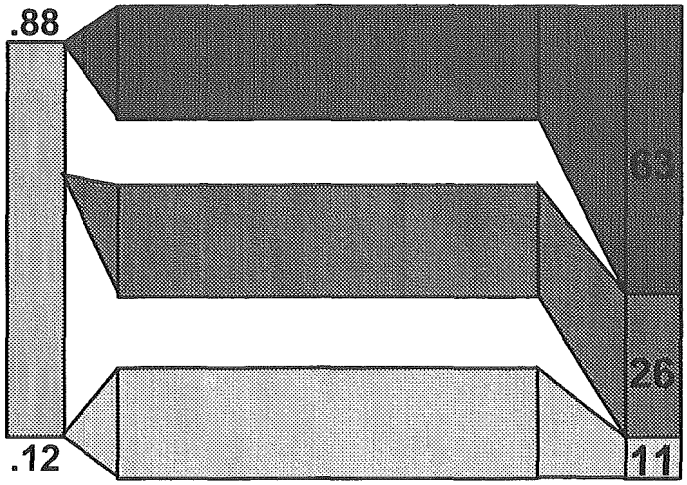
ＲＢＦの結果表示

ーマイニング結果の可視化による分析支援ー

各々の帯は
一つのクラスター
を表している

左側に指し示して
いる位置は各
セグメントの
予測値の平均である



各帯は異種集団に
対応した独立のモデル
を表現しており、
全てを1つのモデルで表現する、
伝統的なアプローチの弱点を回避している。

備考］先のクラスタリングとＲＢＦでは、上記のような帯の表示形式
が採られている。（この表示情報は結果分析を強力に支援する）

ＲＢＦの特長

●何故、ＲＢＦを使うのか？
①複数モードのモデリングを可能にする。(異なった特性を持つ
複数のグループからなるデータに対して各々モデルが作れる)
例えば、スポーツカーに乗る二種類の購入層(若者, お金持ち)
②顧客セグメンデーションを行い、各セグメントの特徴を解析する。
例えば、セグメントは離反の傾向でランク付けできる。
●RBFは、ニューラル・ネットワークに類似しているが、逆伝搬手法(BP法)
より格段に高速処理される。
●ＲＢＦは実用的観点から、さまざまな応用方法がある。

備考]大枠の解析には決定木を用い、詳細な解析はRBFで補完する。

ＲＢＦの特長

●何故、ＲＢＦを使うのか？
①複数モードのモデリングを可能にする。(異なった特性を持つ複数のグループからなるデータに対して各々モデルが作れる)
例えば、スポーツカーに乗る二種類の購入層(若者, お金持ち)
②顧客セグメンテーションを行い、各セグメントの特徴を解析する。
例えば、セグメントは離反の傾向でランク付けできる。
●RBFは、ニューラル・ネットワークに類似しているが、逆伝搬手法(BP法)より格段に高速処理される。
●ＲＢＦは実用的観点から、さまざまな応用方法がある。

備考]一般には、大枠の解析に決定木を用い、詳細な解析にはRBFを適用する。

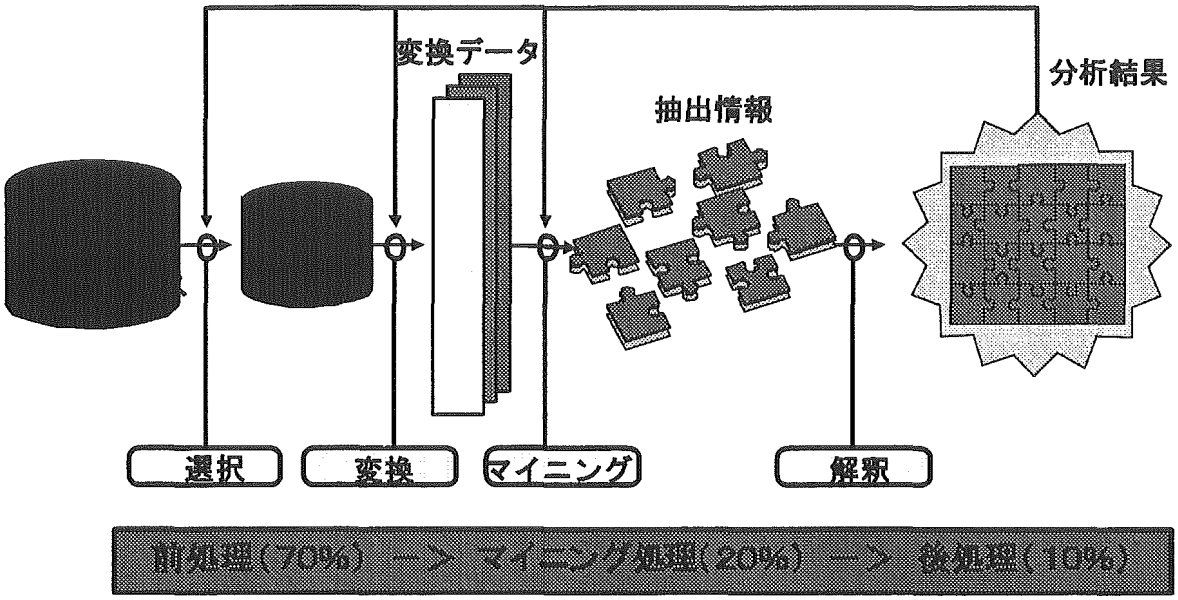
パート2 ： 応用編(1)

データマイニングの実践

データマイニングにより効果的に良い成果を得るには、データマイニング技術に精通した専門家と該当アプリケーションに精通した専門家による緊密な連携が理想である。
統計の専門家が統計カルチャーに準拠して、仮説に基いたデータマイニングを行えば、自ずからが制約を設けることに成り、本来なら求まるものが求まらなくなる。
しかし、統計の専門家がデータマイニングのカルチャーに基ければ鬼に金棒である。
また、データマイニングによる分析ワークは、将に手工業で各問題ごとに独自の解析メソドロジーが望まれる。

データマイニング・プロセス

データマイニングとは、大規模データベースから理解可能で実行可能な未知の有用な情報を自動的に抽出するプロセスのことで重大な意思決定に使用される。



データの種類

1. マスター・データ
   業務上の補助データやアンケートで得られるデータなどである。
   1) 患者データ
      備考] デモグラフィック・データやサイコグラフィック・データも含まれる。
   2) 医師データ
   3) 機材データ
   4) 薬剤データ
2. 品目データ
   1) 在庫データ
   2) 納品データ
   3) 出庫データ
3. 治療データ
   1) 検査データ
   2) 処置データ
4. 調査データ
   1) 検診データ
   2) アンケート・データ
      備考] 誤解や客観性が損なわれる可能性から、特に質問文の表現には注意を要する。
5. 会計データ

