データマイニングと機械学習

櫻井彰人
慶應義塾大学理工学部
管理工学科

データと情報

- 現代社会は大量のデータを産出する
  - データ源：ビジネス、研究、医療、経済、地理、環境、スポーツ、、、、
- 潜在的に価値有るデータ源、しかし、
- 生データは役立たない：自動的に情報を抽出する技術が必須
  - データ：記録された事実
  - 情報：データに隠された規則性

情報が必須

- 例1：体外受精
  - 所与：胚の60の特徴量
  - 課題：生存する胚の選択
  - データ：胚と結果の履歴
- 例2：牛の間引き
  - 所与：牛の700の特徴量
  - 課題：間引きすべき牛の選択
  - データ：牛と農場主の判断の履歴

データマイニング

- 隠れた、これまでは知られていない、潜在的に重要な情報を、データから抽出する
- 必要：データ内のパターンや規則性を抽出するプログラム
- 明確なパターンは予測に有用
  - 問題1：多くのパターンは面白くない
  - 問題2：パターンは不正確かも（完全にみかけだけ、ということも）。データが取り違えられたり失ったりした場合
- パターンに従わない例を発見することも含む
  - 例：不正利用、不正アクセス

機械学習の技法

- データマイニングの技術基盤：事例から構造記述を獲得するアルゴリズム
- 構造記述はパターンを明示的に記述
  - 新しい状態での結果の予測に利用可能
  - 予測が出された理由の理解・説明が可能（こちらの方がより重要）
- 人工知能、統計学、データベース研究にルーツをもつ方法

構造記述

- 例えば、if-then 規則

If 涙産生量 = 少 then 推奨 = しない
Otherwise, if 年齢 = 若い and 乱視 = no then 推奨 = ソフト

年齢	めがね	乱視	涙産生	推奨
若い	近視	なし	少	しない
若い	遠視	なし	通常	ソフト
老眼以前	遠視	なし	少	しない
老眼以前	近視	あり	通常	ハード
…	…	…	…	…

機械は本当に学習するのか？

- 「learning」の辞書的定義
  - 勉強・経験・教えられることにより知識を得る
  - 情報や観測から気が付くようになる
    - 測定できない
  - 記憶する
  - 知らされる、確認する；指示を受ける
    - 計算機ならすでに行っている
- 操作的な定義
  - 事物は、将来の性能が向上するように行動を変更する時、事物は learn するという
    - 靴は学習するか？
- learning は意図を包含するか？

お天気問題

- ゲームを行う条件

見通し	気温	湿度	風	ゲーム
晴れ	暑い	高い	なし	しない
晴れ	暑い	高い	あり	しない
本曇り	暑い	高い	なし	する
雨	普通	普通	なし	する
…	…	…	…	…

If 見通し = 晴れ and 湿度 = 高い then ゲーム = しない
If 見通し = 雨 and 風 = あり then ゲーム = しない
If 見通し = 本曇り then ゲーム = する
If 湿度 = 通常 then ゲーム = する
Ifどれでもない then ゲーム = する

分類と相関規則

- 分類規則：予め指定した属性の値の予測（事例の分類）
  - If 見通し = 晴れ and 湿度 =高い then ゲーム = しない
- 相関規則：任意属性や組合せた属性の値の予測
  - If 気温 = 涼しい then 湿度 = 通常
  - If 湿度 = 通常 and 風 = なし then エーム = する
  - If 見通し = 晴れ and ゲーム = しない then 湿度 = 高い
  - If 風 = なし and ゲーム = しない then 見通し = 晴れ and 湿度 = 高い

お天気問題（複数種属性）

- ゲームを行う条件

見通し	気温	湿度	風	ゲーム
晴れ	85	85	なし	しない
晴れ	80	90	あり	しない
本曇り	83	86	なし	する
雨	75	80	なし	する
…	…	…	…	…

If 見通し = 晴れ and 湿度 >83 then ゲーム = しない
If 見通し = 雨 and 風 = あり then ゲーム = しない
If 見通し = 本曇り then ゲーム = する
If 湿度 <85 then ゲーム = する
Ifどれでもない then ゲーム = する

