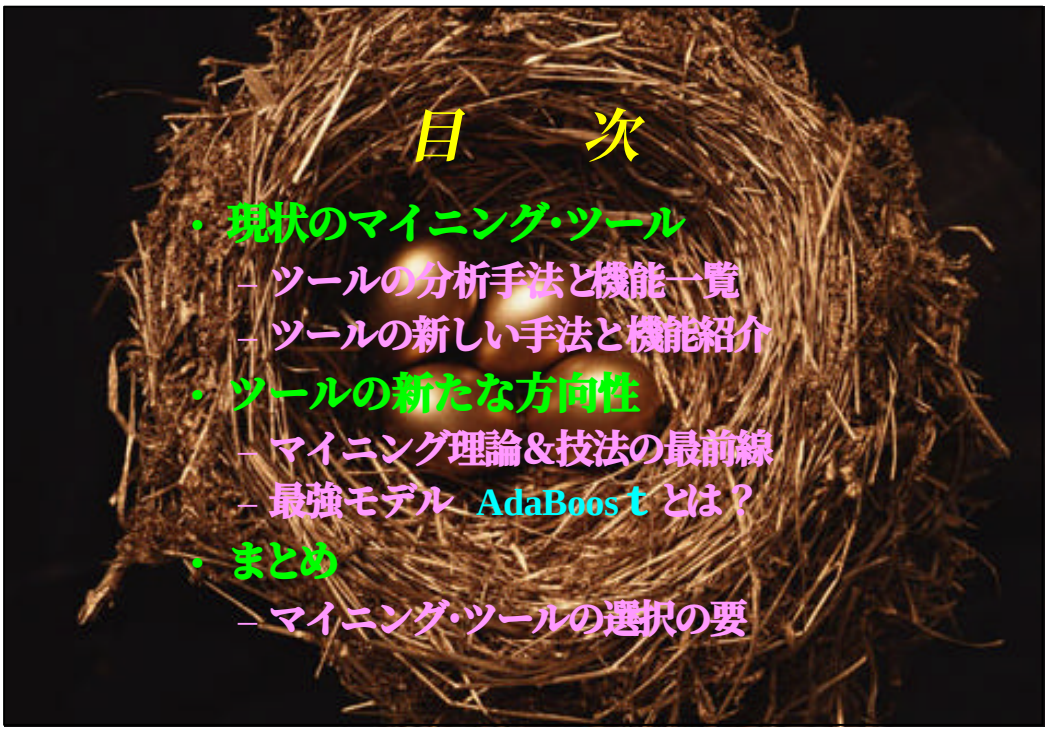

2000年9月1日
SUJ IJ2000
マイニング・ツールの比較評価と
選択のポイント
ベンダー8社のマイニング・ツール評価
株式会社三和銀行
リテール統括部
小　野　潔
kiyosi_ono@sanwabank.co.jp
TEL 03-5252-1518
（本報告は個人的見解である）


目　　次
• 現状のマイニング・ツール
– ツールの分析手法と機能一覧
– ツールの新しい手法と機能紹介
• ツールの新たな方向性
– マイニング理論&技法の最前線
– 最強モデル　AdaBoost とは？
• まとめ
– マイニング・ツールの選択の要

ツールはマイニングに必要不可欠

- 作業効率は3～5倍以上
- 日本では20種類、米国では70種類以上
- 問題点：
  - 大規模ソフトウェアのため価格が高い
  - ベンダーごとに機能や分析手法が異なる
  - 選択を誤れば無用な時間と費用の浪費となる
  - すべてを兼ね備えたツールは存在しない
  - ユーザー本位の「ツール比較レポート」がない