備考1] デモグラフィック・データとは、性別、年齢、学歴、職業、所得、婚歴などが挙げられる。
備考2] サイコグラフィック・データとは、性格、地位、趣味、ライフスタイル、嗜好などが挙げられる。

データの準備

1. データの選択
   分析目的に応じて必要な変数を選択する。
   備考]内情報だけでなく不足なら、外部情報による補完も考慮する。

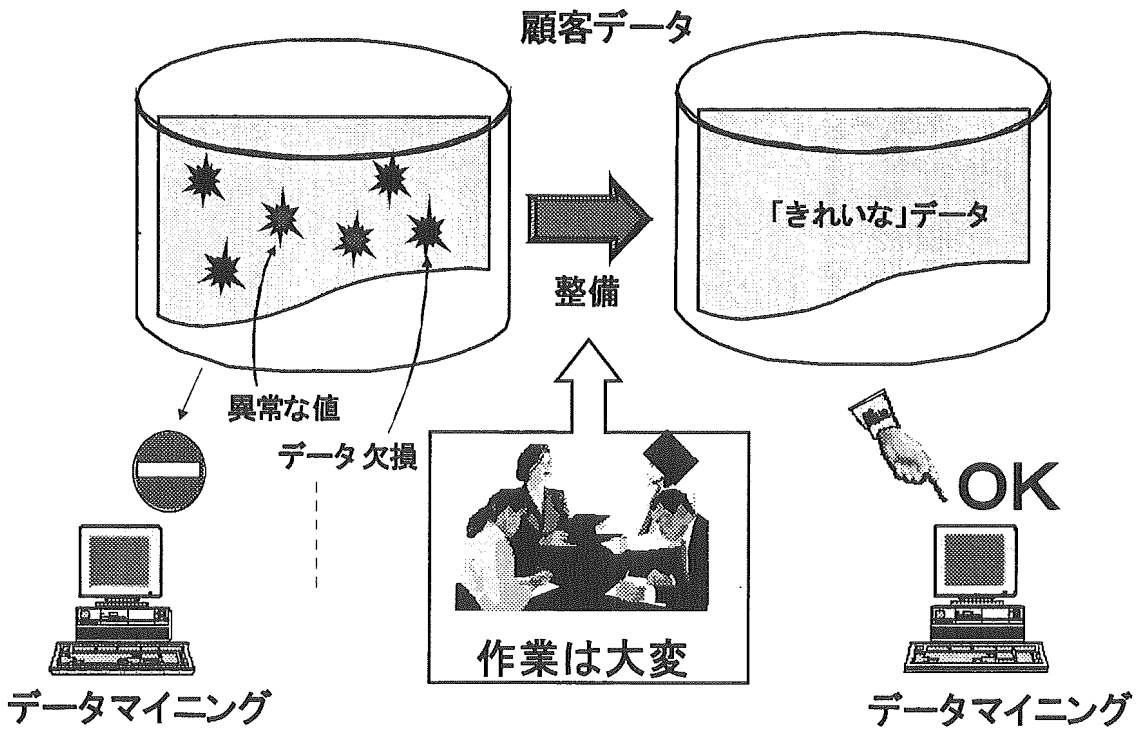
2. データの集計
   1) 時間・・・・・薬剤の投与期間をどういう単位で集計するか。
   2) 空間・・・・・薬剤の投与量をどういう範囲で集計するか。
   3) 対象・・・・・薬剤の種類をどういう区分で集計するか。

3. データの洗浄
   1) 欠損値・・・・・入院患者さんが体重を測定し忘れた。
   2) 外れ値・・・・・入院患者さんが美人看護士さんに興奮し異常体温を示していた。
   3) 不当値・・・・・入院患者さんの体温が37度なのに37．1と記入されていた。

4. データの変換
   1) カテゴリー化
   2) 数値化

5. データの補強
   マイニング処理する上で必要に応じて変数を追加する。

データマイニングで不可欠なデータの整備



特記]不良データでは、有用な結果が得られぬのみか、重要な意思決定を誤らせる。

留意事項

1．データの収集
　1）汚染の除去　　：汚染データの混入があれば、妥当な結果は得られない。
　2）バイアスの除去：データにバイアスを伴えば、妥当な結果は得られない。

