データマイニングの基礎

Multidisciplinary Design Exploration (MDE)
Lecture Series 1

2007年3月23日

東北大学大学院 情報科学研究科／(株)三菱総合研究所
寺邊 正大

本日の内容

1. データマイニングとは？
   - 定義、歴史、手法、応用分野
2. データマイニングを使う
   - 少数の教師つきデータから効率良く学習する
   - 複数のモデルを組み合わせて精度を向上させる
   - 文章情報をマイニングする
3. さらにデータマイニングについて知るために

1. データマイニングとは？

「データマイニング」とは何か？

- データマイニングが身近になった
  - 「データマイニングを使って･･･」
  - 「マイニングする･･･」
  - 「データを掘る･･･」

Q:データマイニングの定義は？

- 3択問題

1. 「大量のデータから目的に沿ったモデル（知識）を掘り起こす技術」
2. 「データが内包する規則や特徴的なパタンを発掘すること」
3. 「汎化能力が高く、新規性が高く、実用的で、理解できる知識を抽出するプロセス」

A:データマイニングの定義は？

- 正解は？

データマイニングとは①

1.「大量のデータから目的に沿ったモデル（知識）を掘り起こす技術」

- 大量のデータ
  - 対象をよく表した大量のデータがあることが望ましい
  - しかし、大量のデータは必須ではない
- 目的に沿った
  - 分析者の「目的に沿った（＝使える知識を提供する）」モデルを発掘するのがデータマイニング
  - モデル化できれば何でも良い、というわけではない

データマイニングとは②

2.「データが内包する規則や特徴的なパタンを発掘する」

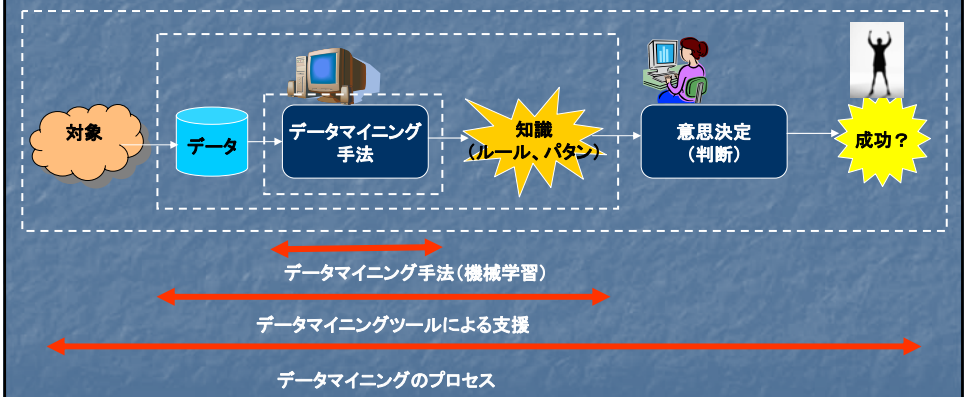
- 機械学習（計算機）の立場からの定義
  データに含まれる規則や特徴的なパタンを発掘する
- 規則やパタンが目的に沿っているか、理解できるか、有用か、ということを決めるのは「人間」
- データマイニングは、人間とは別（全自動）ではありえない

データマイニングとは③

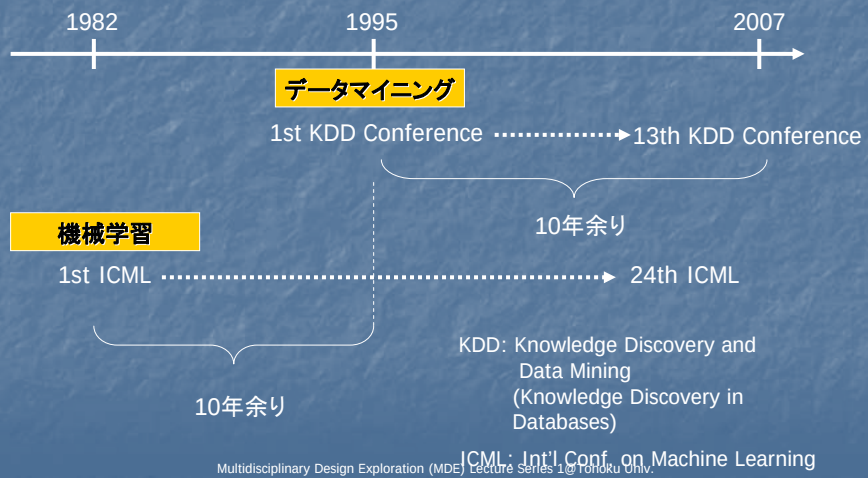
3.「汎化能力が高く、新規性が高く、実用的で、理解できる知識を抽出するプロセス」