結局、ユーザーは「価格」と「分析手法」のみで選択

ベンダーと製品名

販売ベンダー	本社	会社	製品名	Ver	発売日	機能
東芝	日本	メーカー	KINOsuite-PR	1.0	99.04	ルール抽出機能付きニューロ
			KINOsuite-IDTF	1.0	99.04	ファジィ決定木
日本ユニシス	米国	メーカー	MiningPro21（開発は日本）	1.0	99.10	マイニング・エンジン
			分析テンプレート	1.0	99.10	マイニング・ナビ
日本IBM	米国	メーカー	Intelligent Miner	6.1	99.10	マイニング・エンジン
			Relationship Marketing	6.1	99.10	マイニング・ナビ
日本SGI	米国	メーカー	MineSet	3.0	99.07	マイニング・エンジン
日立製作所	日本	メーカー	DATAFRONT	2.01	00.03	マイニング・エンジン
富士通	日本	メーカー	SymfoWARE Parallel Mining Server	4.0	99.12	マイニング・エンジン
			SymfoWARE VisualMiner	4.0	99.12	平行座標ソフト
SAS	米国	統計ソフト	Enterprise Miner	2.02	98.07	マイニング・エンジン
SPSS	米国	統計ソフト	Clementine	5.2	99.05	マイニング・エンジン
			AnswerTree	2.1	97	決定木
			Neural Connection	2.0	95	ニューラルネットワーク
			SmartScore	1.0	99.06	開発支援ツール

分析手法

分　類	分 析 手 法	東 芝	日本ユニシス	日本IBM	日本ＳＧＩ	日 立	富士通	SAS	SPSS
決定木	C4.5	―	○	―	○	―	―	○	―
決定木	C5.0	―	―	―	―	―	―	―	○
決定木	CART	―	―	○	―	―	○	○	○
決定木	CHAID	―	○	―	―	―	―	○	○
決定木	QUEST	―	―	―	―	―	―	―	○
決定木	Pseudo De cition Tree	―	―	―	―	―	○	―	―
決定木	Opt ion Tree	―	―	―	○	―	―	―	―
決定木	ファジィ決定木	○	―	―	―	―	―	―	―
決定木	領域分割決定木	―	―	未定	―	―	―	―	―
決定木	その他機能拡張決定木	―	○	―	―	―	―	―	―
ニューロ	Back　Propagation	○	次期	○	―	△	△	―	―
ニューロ	Radical　Basis　Function	―	―	○	―	―	―	○	○
ニューロ	ベイジアン・ネットワーク	―	―	―	―	―	―	―	○
ｸﾗｽﾀﾘﾝｸ	コホーネン・ネットワーク	―	次期	○	―	次期	―	次期	○
ｸﾗｽﾀﾘﾝｸ	K-means法	―	○	―	○	―	―	○	○
ｸﾗｽﾀﾘﾝｸﾞ	Ward法	―	次期	―	―	―	○	○	○
ｸﾗｽﾀﾘﾝｸﾞ	コンドルセの手法	―	―	○	―	―	―	―	―
ｸﾗｽﾀﾘﾝｸﾞ	概念クラスタリング	次期	―	―	―	―	―	―	―
ｱｿｼｴｰｼｮﾝ	Apriori	次期	○	○	○	―	○	○	○
ｱｿｼｴｰｼｮﾝ	Generalized Rule Induction	―	―	―	―	―	―	―	○
ｱｿｼｴｰｼｮﾝ	順序ｱｿｼｴｰｼｮﾝ	―	―	○	―	―	―	○	○
ｱｿｼｴｰｼｮﾝ	類似時系列パターン	―	―	○	―	―	―	―	―
ﾙｰﾙ抽出	ﾆｭｰﾗﾙﾈｯﾄﾜｰｸ、MLP	○	―	―	―	―	―	―	―
ﾙｰﾙ抽出	ﾙｰﾙｲﾝﾀﾞｸｼｮﾝ	―	―	―	―	○	―	―	―
最小近傍法	MBR	―	―	―	―	○	○	次期	―
回帰分析	ロジスティック回帰分析	―	―	―	○	―	―	○	○
ﾃｷｽﾄﾏｲﾆﾝｸﾞ	テキストマイニング	△	△	△	―	未定	△	次期	―
ﾃｷｽﾄﾏｲﾆﾝｸﾞ	Concept Base Search	―	△	△	―	―	△	―	―