2．項目の選択
　選択項目に不足があれば、期待の結果が得られなく成る。

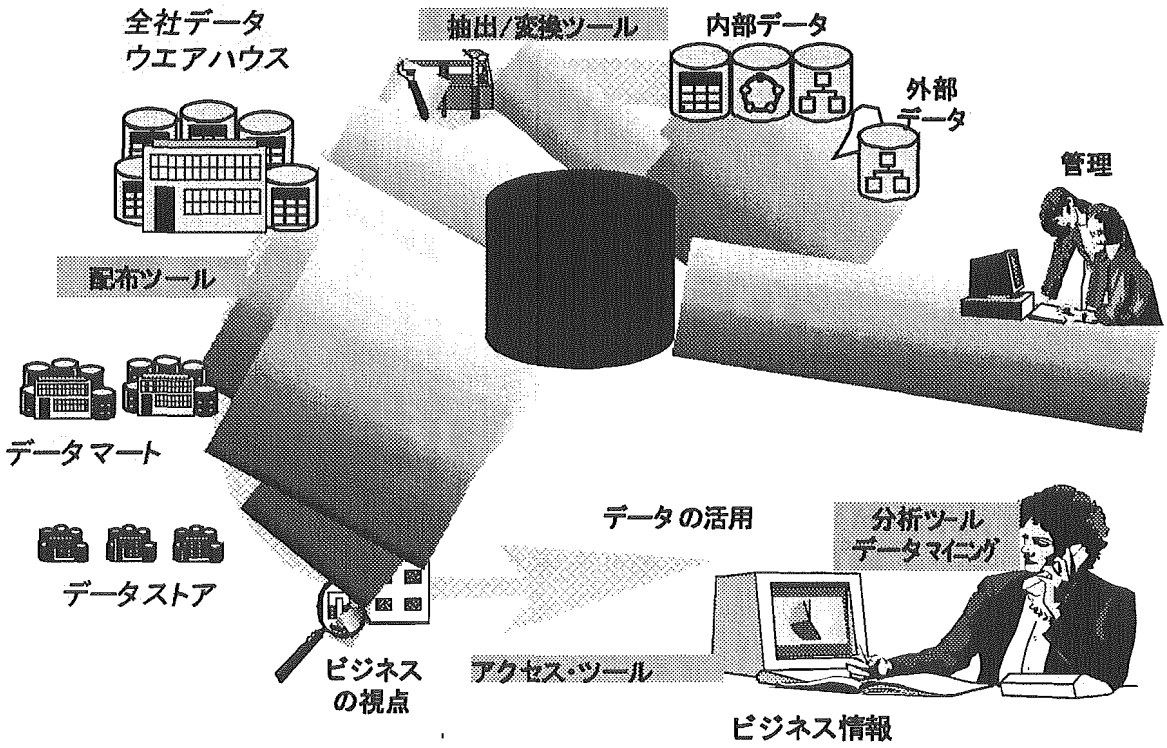
3．データ量
　必要データ量を欠けば、結果は保証されない。

4．データの整備
　データ・クレンジングを怠れば、不当な結果へ繋がる。

5．データ加工
　1）データ表現に不備があれば、妥当な結果は得られない。
　2）データに間違いがあれば、妥当な結果は得られない。

6．不可解な結果
　結果が期待に反した時、背後に意外な知識が隠れてないかの発想を持ちたい。
　注］意外な知識の発見にこそ真骨頂があり、新たな知識の発見への入口かも知れない。

データウェアハウス構築が理想



備考］DWHは業務系ＤＢと異なり、分析用としてＤＢ構築され、データストアの構築が理想である。

手法の選択

同じデータと同じツールでも結果に格差を生じる。仮え同じ結果が得られるにしても、採用手法は簡単なほどベターで、活用スキルは軽視できず、特に手法の選択には留意したい。
実際、手法には、各々異なる特性があるから、基礎技術の特性は習得して置くことが望まれる。

1. データ
   データのタイプ、クオリティ、ボリュームなどにより候補手法が絞られる。

2. 目的
   目的は解析なのか予測なのか、或いは両方なのかで手法が絞られる。

3. テーマ
   テーマにより、手法の選択や組み合わせ、並びに適用手順が決まる。

4. ターゲット
   目標とする結果の粒度により手法が決まる。

備考]仮え同じ手法であっても、インプリメンテーションにより、探索の性能や能力に格差が見られるから注意を要する。

統計解析とデータマイニングの補完関係

1. 変数の代表処置

1) 数値変数:相関が強い。
   相関分析を行い、相関係数が大きい関係の変数がある場合、何れか1つの変数を代表として採用する。

2) カテゴリ—変数:独立性が強い。
   カイ二乗検定を行い、カイ二乗値が小さい関係の変数がある場合、何れか1つの変数を代表として採用する。

2. 変数の代替処置と欠損値問題の回避

次の何れかを適用して、代替変数を求めて使用する。
1) 主成分分析
2) 因子分析
注]Intelligent Minerのクラスタリングと RBFでは、オリジナル変数を補助変数としての設定により、オリジナル変数と代替変数の関係を把握できる。

注]サンプル数と比較して、変数の数が極端に多い場合、変数の低減を図るに際して、例えば平均の差に基くゴセットのT値などでは、局所的な違いに基く影響を解明するデータマイニング・カルチャーから逸れて、得られる結果も得られなくなる。

評価と検証

1．モデルの評価

手法やパラメータを種々変えて幾つかの候補モデルを生成し、期待効果を比較して最善モデルを決定する。

補足］生成モデルを各種評価基準に基づいて査定する。

2．施策案の評価

最善モデルと新データを用い、種々パラメータを変更してシミュレートし、得られた候補結果を比較して最善施策を選択する。

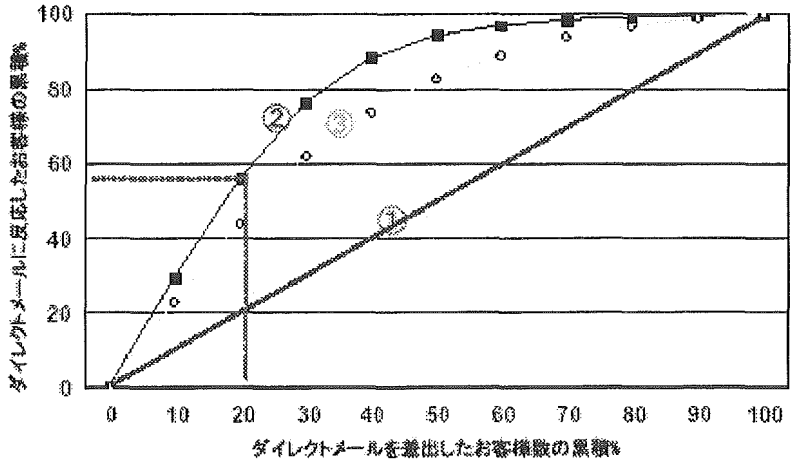
補足］該当アプリケーションに精通した専門家による、実務特性に基づく評価は必須である。

3．実施と検証

最善施策を実地に適用して、適用結果で最終的な評価が決まる。

補足］適用結果のデータを収集し、逐次改善策を採ることもある。

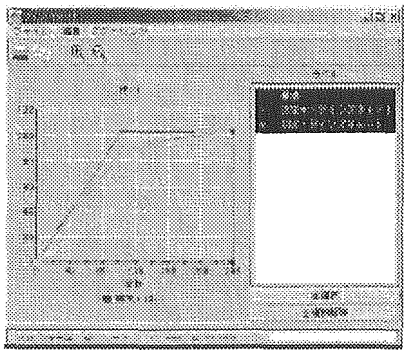
ケインズチャートによる効果の評価



①：マイニングの予測モデルを使わないで、ランダムにダイレクトメールを差し出すと、ほぼ差出数に比例して反応があります。

②：マイニングの予測モデルにより、ダイレクトメールに対してレスポンスが高い確率で期待されるお客様を予測。これにより、20%のお客様にダイレクトメールを出すことにより、潜在顧客の60%近くをカバーするとシミュレーションできます。

③：②と別の項目・手法等を使って予測。このモデルもある程度の正確な予測はしていますが、青い線のモデルよりは精度が劣ります。



マイニングを失敗しない十ヶ条

1. 選択データ項目はテーマに適合しているか。
2. データ項目が不足なら補強できないか。
3. データ量が不足なら補充できないか。
4. 不当なデータは混在してないか。
5. 不適切なデータ表現を採ってないか。
6. データ変換は妥当に行われているか。
7. データ項目の絞込みは適正に行われているか。
8. 採用アルゴリズムはテーマやデータに対して適性か。
9. 解析メソドロジーに改善の余地はないか。
10. データ収集に際して問題はなかったか。

備考]ゴミからはゴミしか出ないと云う問題とは別に、仮説を設けたアプローチでは、発見できるものを発見できなくするが、良質なデータが必要量提供されてない場合、最低限の仮説を想定することも仕方ない。