- データマイニング分野の創始者の１人であるU. Fayyadによる定義
- 「・・・理解できる知識」→最終的には「人間が使う」
- 計算機（機械学習）が抽出した規則性やパタンが「使える」ものであって、はじめて「データマイニング」

データマイニングプロセス



データマイニングの歴史



主な適用事例①

金融	住宅ローンの潜在顧客発掘 ダイレクトメールの送付先顧客の発掘 クレジットカードの不正利用パタン推定 社債格付け推測
流通・小売り	消費者行動パターンの分析 消費者の併売パタンの分析（リコメンデーションシステムへの応用） オークションサイトでの不正出品検出
製造	顧客クレーム情報に基づく設計・製造時の品質改善 顧客意見情報の分析に基づく新製品開発

（参考：元田他，データマイニングの基礎，オーム社，2006）

主な適用事例②

通信	網管理のための負荷状況把握・障害診断 電話網マーケティングのための通信トラフィック分析 アクセスログ分析に基づく不正アクセス検出 Spamメールフィルタリング
運輸	ヒヤリハット報告に基づくリスクマイニング
製薬・医療	医薬品安全性情報からの注目すべき副作用の抽出 化学化合物分子構造と生理活性の相関解析
スポーツ	バスケットボールの攻撃パターン分析 アメフトの攻撃パターン分析

（参考：元田他，データマイニングの基礎，オーム社，2006）

学習とは？

- データからデータが内包する特徴（規則性やパタン）モデルとして抽出すること
  - 規則性やパタンの表現→知識表現

教師つき学習と教師なし学習①

- 学習は、「教師つき学習」と「教師なし学習」に分けることができる
- 教師つき学習(Supervised-Learning)
  データ（事例）と抽出されるべき規則性やパタンが「解答」として与えられており、より正しく解答できるモデルを抽出する
- 教師なし学習(Unsupervised-Learning)
  データ（事例）のみが与えられており、アルゴリズムが持つ尺度（評価基準、ものさし）に従い、そこに含まれる規則性やパタンをモデルとして抽出する

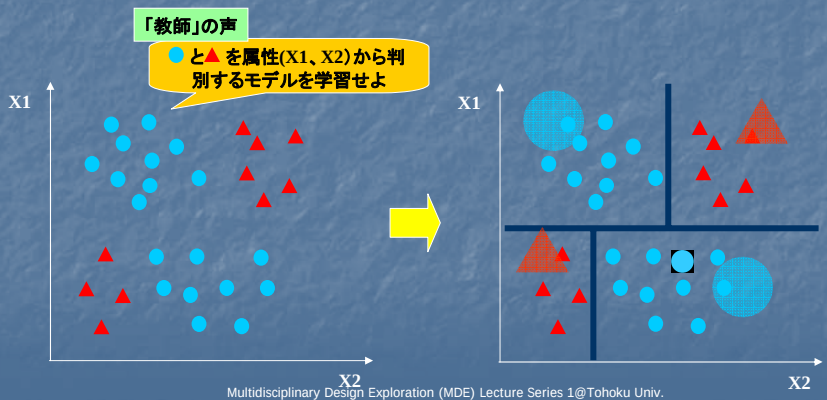
教師つき学習と教師なし学習②

（例）Webニュース記事を分類する

- 教師つき学習
  - ニュース記事と分類されるべきクラス（カテゴリ）が準備されている
  - 与えられたニュース記事が正しいクラス（政治、スポーツ・・・）に分類されるように学習する
- 教師なし学習
  - ニュース記事がデータとして準備されている
  - 記事に含まれる単語の内容等に基づき、記事間の類似度を評価し、類似した記事の群を抽出する（結果として、政治に関係する記事の集合が抽出される）