機能一覧

	東芝	日本ユニシス	日本IBM	日本ＳＧＩ	日立	富士通	SAS	SPSS
並列処理	◇	―	○	○	◇	○	―	―
UNIX版	―	○	○	○	○	○	○	○
NT版	○	○	○	○	○	○	○	○
LINUX版	―	―	―	次期	―	―	未定	―
メインフレーム版	―	―	○	―	―	―	未定	―
リフト図（モデル的中率の比較）	―	次期	○	―	未定	―	○	○
説明属性の選択機能	―	―	―	○	―	○	○	○
高度な多次元散布図	―	―	―	◎	△	◎	◎	○
ＫＤＤプロセスの可視化	△	△	△	△	△	△	○	○
対費用効果を含めた収益計算	―	―	○	○	―	―	○	○
初心者向けのナビゲータ機能	未定	○	◎	―	次期	―	―	―
プロセスのC or XLM言語変換	―	―	―	―	―	―	次期	○
レポート機能	―	○	―	―	―	―	○	◎
日本語のマニュアル本	―	○	○	―	―	―	○	○
定期セミナーの開催	―	○	○	―	―	―	○	○
専門の質問応答セクション	△	△	△	△	△	△	○	○
価格 （Aは３００万円未満、Cは1000万円以上）	B	A	C	B	A	A	C	B

ツールの新しい機能

①プロセスのC　or XML言語変換機能
　　XMLはインターネットの標準言語になる
②運用形態
　　クライアント･サーバー型
　　シングル･アローン型
③インターフェース用開発支援ツール
　　モデル更新に伴うインターフェース画面の変更を
　自動的に行なう
④マイニング ナビゲーター機能
　　初心者でもある程度のマイニングが可能
　　仕組はシステム部門が前もって加工データを用意