パート2 ： 応用編(2)

データマイニングの事例

一般分野 （添付資料省略）
①競馬の着順予測
②品質管理
③個々人の購入額予測
④RFMモデルとの比較事例
⑤成長顧客と衰退顧客の識別
⑥季節変動分析
医療分野 （添付資料省略）
①肝疾患患者の予後予測
②SNP解析による薬物感受性分析

各種分野における適用事例

ーデータマイニングー

業種	共通アプリケーション	業種特化アプリケーション
流通	ターゲットマーケティング	併売分析、棚割分析、購入パターン分析、資金配分、出店計画、...
金融	リスク分析	ポートフォリオ分析、株価予測、金利予測、顧客格付、与信分析、...
保険	宣伝効果分析	個客リスク予測、最適保険料決定、利益分析、保険支払管理、...
通信	販促効果分析	離反分析、脱落分析、利益率分析、不正使用検出、競合情報分析、...
運輸	在庫管理	価格決定、マーケティング応答分析、イールド分析、不正防止、...
医療	研究結果分析	疾病診断、疾患要因分析、処方箋分析、発病名予測、遺伝子分析、...
製造	需要予測	不良要因分析、工程分析、クレーム特性分析、受注分析、新製品開発、...

適用分野のサンプル

ターゲット・マーケティング（家電量販店）

購入した商品の特性や組み合わせなどの購買行動によって、顧客を分類しました。例えば、「新しいもの好き」「機能重視」「オーディオが趣味」などといったライフ・スタイルに基づいた顧客層を見つけ、それぞれの層の価値観やニーズを類推して、マーケティング施策に結び付けています。

ダイレクトメール反応予測（通販会社）

予測手法を用いて、膨大な顧客データベースからダイレクトメールに反応する可能性の高い顧客を的確に予測して、ヒット率の大幅な向上を実現しています。従来、RFM分析だけでは十分に把握しきれていなかった、隠れた優良顧客を見出すことができるようになりました。

優良顧客の離反防止（通信会社）

通信業界等では、優良顧客の離反防止が大きな関心事のひとつとなっています。顧客の離反の兆候となる事象をデータマイニングにより的確に把握することができます。離反の兆候のある優良顧客を発見し、その特性を把握し、離反を防止するための利便性・満足度向上施策の策定につなげています。

インターネットマーケティング（ネット通販）

Webログにより、ある顧客がどういったページを見ているのかを把握することが出来、それを分析に取り入れることでインターネットならではのCRMを実現できます。従来の顧客属性や購買データを補強することで、より一層顧客を理解することができます。

需要予測（家電メーカー）

需要の変動が激しい季節性の家電商品に対して、需要の変動要因を導き出すことで、将来の需要予測精度を高め、市場環境の変化に即応することができます。

品質管理（半導体製造）

半導体製造工程において、製造工程履歴データや不良品検出データなどから、品質悪化の原因を発見し、歩留まりの改善と生産効率の向上をもたらします。

データマイニング・ツール

データマイニング・ツールには種々の異なるタイプがある。

汎用データマイニング・ツール Intelligent Miner

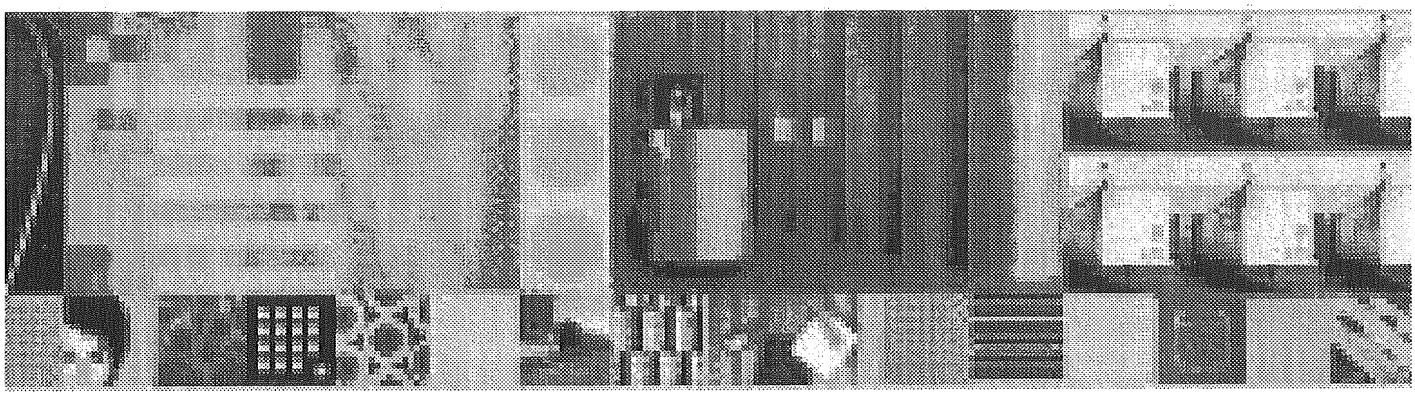
Intelligent Minerは、複数のIBM研究所で開発された単独機能ツールを統合してパッケージングし、外販用として提供している汎用マルチタスク・ツールで、初版の提供から約10年を経過し、現行版は種々改良強化を経たVer8．1である。
一方、汎用データマイニング・ツールの比較評価について、種々の報告書が出されているが、Intelligent Minerについて深い査定を極めた報告書を見掛けたことがない。