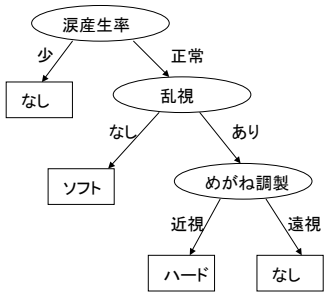
コンタクトレンズ・データ

年齢	めがね	乱視	涙産生	推奨
若い	近視	なし	少	しない
若い	近視	なし	通常	ソフト
若い	近視	あり	少	しない
若い	近視	あり	通常	ハード
若い	遠視	なし	少	しない
若い	遠視	なし	通常	ソフト
若い	遠視	あり	少	しない
若い	遠視	あり	通常	ハード
老眼以前	近視	なし	少	しない
老眼以前	近視	なし	通常	ソフト
老眼以前	近視	あり	少	しない
老眼以前	近視	あり	通常	ハード
老眼以前	遠視	なし	少	しない
老眼以前	遠視	なし	通常	ソフト
老眼以前	遠視	あり	少	しない
老眼以前	遠視	あり	通常	しない
老眼	近視	なし	少	しない
老眼	近視	なし	通常	しない
老眼	近視	あり	少	しない
老眼	近視	あり	通常	ハード
老眼	遠視	なし	少	しない
老眼	遠視	なし	通常	ソフト
老眼	遠視	あり	少	しない
老眼	遠視	あり	通常	しない

規則集合

If 涙産生 = 少 then 推薦 = しない
If 年齢 = 若い and 乱視 = なし and 涙産生 = 通常 then 推薦 = ソフト
If 年齢 = 老眼以前 and 乱視 = なし and 涙産生 = 通常 then 推薦 = ソフト
If 年齢 = 老眼 and めがね = 近視 and 乱視 = なし then 推薦 = なし
If めがね = 遠視 and 乱視 = なし and 涙産生 = 通常 then 推薦 = ソフト
If めがね = 近視 and 乱視 = あり and 涙産生 = 通常 then 推薦 = ハード
If 年齢 = 若い and 乱視 = あり and 涙産生 = 通常 then 推薦 = ハード
If 年齢 = 老眼以前 and めがね = 遠視 and 乱視 = あり then 推薦 = しない
If 年齢 = 老眼 and めがね = 遠視 and 乱視 = あり then 推薦 = しない

本問題に対する決定木



アヤメの分類

	がく弁の長さ	がく弁の幅	花弁の長さ	花弁の幅	型
1	5.1	3.5	1.4	0.2	Iris setosa
2	4.9	3	1.4	0.2	Iris setosa
...					
51	7	3.2	4.7	1.4	Iris ve io
52	6.4	3.2	4.5	1.5	Iris ve
...					
101	6.3	3.3	6	2.5	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

If 花弁の長さ < 2.45 then Iris setosa
If がく弁の長さ < 2.10 then Iris versicolor
. . .

計算機性能の予測

- 例: 209種の計算機構成

	サイクル時間 (ns)	主記憶 (Kb)		キャッシュ (Kb)	チャンネル		性能
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
108	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

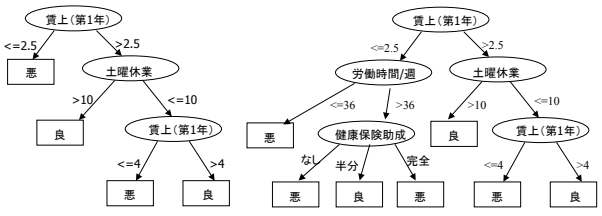
- 線型回帰式

PRP = –55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
+ 0.6410 CACH – 0.2700 CHMIN + 1.480 CHMAX

労使交渉データ

属性	型	1	2	3	...	40
継続期間	(年数)	1	2	3		2
賃上げ(第1年)	百分率	2	4	4.3		4.5
賃上げ(第2年)	百分率	?	5	4.4		4
賃上げ(第3年)	百分率	?	?	?		?
生活費保証	[none, tcf, tc]	none	tcf	?		none
労働時間/週	時間数	28	35	38		40
年金	[none, ret-allw, empl-cntr]	none	?	?		?
stand-by pay	百分率	?	13	?		?
変則勤務手当て	百分率	?	5	4		4
教育手当て	[あり、なし]	あり	?	?		?
土曜休業	休日数	11	15	12		12
休暇	[平均以下、平均、平均以上]	平均	平均以上	平均以上		平均
長期傷害助成	[あり、なし]	なし	?	?		あり
歯科診療保険助成	[なし、半分、完全]	なし	?	完全		完全
死別助成	[あり、なし]	なし	?	?		ari
健康保険助成	[なし、半分、完全]	なし	?	完全		半分
対応	[良い、悪い]	悪い	良い	良い		良い

労使交渉データの決定木



大豆の分類

	属性	値の個数	値の例
環境	発生日	7	7月
	降水	3	平均以上
...			
種子	状態	2	正常
	徴の成長	2	なし
...			
果実	さやの状態	4	正常
	斑	5	?
葉	状態	2	異常
	斑の大きさ	3	?
...			
茎	状態	2	異常
	Stem lodging	2	yes
...			
根	状態	3	正常
診断		19	Diaporthe stem canker

領域知識の役割

If 葉の状態 = 正常 and
茎の状態 = 異常 and
茎の胴枯れ病 = 土壌線以下 and
胴枯れ病病変部位の色 = 褐色
then
診断 = rhizoctonia root rot

If 葉の奇形 = なし and
茎の状態 = 異常 and
茎の胴枯れ病 = 土壌線以下 and
胴枯れ病病変部位の色 = 褐色
then
診断 = rhizoctonia root rot

実地に使用された適用事例

- 学習結果・学習方法が実応用された例
  - 携帯電話のchurning発見・防止
  - 飛行物体の自動分類
  - 不正検出
    - 医療行為（レセプトから）、マネーロンダリング
  - 文書分類（ニュースグループ、電子メイル、ドキュメント）
  - ....