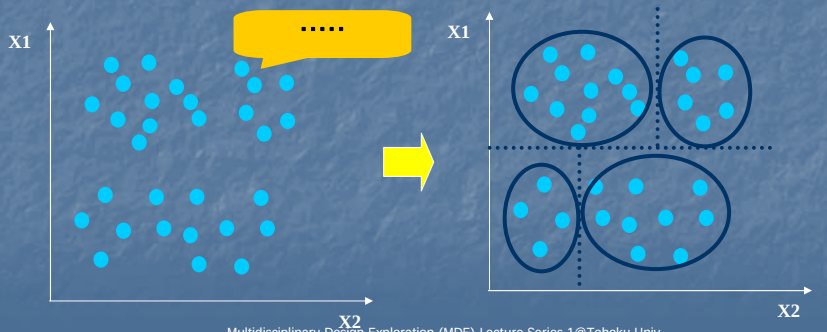
教師つき学習と教師なし学習③

- 教師つき学習：データに、「予測したい内容」（教師、クラス）が含まれる
  - 教えられた事例を正しく分類できるモデルを学習する



教師つき学習と教師なし学習④

- 教師なし学習：訓練データに、「予測したい内容」（教師）が含まれない
  - 類似した（距離が近い）データをまとめる



データマイニング手法

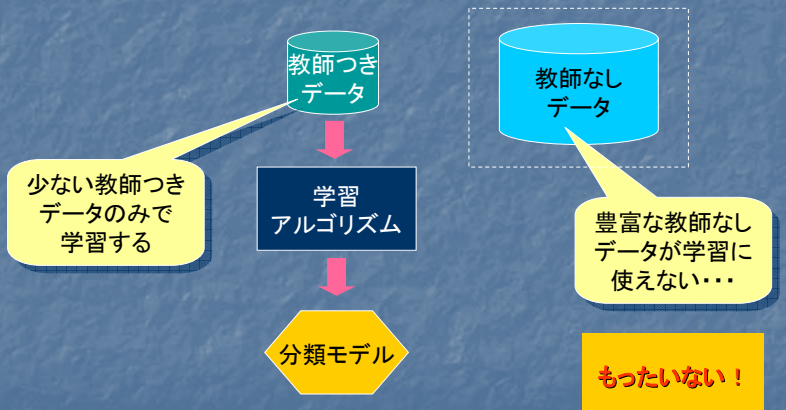
タスク		代表的な手法	教師つき／教師なし
予測(Prediction)	事例の特徴から、値を予測する	▪回帰木 ▪ニューラルネットワーク ▪SVM (Support Vector Machine)	教師つき
分類(Classification)	事例の特徴から、分類されるべきクラスを予測する	▪決定木 ▪ニューラルネットワーク ▪SVM	教師つき
相関(Association)	事例の中で頻度高く共起している特徴を抽出する	▪相関ルール	教師なし
クラスタリング(Clustering)	事例のうち類似したものをクラスタにまとめる	▪K-means ▪ニューラルネット（自己組織化マップ）	教師なし