データマイニングの真の威力発揮へ ！

IBM

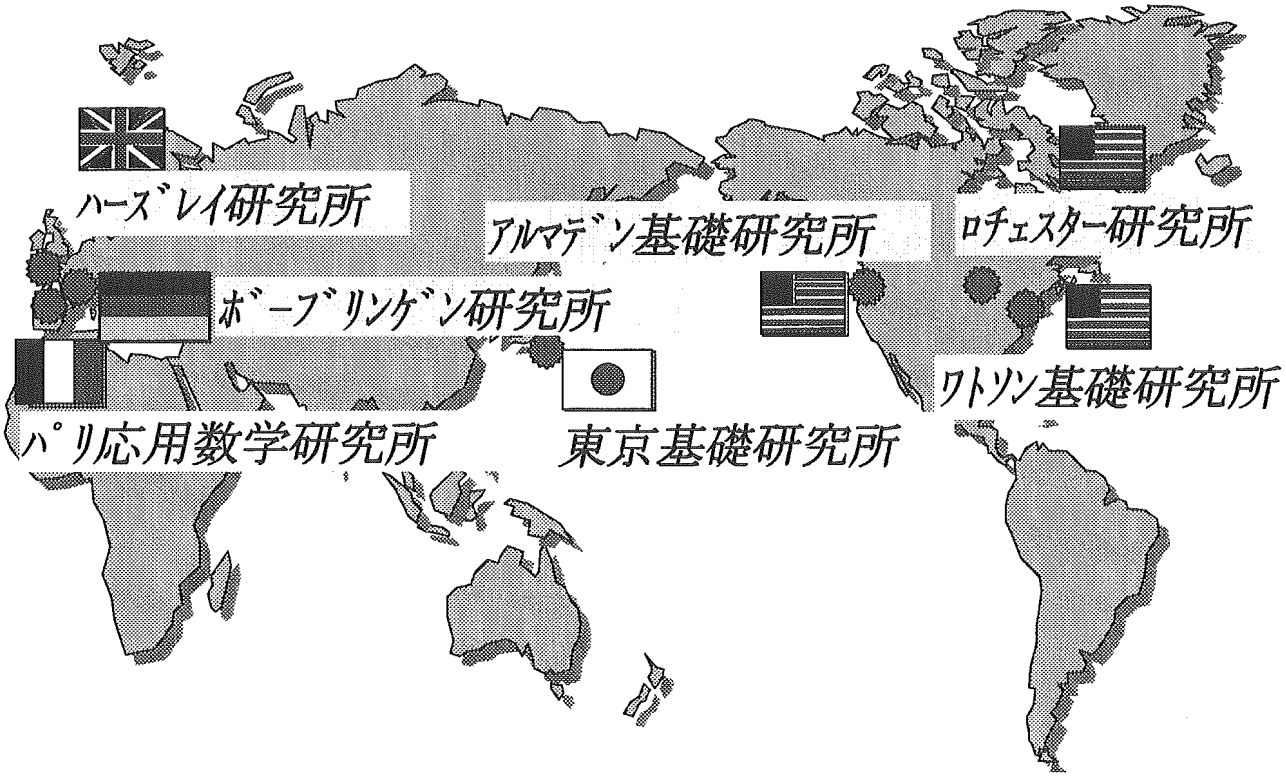
IBM Intelligent Miner for Data

Version 8.1



本格的なデータマイニングに適性がある

ＩＢＭ研 究 所



備考1]データマイニングに関与しているIBM研究所のリストアップです。
備考2]IMの開発に関与する研究所はダーク・グリーンで表示されています。

Intelligent Minerの概要

1. 豊富なデータマイニング機能

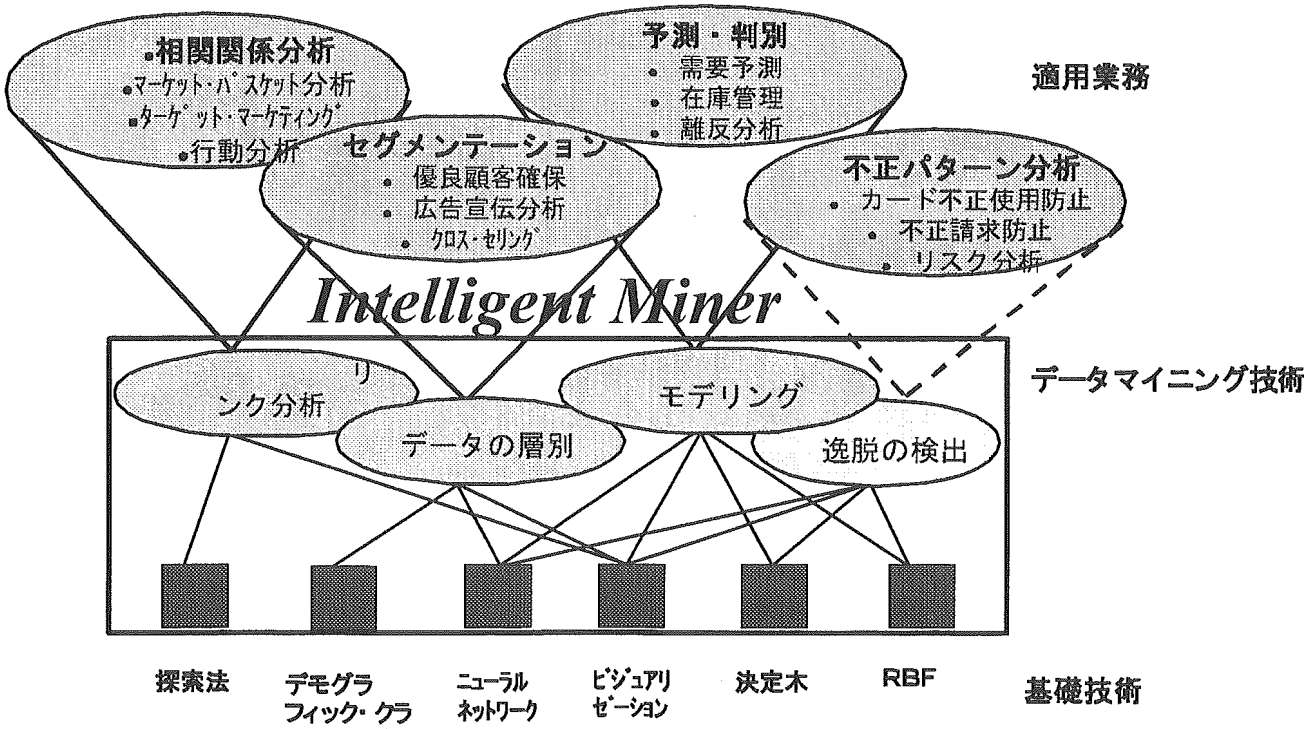
- クラスタリング(Clustering)
- クラス判別(Classification)
- 数値予測(Value Prediction)
- アソシエーション(Associations)
- 時系列パターン分析(Sequential Patterns)
- 類似時系列パターン分析(Similar Time Sequences)

備考]IBMの研究所で開発した種々の独自技術に基く優れたスケーラビィリティを実現している。

2. 強力な補助機能

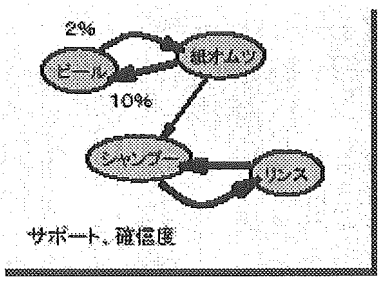
- GUIによるマイニング結果の解釈容易な視覚化表示
- 統計機能(因子分析、主成分分析、線形回帰、単変量ｶｰﾌﾞ･ﾌｨｯﾃｨﾝｸﾞ、2変量統計)
- 複数のマイニング処理をシーケンス定義して一括処理する自動実行機能
- 強力な高効率の並列処理による独自のスケーラビリティ実現
- API機能による種々プログラムの連携開発を支援
- データ・ベース(DB2)とフラット・ファイルの2形態のデータ入力に対応
- 種々の主要DBに対して透過的アクセスが可能(DataJoiner、Relational Connect)
- 幅広いプラットホームに搭載可能
- PMMLに準拠したモデルを生成すると共に他社データマイニング・ツールとの連携
- DB2からの入力形態に対してプリプロセッサーを提供(幅広いデータ加工機能)
- Modeling、Visualization、並びにScoringといった製品ラインナップの充実