ローン申込み処理

- 所与：経済状況・個人情報に関する質問表
- 問題：貸してよいか？
- 単純な統計的方法で90%は処理できる
- 境界線上のものはローン担当者に
- しかし：境界線上で認可した50%は不履行
- 解(?)：境界線上はすべて拒絶
  - NO! 境界線上の顧客は利用が多い

機械学習を用いて

- 境界線上の1000訓練データ
- 20属性：年齢、（現在の）勤続年数、（現在の）居住年数、（銀行との）取引年数、他のクレジットカード状況、、、、、
- 学習結果（規則）は、境界線上の事例の2/3 を正しく予測
- その上：顧客に決定を説明するのに、その規則を用いることができた

画像のスクリーニング

- 所与：沿岸海洋上レーダー衛星画像
- 問題：画像から石油による光沢面を発見
- 石油面：は暗い領域で大きさや形が変わる
- 易しくはない：暗領域は気象条件（強風等）によっても生じる
- 高度に訓練された専門家によるコスト高な処理が必要

機械学習を用いて

- 正規化された画像から暗領域を抽出
- 属性：領域の大きさ、形、範囲、強さ、明瞭さ、境界のぎざぎざ度合い、他領域の近さ、背景に関する情報
- 制約：
  - 訓練データの不足（石油漏れはまれ）
  - データが不均衡：暗領域の殆どは石油ではない
  - 同一画像の領域は一群をなす
  - 要請：false-alarm 率を調整可能に

需要予想

- 電力会社：電力需要の将来予想が必要
- 毎時の電力需要の最小・最大値が正確に予測できれば大きな節約となる
- 所与：正常気象のもとでの静的需要モデル（手で作成）
- 課題：気象条件に合わせて変更する
- 静的モデル構成要素：年間の基礎需要、年間の需要曲線、休日の影響

機械学習を用いて

- 「最も良く似た」日を用いて予測修正
- 属性：温度、湿度、風速、雲量。実際の需要と予測値との差。
- 最も良く似た3日の平均差異が静的モデルに加算
- 類似性判定関数中の属性毎の荷重は、線型回帰で求める

機械欠陥の診断

- 診断：エキスパートシステムの特異領域
- 所与：装着台上の様々な位置での振動のフーリエ解析結果
- 問題：欠陥の同定
- 電動機や発電機の予防的保守
- これまで：専門家か手で作った規則を用いた診断

機械学習を用いると

- 利用可能：専門家が診断した 600 欠陥
- ~300 は不充分。残りを学習に使用
- 属性の追加：
  因果関係を表す領域知識を具現する中間概念
- 初期の規則に専門家は満足せず：
  専門家の領域知識につながらないため
- 背景知識を追加：
  満足できる、より複雑な規則を得る
- 学習結果は手作り規則より高性能

マーケティング Ⅰ

- 企業は大量のマーケティングや販売データを記録している
- 可能な応用：
  - 顧客ロイヤルティ：顧客の行動変化から逃げやすい顧客を発見する（銀行、電話会社等）
  - 特別サービス：利益をもたらす顧客の発見（例：休暇シーズンに資金が必要なクレジットカード優良顧客）

マーケティング Ⅱ

- バスケット分析
  - 一取引中に発生しやすい、品目のグループを発見する技法（POSデータ解析によく利用）
- 購入パターンの履歴分析
- 潜在顧客の同定
  - プロモーション用DMをフォーカスする（目標を絞ったキャンペーンはマスマーケティングより効果的）

機械学習と統計学

- 歴史的差異(大幅に単純化した):
  - 統計学: 仮説検定、単純モデル&確率論
  - 機械学習: 正しい仮説の発見、複雑モデル&確率的根拠が弱い
- しかし: 重なりは多い
  - 決定木: C4.5 や CART
  - 最近接法(nearest-neighbor)
- 今日: 収束してきている
  - 機械学習の多くは統計的手法を利用

探索としての一般化

- 帰納学習: データを記述する「概念」の発見
- 例: 記述言語として規則集合を用いる
  - 巨大であるが、有限の探索空間
- 単純な開放: 概念を数え上げ、訓練例に合わない記述を捨てていく
  - 残った記述が目標概念を含む