2. データマイニングを使う

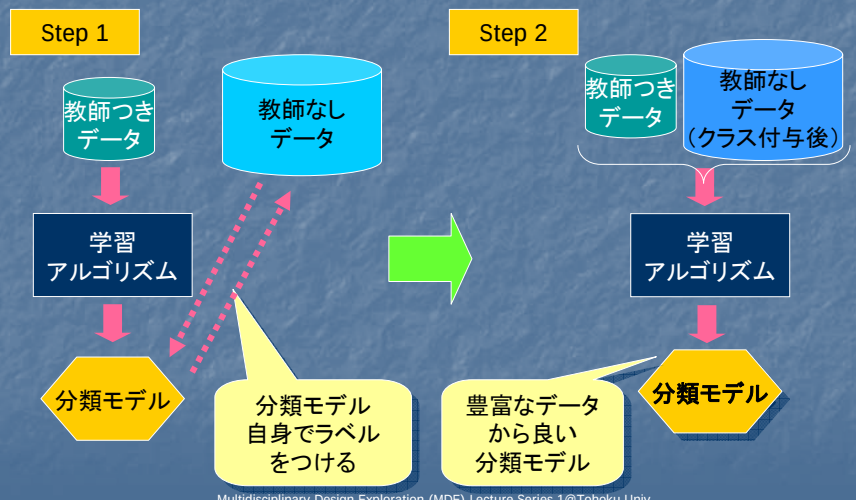
少ない教師つきデータから学習する ～半教師つき学習～①

- 分類モデルを作りたい！
  - 分類学習を行う上では、教師つきデータが多く必要
  - 教師つきデータを作るにはコスト（手間）がかかる
- 例：Webページの分類
  - 教師つきデータ：Webページ＋分類されるべきカテゴリ
  - 教師なしデータ：Webページ
- 教師つきデータを準備するのはコストがかかるが、教師なしデータは豊富にある
- 分類学習に豊富な教師なしデータを活用：
  半教師つき学習(Semi-Supervised Learning)

少ない教師つきデータから学習する ～半教師つき学習～②



少ない教師つきデータから学習する ～半教師つき学習～③



少ない教師つきデータから学習する ～半教師つき学習～④

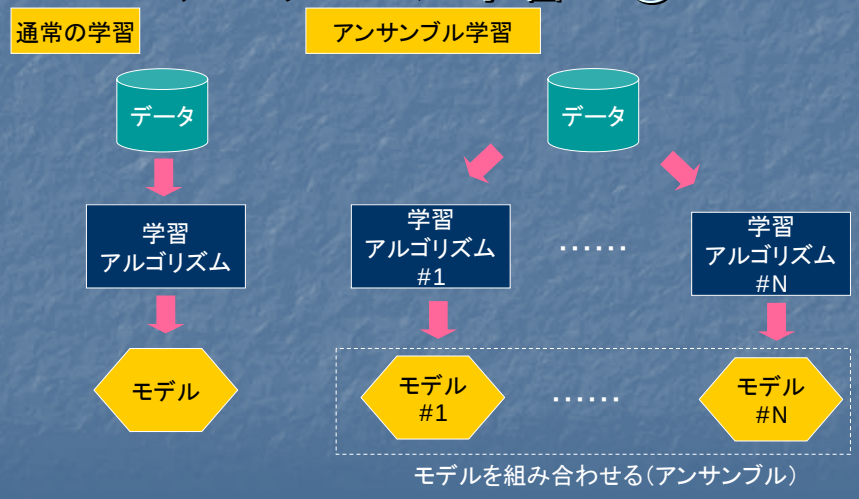
- 半教師つき学習
  人間の専門家が正解を付与した少ない教師つきデータと多数の教師なしデータから効率よく精度のよい学習をする方法
- 効率を向上させるための手段
  能動学習を取り入れ、学習アルゴリズムが正解が分からない／自信を持てないものだけに人間が正解を付与する
- 代表的な手法
  - Co-Training/Tri-Training/Democratic-Learning
  - Co-Testing　（Tri-Testing)

3人寄れば文殊の知恵 ～アンサンブル学習～①

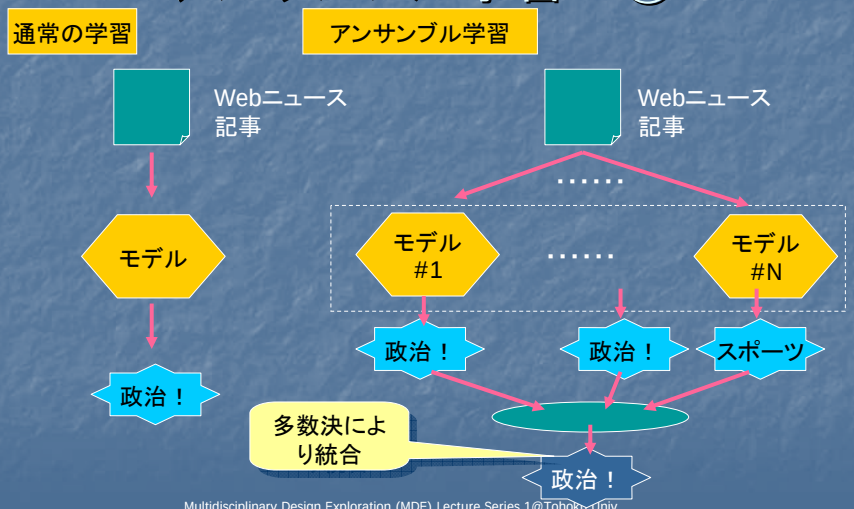
- 単独のモデルによる予測精度の改善には限界がある
- 人間は、複数の人の意思決定を組み合わせることにより、意思決定の質を高めている（3人寄れば文殊の知恵）