データマイニングの基本機能



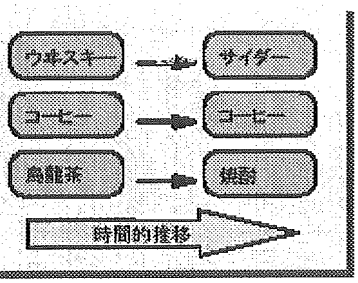
IMが提供する6つのデータマイニング技法

相関関係 (Association)



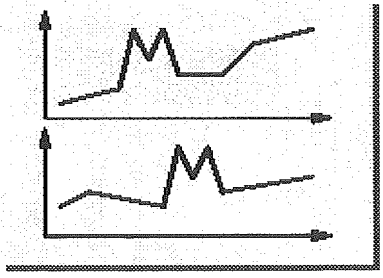
紙オムツを買う顧客の10%は同時にビールも購入する

時系列パターン (Sequential Pattern)



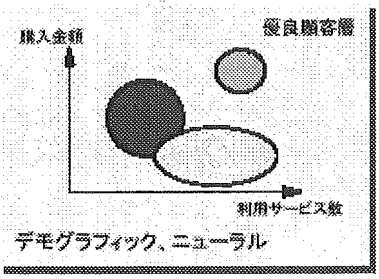
ウィスキーを買う顧客の20%は次回以降の買い物でサイダーを買う

類似時系列 (Similar Time Sequesce)



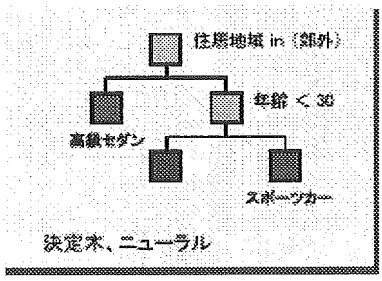
過去の商品別時系列売上データから新商品の売上パターンを予測する

クラスター分割 (Clustering)



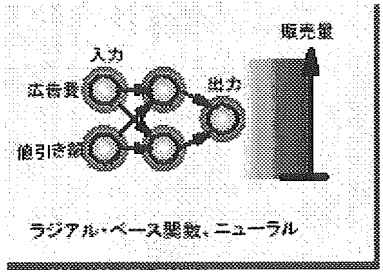
100万人の顧客を50のグループにセグメント化する

クラス判別 (Classification)