概念空間の数え上げ

- お天気問題の探索空間
  - $4 \times 4 \times 3 \times 3 \times 2 = 288$ の可能な規則
  - 14規則以下としても 2.7×10^{24} の規則集合が可能
- 改善法: 候補削減法(candidate-elimination)
- 他の実際上の問題
  - 2個以上の規則が最後まで残る
  - 残る規則がなくなる
    - 記述言語では目標概念が記述できない
    - データにノイズがある可能性

バージョン空間

- データに矛盾しない概念記述のなす空間
- 2つの集合で完全に規定できる
  - L: 正例をすべて覆い、負例を全く覆わない記述の内、最も特殊なもの
  - G:正例をすべて覆い、負例を全く覆わない記述の内、最も一般的なもの
- L と G のみを保守し更新すればよい
- しかし: 計算量的には、まだ、高価
- そして: 他の問題は解決しない

バージョン空間の例

- 所与: 赤または緑の牛または鶏

L = {} G = {< *, * >}

<緑, 牛> : 正例

L = {<緑, 牛>} G = {< *, * >}

<赤, 鶏>: 負例

L = {<緑, 牛>} G = {< 緑, * >, <*, 牛>}

<緑, 鶏>: 正例

L = {<緑, *>} G = {< 緑, * >,}

候補削除アルゴリズム(1/3)

- *G*←*H*の一般化が極大である仮説の集合
- *S*←*H*の特殊化が極大である仮説の集合
- 各訓練事例 *d* 毎に次のことを行なう

候補削除アルゴリズム(2/3)

- *d* が正例であるとき
  - *d* と不整合な仮説を *G* から除去
  - *S*中の*d* と不整合な仮説 *s* につき
    - *s* を*S* から除去
    - *s* の一般化が極小である *h* を全て *S* に加える
      - *h* は *d* と整合的であり、
      - *G* の中に *h* より一般なものがある
    - *S*から、*S* 中の他の仮説より一般な仮説をすべて除去する

候補削除アルゴリズム(3/3)

- *d* が負例であるとき
  - *d* と不整合な仮説を *S* から除去
  - *G* 中の*d* と不整合な仮説 *s* につき
    - *s* をG から除去
    - *s* の特殊化が極小である *h* を全て *G* に加える
      - *h* は *d* と整合的であり、
      - *G* の中に *h* より特集なものがある
    - *G*から、*G* 中の他の仮説より特殊な仮説をすべて除去する

バイアス

- 学習システムを作る時の最重要な決定
  - 概念を記述する言語
  - 概念の空間を探索する順序
  - 個別の訓練データへの過剰適応を避ける方法
- これらが探索の「バイアス」を構成する
  - 言語バイアス
  - 探索バイアス
  - 過剰適応回避バイアス

言語バイアス

- 最重要な問: 言語は万能か、それとも学習できるものを制限しているか？
- 言語は、(可能な)例集合がすべて記述できるなら、万能という。
- 言語は、論理的ORが記述できれば万能(論理的ANDは記述できると想定)
  - 例: 規則集合
- 領域知識を用いて、探索対象から、a priori にある概念を取り除くことができる

探索バイアス

- 探索のヒューリスティックス
  - Greedy探索: 一歩ごと、一歩としては最適
  - ビーム探索: 複数の選択肢を保持
  - .....
- 探索方向
  - 一般から特殊へ
    - 条件を付加しながら規則を特殊化していく
  - 特殊から一般へ
    - 個別事例を規則へ一般化していく

過剰適応回避バイアス

- 探索バイアスの一つとして見える
- (ある規則を受け入れる)評価基準の変更
  - 例: 単純さと誤差の多さ
- 探索戦略の変更
  - 例: 枝刈(記述の簡単化)
    - 先刈: 過剰に複雑な規則の探索に進む前の単純な規則にとどまる
    - 後刈: 複雑な記述を生成した後に、それを簡単化する

データマイニングと倫理 Ⅰ

- 実応用には、多数の倫理的問題がある
- データマイニングはしばしば「区別」に利用
  - ローン申請：ある情報（性、宗教、人種等）を用いることは反倫理的
- 倫理的状況は応用に依存
  - 例：医療応用では同一の情報は利用可
- 属性は問題となる情報を含んでいる
  - 例：地域コードは人種と相関しているかもしれない

データマイニングと倫理 Ⅱ

- 実応用に関する重要な問題
- 誰が、データにアクセスできるのか
- 何の目的で、データが収集されたか
- どんな種類の結論が、合法的に導出できるのか
- 警告を、結果につけるべき
- 純粋に統計的議論だけは、常に不充分

レポート課題

- 「セブンイレブンの論理思考」（週刊東洋経済2003年4月12日号 pp.28-33）を読んで、データマイニングの実応用上の可能性と限界について、論じなさい