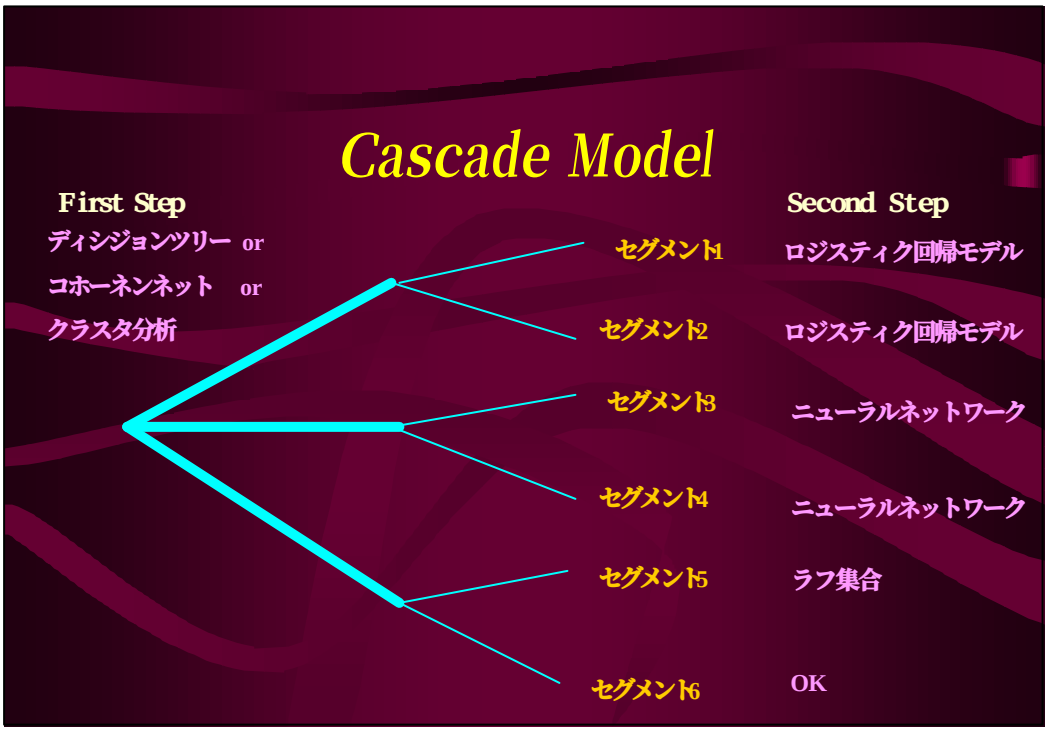
注目の分析手法と技法

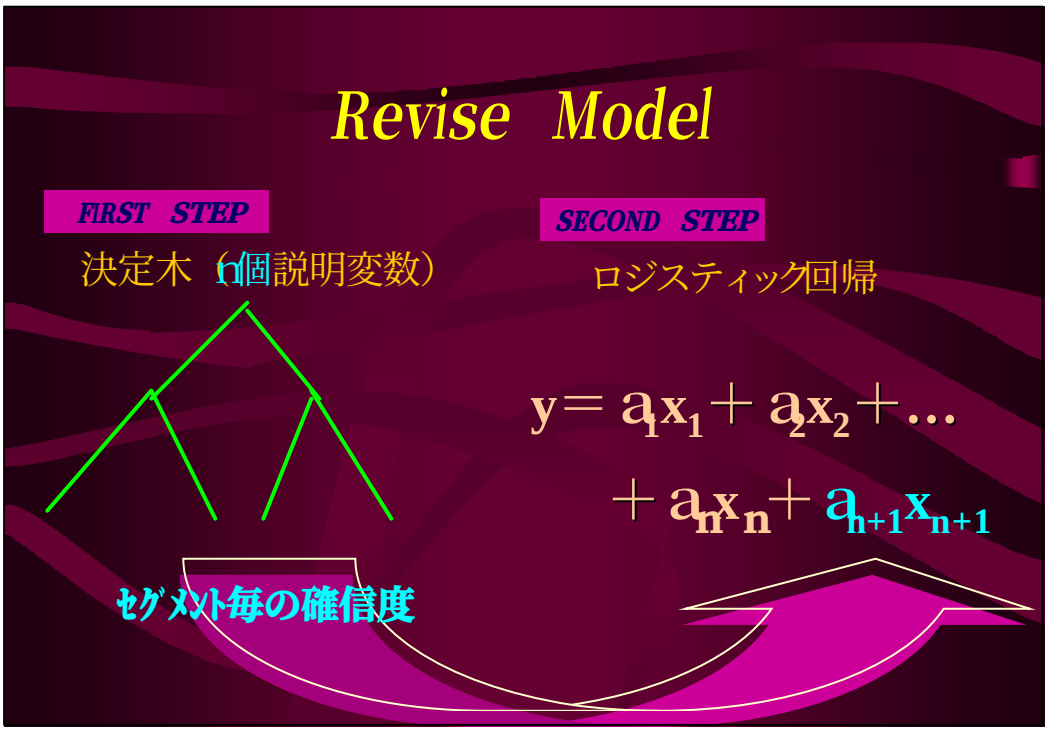
- MBR(Memory Based Reasoning,記憶に基づく推論)
  - 最小近傍(Nearest Neighbor)法
  - モデルの更新が必要ない
  - 欠点 :ブラックデータが数千個必要
- Rough Sets
  - 欧米ではファジィー集合より使われている
- SVM(Support Vector Machines)
  - key word: カーネル法　マージン
  - AdaBoosting　との関係
- Cross Validatin(=Jack knife)　精度を高める技法
  - k-fold cross-validation　(leave-one-out,)
  - multiple cross-validation