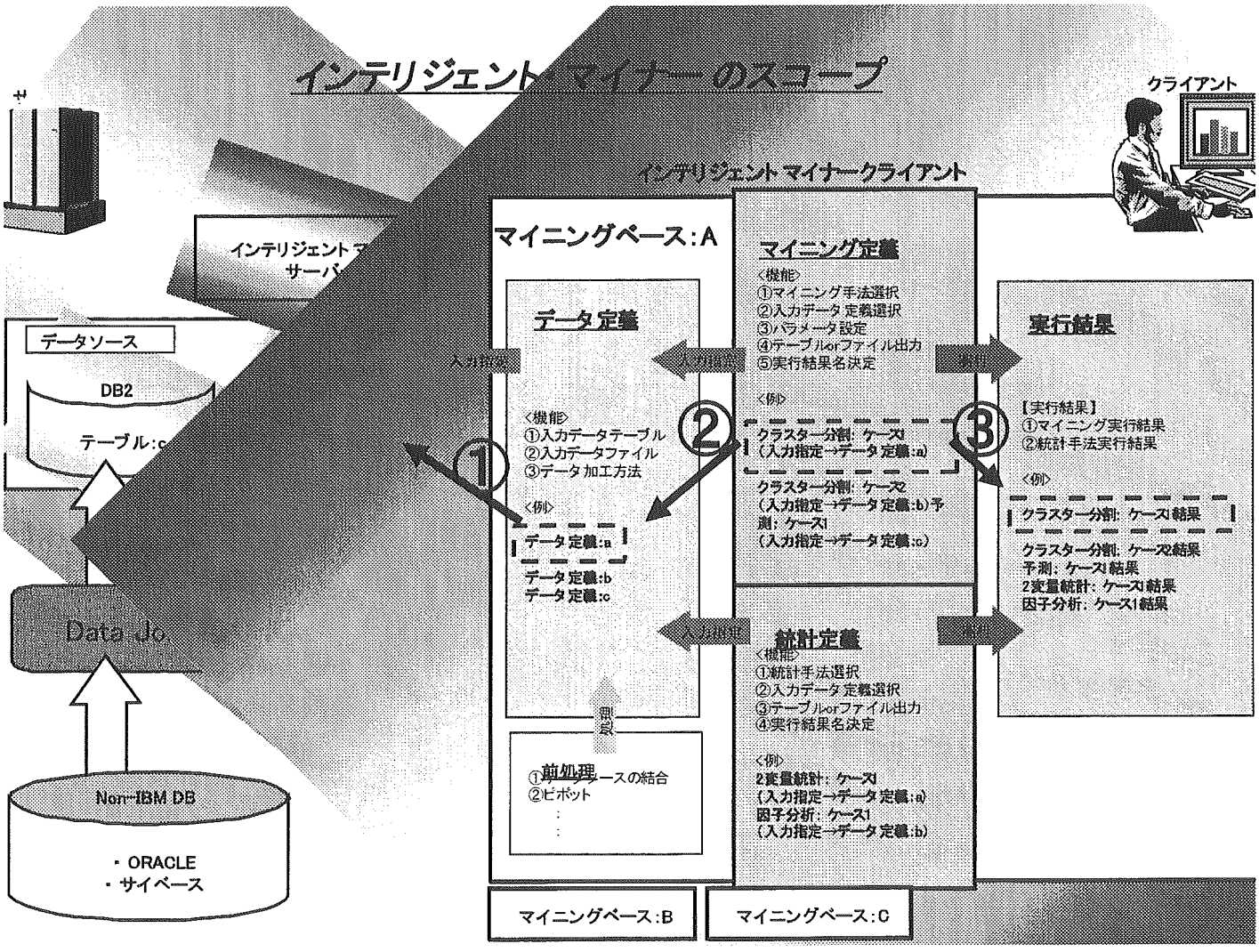


派手なスポーツカーを買うのは郊外に住む若いプロフェッショナル で、高級セダンを買うのは裕福な年配の人々である

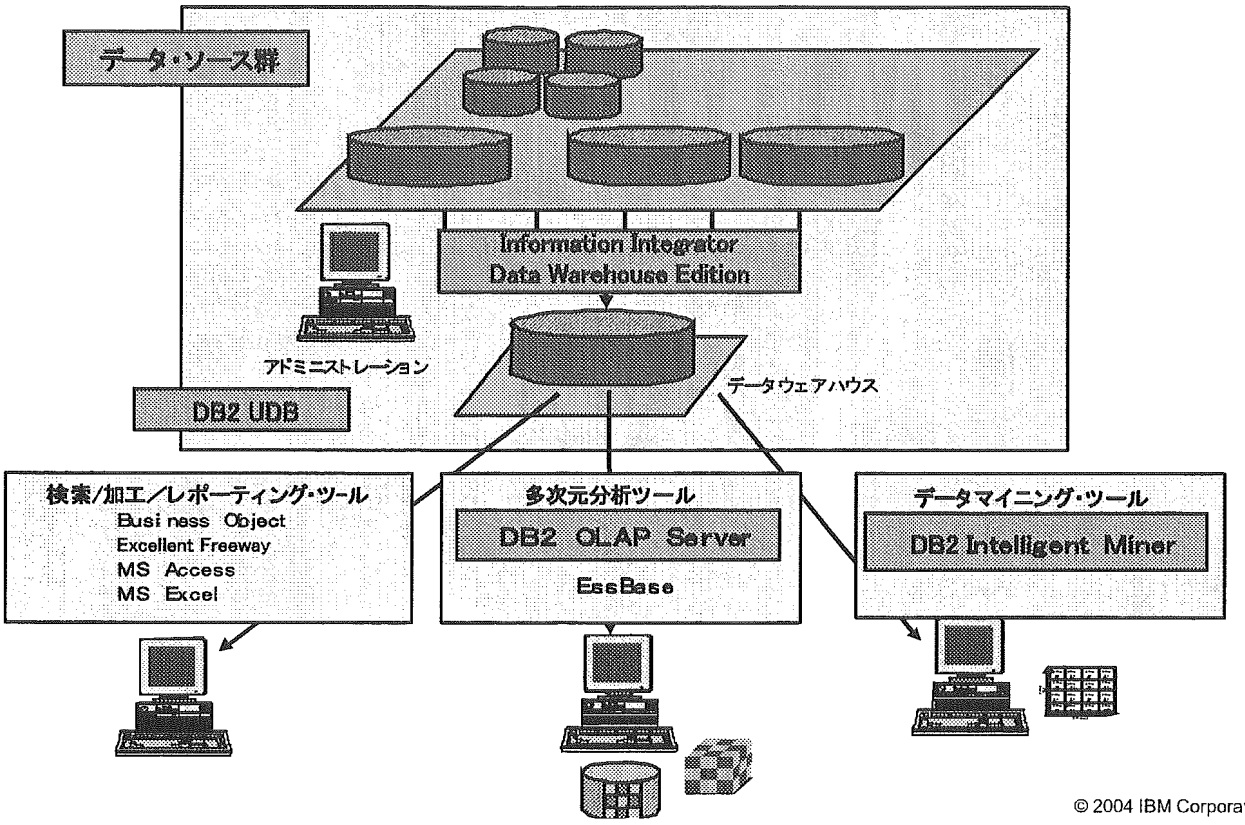
値の予測 (Value Prediction)



