い出力を出しているニューロンが勝ち残り、学習は勝者に対してのみ行われて結合重みが修正される競合学習の代表モデルである。(原理:winner-take-all)
注]実際には、勝者ニューロンkの近傍ニューロンも併せて修正する。

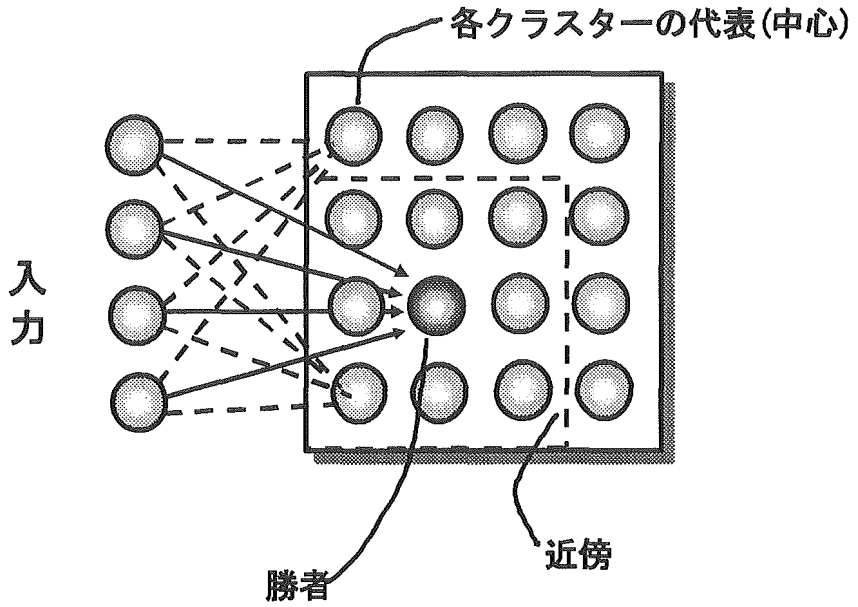
備考1] コホーネン・クラスタリング法、ニューラル・クラスタリング法、SOMとも呼ばれる。
備考2] 測定尺度に、ユクリッド距離を用いるが、カテゴリー・データも許容する。

コホーネンの学習法

入力空間におけるデータの関係を出力空間で極力保持できる学習を行う学習法でである。すなわち、実データに含まれる特徴をネットワークが自己組織的に表すという特徴を持っている。



コホーネン特徴マップ



入力データ(入力ベクトル)にたいして、最も近いニューロンが勝利ニューロン(勝者)として決定され、勝利ニューロンは入力ベクトルに近づくように重みが修正される。また、同時に近傍のニューロンの重みも調整される。

コホーネンの学習則(1)

○ニューロンの内部ポテンシャルPは次式で計算される。

$Pij = 1/(D(wi, xj))$

ここに、wi=(wi1, wi2, ・・・, win)は各ニューロンの結合重みベクトル、xj=(xj1, xj2, ・・・, xjn)は入力データベクトル、nは入力データの次元数である。

○両ベクトルの相違度を測る測度関数Dは、一般にユークリッド距離が用いられる。

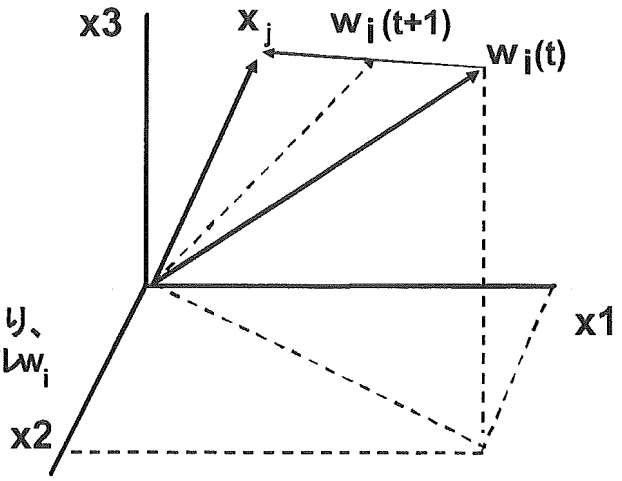
$(D(wi, xj) = |wi-xj|$

○最大の内部ポテンシャルを示すニューロンを見付ける。
これは入力データに最も類似した結合重みのニューロンを検出することを意味している。

コホーネンの学習則(2)

ー勝者ニューロンと近傍ニューロンの更新ー

ニューロンkに加え、
全ての近傍 i について、
次式のコホーネン学習則により、
ニューロンの結合重みベクトルw_i
は修正される。



$$w_i(t+1) = w_i(t) + a(t) \| x_j - w_i(t) \| \qquad i \in N(t)$$

a (t)： 学習回数 t における学習係数 （1 ≧ a ＞ 0）

コホーネンの学習則(3)

○入力データ空間において、入力データが多く存在する領域では、ニューロンは勝ち残る可能性が高く、結合重みベクトルを幾度も引き寄せて、結合重みベクトルが密集して来る。

○結合重みベクトルによる入力データ空間の分割は、結合重みベクトルの密集領域では部分空間が小さくなり、希薄領域では大きくなる。

○各部分空間における入力データ数の一様化が進み、入力データの分布に適応したベクトル量子化(分割)により、分布を無視した一様量子化に比べて良い結果が期待できる。
備考]データ分布に極端な偏りがあれば、良好な結果が得られない可能性は残る。

○多次元入力データ間の類似関係が、ニューロン・マップ(一次元、或いは二次元)上に自己組織的な形成が進み、入力データの分布状態を上手く反映したマップが作られ、入力データ空間の相互関係を保存した写像が行われる。

コホーネン特徴マップ法の特徴

●処理概要：
- -測度としてユークリッド距離を採用する。
- -近傍関数は学習と共に縮小する。
- -学習係数は最初は大きく始め次第に小さく設定する。

●長所：
- -高質のクラスタリング結果が期待できる。
- -クラスターが空間的に順序付けられる。

●短所：
- -計算量とメモリー所要量が多い。
- -クラスター数はユーザーが与えなければならない。

備考]ここでのマップは写像と地図の2つの意味を兼ねている。

デモグラフィック・クラスタリング法

●カテゴリー値と連続値の混在を許容する。

●クラスターの最適個数は自動的に決定される。

●データの一致度を投票して評価する手法である。

●評価基準：
- -同じクラスターの属性値は類似している。
- -違うクラスターの属性値は類似していない。

●簡単そうであるが、実は一筋縄ではいかない。

備考]数値とカテゴリーの両方について、類似性を評価できる点に大きな意義を持つコンドルセの評価基準を採用している。
補足]Condorcetはフランスの数学者で、データの一致度を投票して評価する手法を提唱した。

類似度の定義

●カテゴリー値については、同じ値の場合にのみ類似とする。
●数値については、どのくらい離れていているものを同じとするかをアルゴリズムが決める。

	名前	年齢	学歴	地域	結婚
	香苗	38	中卒	東京	既婚
	小太郎	42	博士	神奈川	既婚
類似度：	無意味	類似	相違	相違	類似

デモグラフィック・クラスタリング法の処理手順

●クラスター無しの状態から始める。
●データベースの各レコードについて
  -適切なクラスターに割振る。
  -適切なクラスターがない場合は新クラスターを作る。
●この後、再度データベースの各レコードについて
  -他の全てのクラスターに割振ってみる。
  -それ自身の新しいクラスターを割振ってみる。
  -既に属しているクラスターに残してみる。
  -これらの３パターンのうち最適なケースを選ぶ。