マイニング理論の最前線

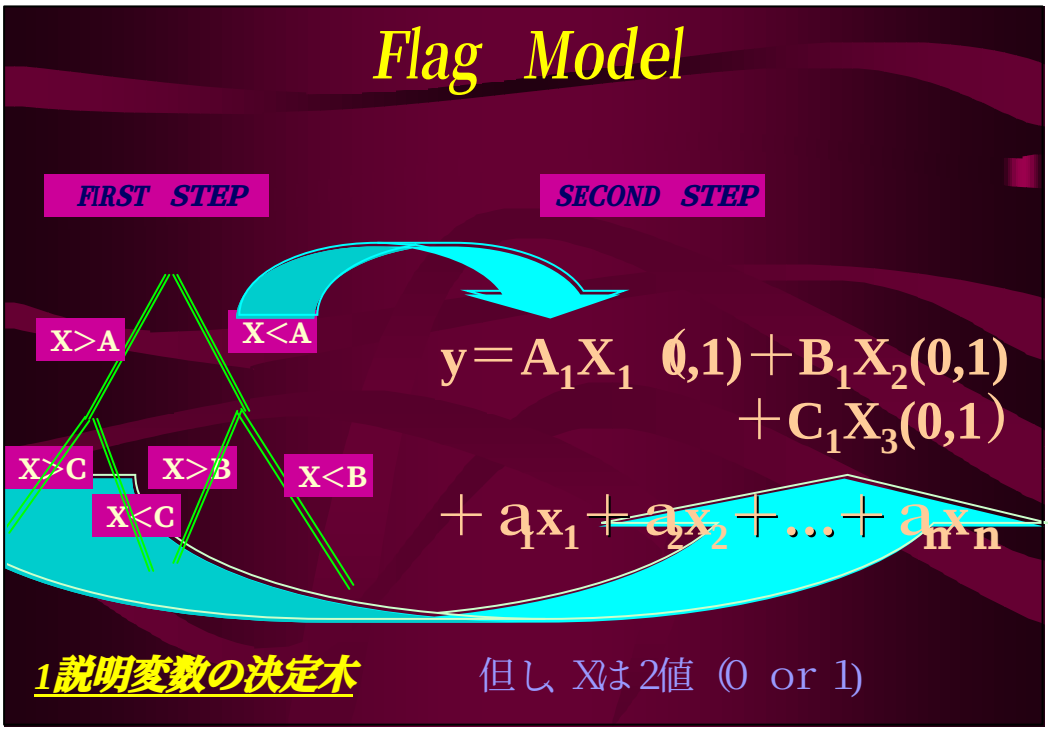
- 分類器の組み合せ
  - ①Hybrid Model
    - Cataract Model (連続型)
    - Cascade Model (直列型)
    - Revise Model(補正型)
    - Flag Model (CARTを利用してｶﾃｺﾞﾘ･ﾌﾗｸﾞ作成)
  - ②Committee Model
    - Bagging (並列型)
    - AdaBoosting (モデルの重み付け結合型)
    - ECOC (Error Correcting Output Codes)
- SAS のEnsemble Model の構成
  - Combined Model
  - Stratified Model
  - Bagging Model
  - Boosting Model

Cataract Model


Cascade Model
First Step
ディシジョンツリー or
コホーネンネット or
クラスタ分析
セグメント1
セグメント2
セグメント3
セグメント4
セグメント5
セグメント6
Second Step
ロジスティク回帰モデル
ロジスティク回帰モデル
ニューラルネットワーク
ニューラルネットワーク
ラフ集合
OK


Revise Model
FIRST STEP
決定木（n個説明変数）
SECOND STEP
ロジスティック回帰
y= a1x1 + a2x2 + …
+ anxn + an+1xn+1
セグメント毎の確信度

Flag Model



1説明変数の決定木　　但し、Xは2値（0 or 1）

Committee Model

	ディシジョンツリー	ニューラルネットワーク	ロジスティク回帰
顧客A	90%	95%	70%
顧客B	15%	10%	11%
顧客C	70%	50%	40%
顧客D	60%	65%	55%
顧客E	10%	15%	20%