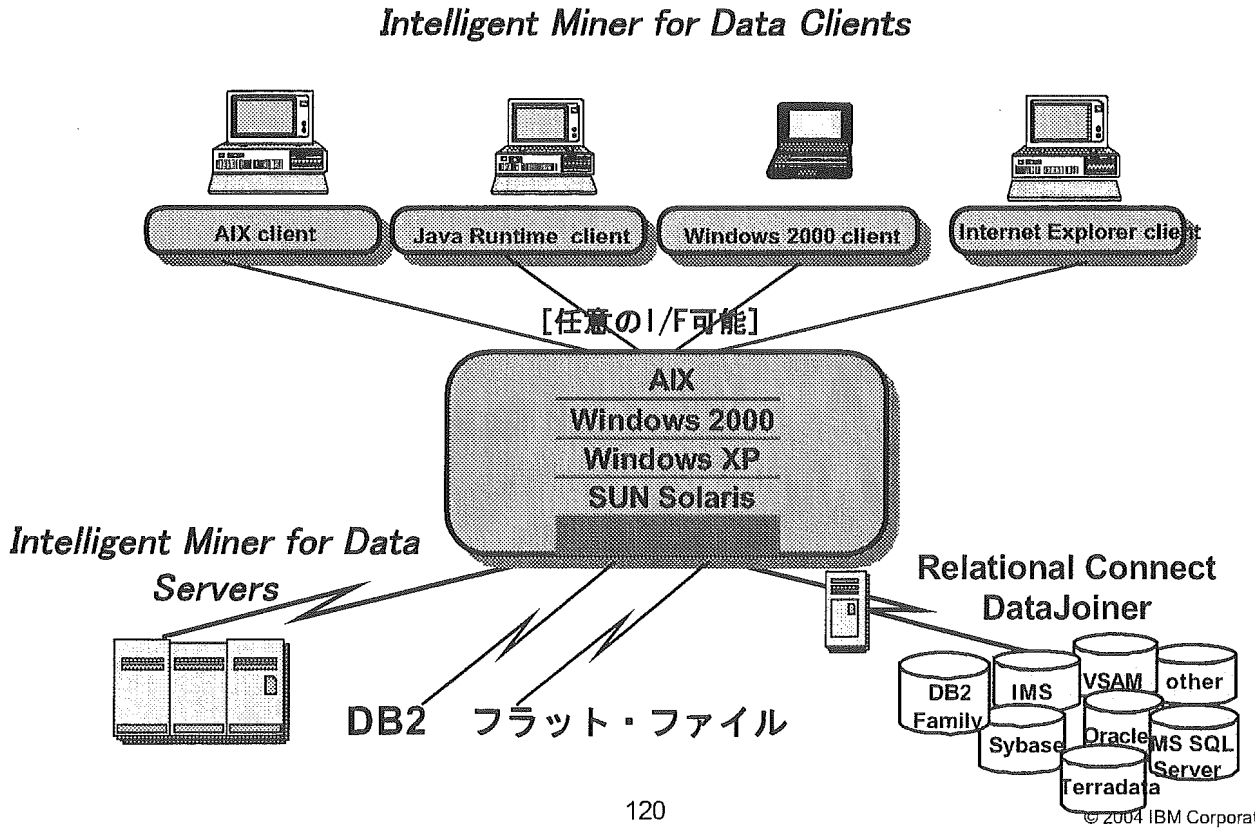
広告、値引き、陳列をパラメータとして、インスタントコーヒーの販売量を予測する



BIソリューションの構成要素例

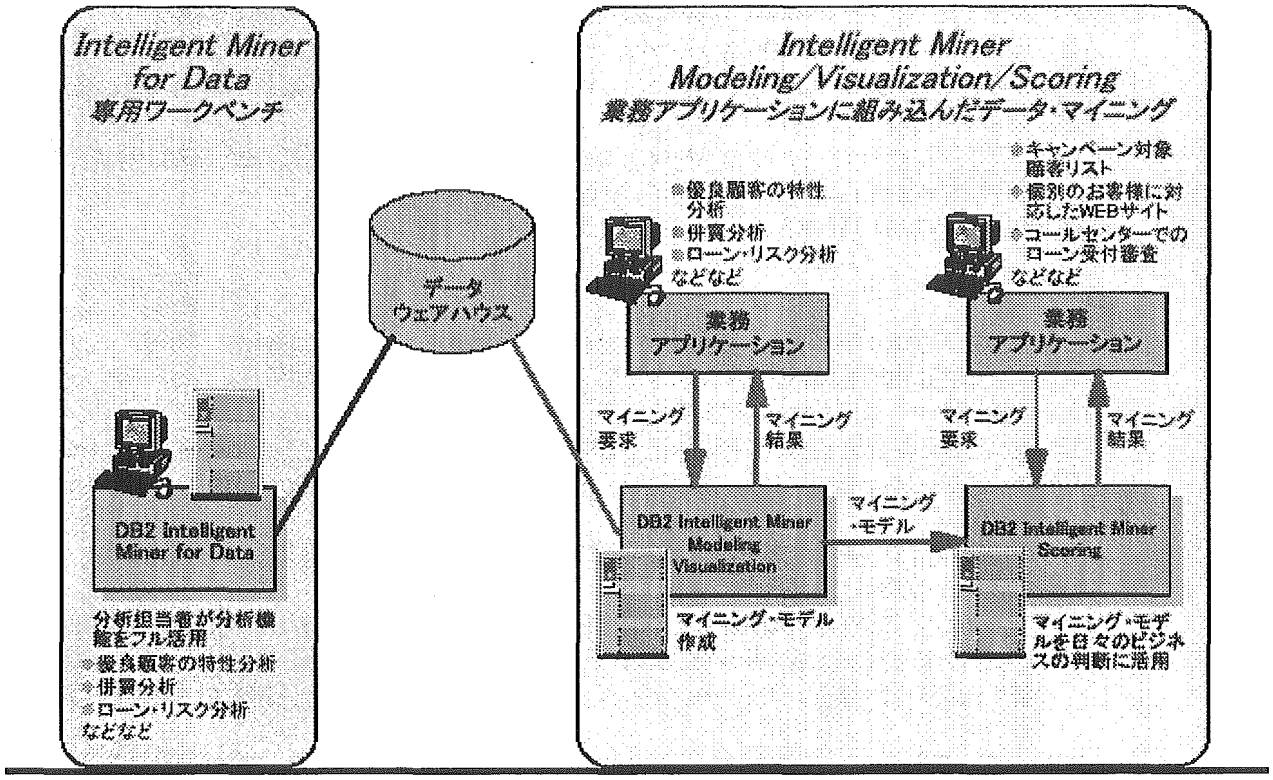


拡張性と柔軟性



Intelligent Miner V8.1ラインナップ

従来からのワークベンチツールに加え、Intelligent Minerをマイニングエンジンとして活用したアプリケーション開発をさらに容易に実現するために、DB2エクステンダーなどの形で提供される製品群が加わりました。



IM／Modeling/Visualization/Scoringの活用プロセス

Intelligent Miner Modeling / Visualization / Scoringは、Intelligent Miner for Data上でも行なわれているデータマイニングの次の3つのステップを、より容易に開発アプリケーションに組み込むことが出来ます。

ステップ	製品名	例
モデルの作成	DB2 Intelligent Miner Modeling（V8からの新製品）	ローンの与信にあたって、年齢、年収、職業、持ち家状況等、どのような項目を使って予測を行なうか等を決め、作成する。
作成されたモデルの確認	DB2 Intelligent Miner Visualization（V8からの新製品）	上記で作成された決定木などの予測モデルがどのようなものか、GUIの助けによって確認し、評価する。
新しいデータへのモデルの適用	DB2 Intelligent Miner Scoring	確認された予測モデルを用いて、新しくローンを申し込んできたお客様に対して、与信可否をリアルタイムに評価する。

参考文献

1. 論文

1) Teuvo Kohonen, The Self-Organizing Map, Proc. IEEE, vol.78, no. 9, Sept. 1990, pp. 1464-1480.
2) Manish Mehta, Rakesh Agrawal and Jorma Rissanen, SLIQ:A Fast Scalable Classifier for Data Mining, IBM Almaden Research Center, 1996.
3) Networks for Approximation and Learning, Tomas Poggio, Proc. IEEE, Vol. 78, No. 9, Sep. 1990, pp.1481-1497.
4) Fast Algorithms for Mining Association, R. Agrawal and R. Srikant, Proc. Of the 20th BM Internal Conf. on Very Large Databases, 1994.
5) Intelligent Miner User Guide, Ver. 1, Rel. 1, 1996.

2. 書籍

1) Bigus J., Data Mining with Neural Networks, NY, McGraw-Hill, 1996.
邦訳:株式会社社会調査研究所, 日本アイ・ビー・エム株式会社共訳, ニューラル・ネットワークによるデータマイニング, 日経BP社, 1997.
2) Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees and Alessandro Zanasi, Discovering Data Mining From Concept to Implementation, Prentice Hall, Inc., 1997.
邦訳:河村佳洋, 福田剛志監訳, データマイニング活用ガイド 概念から実践まで, 株式会社トッパン, 1999.
3) Michael J.A. Berry and Gordon Linoff, DATA MINING TECHNIQUES:FOR MARKETING, SALES, AND CUSTOMER SUPPORT, John Wiley & Sons, Inc., 1997.
邦訳:SAS インスティテュート・ジャパン、江原 淳、佐藤栄作、海文堂出版株式会社、1999.

備考]上記文献は参考提示したもので、飽くまでも関連文献の極く一部に過ぎない。