- アンサンブル（コミッティ）学習
  複数のモデルを抽出し、それらを組合せ（アンサンブル）、予測結果を統合することを試みる

3人寄れば文殊の知恵 ～アンサンブル学習～②

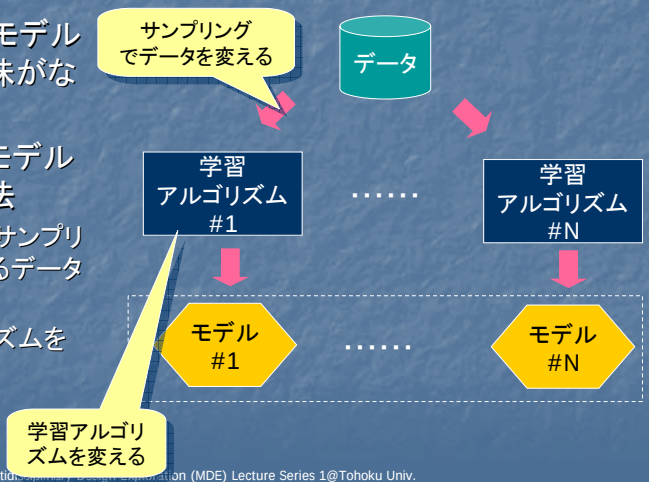


3人寄れば文殊の知恵 ～アンサンブル学習～③



3人寄れば文殊の知恵 ～アンサンブル学習～④

- 組み合わせるモデルが同じでは意味がない
- 互いに異なるモデルを導出する方法
  - 元データからサンプリングして与えるデータを変える
  - 学習アルゴリズムを変える



３人寄れば文殊の知恵 ～アンサンブル学習～⑤

- アンサンブル学習
  多様な知識／視点（モデル）を組み合わせることにより、予測精度をあげる手法
- 多様な視点を産み出す手段
  - データを変える（分析の情報源を変える）
  - 学習アルゴリズムを変える（分析方法を変える）
- 代表的な方法
  - Boosting
  - Bagging
  - Random Forest

文章をマイニングする ～テキストマイニング～①

- 通常扱われるデータ
  構造化されたデータ

outlook	temprature	humidity	windy	play
sunny	85	85	FALSE	no
sunny	80	90	TRUE	no
overcast	83	86	FALSE	yes
rainy	70	96	FALSE	yes
rainy	68	80	FALSE	yes
rainy	65	70	TRUE	no
overcast	64	65	TRUE	yes
sunny	72	95	FALSE	no
sunny	69	70	FALSE	yes
rainy	75	80	FALSE	yes
sunny	75	70	TRUE	yes
overcast	72	90	TRUE	yes
overcast	81	75	FALSE	yes
rainy	71	91	TRUE	no

- 文章（テキストデータ）をマイニングしたい！
  テキストマイニング

文章をマイニングする ～テキストマイニング～②

- 文章をマイニングするためのアプローチ
  - 既存の学習アルゴリズムは、構造化データであれば利用できる

  - テキストデータを構造化データに変換する
  - テキストデータに対して形態素解析を行い、単語（と頻度）組のデータに変換する

文章をマイニングする ～テキストマイニング～③

- 形態素解析ソフト
  chasen (http://chasen.naist.jp/hiki/ChaSen/)
  mecab (http://mecab.sourceforge.net/)