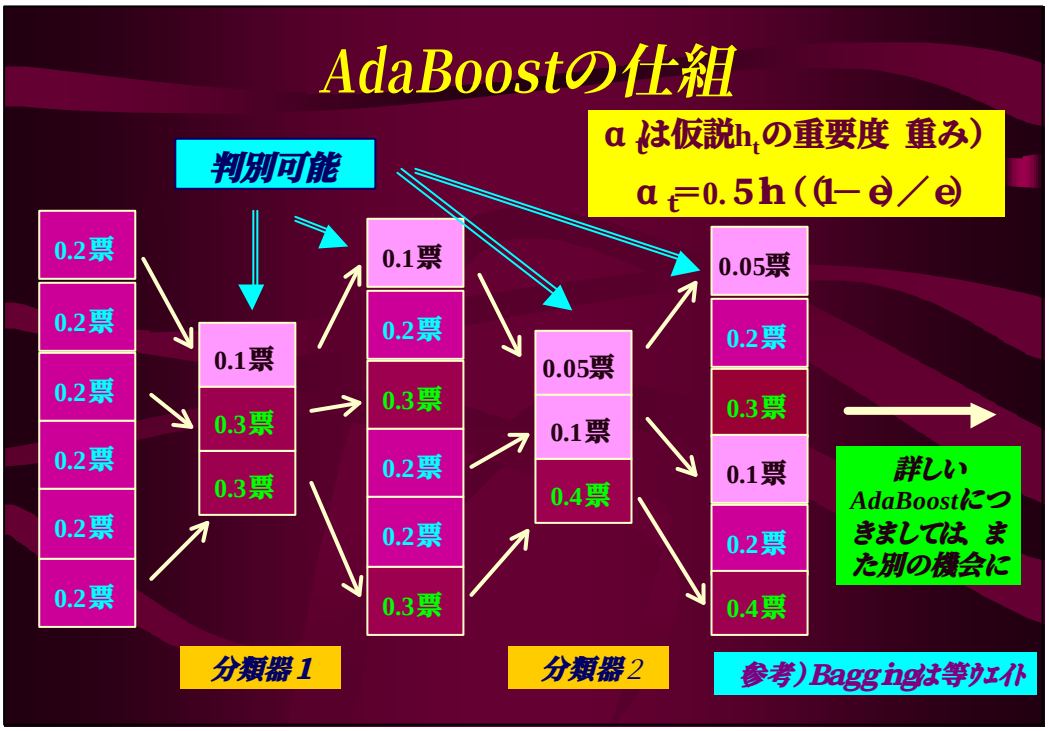
多数決、線形結合、最小値、最大値、平均　等

	総合点
顧客A	70%
顧客B	12%
顧客C	50%
顧客D	55%
顧客E	17%

同じ分類器

Bagging, Boosting

AdaBoostの仕組



最強のAdaBoost

- ***精度の低い弱学習分類器を精度の高い強学習分類器にするためのアルゴリズム***
  - 例えば、10〜100個の複数の分類器の多数決
  - 簡単に言うと、***AdaBoost***は分類機のターボエンジン
- 今、最も米国の学会ではホットな話題
  - 数学的な裏付け？
  - 過学習が起きない？
- エンジン（分類器）によって最適ターボが違う
  - 現在、多値、Local Classificaion 、雑音、データ不足に対応した新しい AdaBoostモデルが現われた

ツールの選定基準（最後に）

1.マイニングの成功は、良いデータ・分析者・ツールが必要不可欠

マイニング・ツールは未発達である
最初はベンダーによるコンサルトが望ましい

2.ツールの決定要素は？

使用目的、ベンダーのコンサルタント能力、運用形態、分析者の素質等

3.ユーザーは新しい分析手法や機能を絶えずチェック

これからツールには、Bagging，AdaBoostが必要
SAS/EM ver4.0 は頭一つリードしているようだが
他社ツールの追随確実