ドンキ連続放火、無期懲役の判決 さいたま地裁
2007年03月23日10時10分
　04年12月にさいたま市緑区の「ドン・キホーテ浦和花月店」で3人が焼死した火災など7件の連続放火事件で現住建造物等放火などの罪に問われた、同市中央区大戸6丁目、無職渡辺ノリ子被告（49）の判決公判が、23日、さいたま地裁であった。飯田喜信裁判長は求刑通り、無期懲役を言い渡した。
　起訴状によると、渡辺被告は04年12月13、15日に、さいたま市の「ドン・キホーテ」や「サティ」など大型量販店4カ所で、トイレの個室や売り場の
……

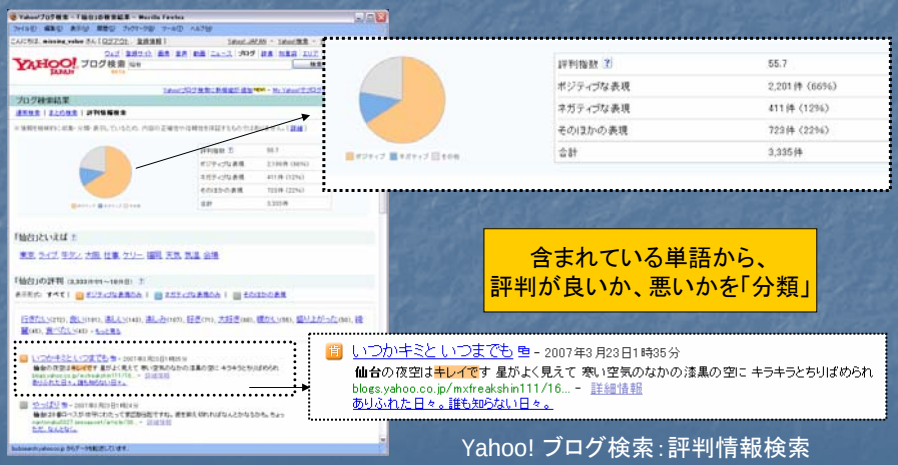
テキストデータ

単語	頻度
放火	7
無期	3
地裁	3
被告	3
…	…

単語頻度データ

一般のデータマイニングへ

文章をマイニングする
～テキストマイニング～④



Yahoo! ブログ検索：評判情報検索
（http://blog-search.yahoo.co.jp/）

その他のトピック

- 時間経過にともない変化する対象をモデル化する
  Concept-Drift (Concept-Change)
- 逐次提供されるデータから学習する
  Data stream、On-Line学習(cf. Batch学習)
- グラフ構造のパターンを抽出する
  Graph Mining

3. さらにデータマイニングについて知るために

データマイニングの情報源 ～ツール～

- WEKA
  - http://www.cs.waikato.ac.nz/ml/weka/
- MUSASHI
  - http://musashi.sourceforge.jp/

データマイニングの情報源 ～テキスト／解説記事～

- 新書
  - 岡崎、数式を使わないデータマイニング入門 隠れた法則を発見する、光文社、2006
- 書籍（入門書・教科書）
  - ベリー他、データマイニング手法―営業、マーケティング、CRMのための顧客分析、海文堂出版、2005
  - 元田他、データマイニングの基礎、オーム社、2006
  - I. H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann, 2005
- 論文誌特集
  - 計測と制御（計測自動制御学会誌）：特集 データマイニングの最前線, Vol.41, No.5, 2002
  - システム/制御/情報（システム制御情報学会誌）：データマイニング特集, Vol.46, No.4, 2002

データマイニングの情報源② ～国際会議～

- ACM KDD Conference
  - http://www.kdd2007.com/
- IEEE ICDM (International Conference on Data Mining)
  - http://www.ist.unomaha.edu/icdm2007/
- SIAM Conference on Data Mining
  - http://www.siam.org/meetings/sdm07/
- PKDD (European Conference on Principles and Practice of Knowledge Discovery in Databases)
  - http://www.ecmlpkdd2007.org/
- PAKDD (Pacific-Asia Conference on Knowledge Discovery and Data Mining)
  - http://lamda.nju.edu.cn/conf/PAKDD07/

ご静聴ありがとうございました