データマイニングとは？
概論

2011年11月18日
北海道大学大学院 情報科学研究科
湊ERATO連携講座大学院特論科目（1）
集中講義「大規模離散計算科学特論」
客員教授　鷲尾隆（大阪大学産業科学研究所）

1

講義内容

- データマイニングとは？ ― 概論 ―
- グラフ同型・部分グラフ判定の原理
- 頻出部分グラフマイニングの原理と応用
- グラフ検索の原理と応用
- 統計的因果推論の原理と応用

2

概論の内容

- データマイニングとは？
  - データマイニング研究の歴史的変遷
  - データマイニングとは何をする技術か
  - データマイニングと他技術との関係
  - データマイニング技術の俯瞰
- 世界の基礎研究コミュニティ
- データマイニングツールと実用の現状

3

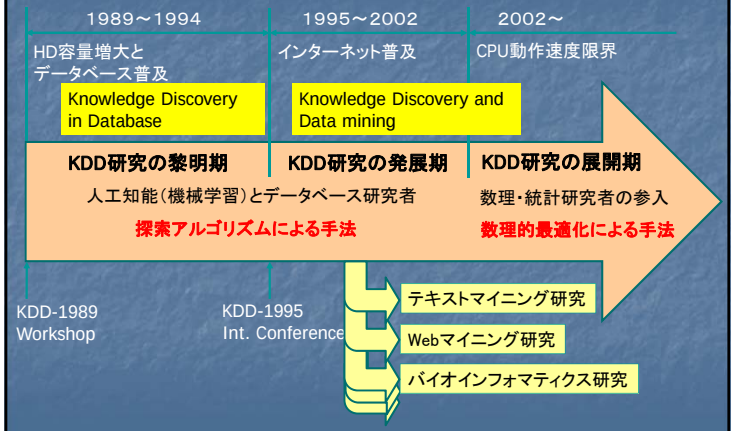
データマイニングとは？

- 人によって意味していることが違う。
  - 十数年くらい前から、書店にいくと、Excelでデータマイニングとか、いろんな本が並んでいる。でも中には相関解析、重回帰分析、判別分析など、従来の統計手法による分析方法を解説しているだけのものもある。。。。
  - データベースや検索エンジンで情報検索して、いろんな情報を見つけたり、傾向を調べることも行われている。。。。
- データマイニングは、統計分析や情報検索と変わらないのか？

- まずは，この素朴な疑問についてお話したい。
  - ただし、人によって意見が違う部分もあるので、一部私見が入る。
  - 日進月歩で技術が発展し、その概念自体が急速に変化している.

4

データマイニング研究の歴史的変遷



データマイニング研究の歴史的変遷 2つのKDD

- **KDD研究の黎明期（1989－1994）**
  - Knowledge Discovery in Database
  - 1989年　世界ではじめてのKDD(Knowledge Discovery in Database)−1989ワークショップ会議開催@デトロイト
    ～1994年　KDD−1994まで4回のワークショップ会議開催
    - 主に人工知能（機械学習）研究者とデータベース研究者が一緒に主催し，研究分野として確立した．
      発端となった研究動機：　計算機に蓄積される膨大だけど雑多（不均一）なデータから何か有用な知識を見つけることはできないか？
    - 統計研究者は，それまで不均一なデータはあまり相手にしなかった。膨大でも均一なデータなら従来の統計解析手法で十分と考えた（？）

データマイニング研究の歴史的変遷 2つのKDD

- **KDD研究の発展期（1995－2002）**
  - Knowledge Discovery and Data mining
  - 1995年　インターネット元年と言われる年
    第1回KDD(Knowledge Discovery and Data mining)国際会議開催
    - はじめて**公式に？データマイニングという言葉が使われる。**
    - データベースだけでなく、ネットワークを含めより広く膨大で不均一なデータから有用な知識を**掘り出す方法を研究**する分野と位置づけられた。
  - ～2002年の第9回KDD国際会議くらいまでは，引き続き人工知能（機械学習）とデータベース研究者による**探索アルゴリズム手法が中心**。
    - この間に，インターネットの爆発的普及にも後押しされブームの到来！
      他学会のICDM, SDM, 欧州中心のPKDD，アジア太平洋中心のPAKDD国際会議開催が開始され，機械学習中心のICML, ECML，データベース中心のSIG-MOD, VLDBなどの国際会議もデーマイニング研究を扱い出した。
      日本でも人工知能学会などを中心に研究や講演会、セミナーが広がる。
  - 発見より発掘するイメージなので、データマイニングという言葉が普及。

データマイニング研究の歴史的変遷 CPU動作速度限界と専門分野分化

- **KDD研究の展開期（2003～現在）**
  - 数理・統計研究者の参入
    - 対象データはより膨大でより複雑化，CPU動作速度の向上は頭打ち
    - 一般に計算コストの大きい探索アルゴリズムではなく，**数理的最適化による手法**の研究へと展開
    - 数理モデル中心のNIPS，統計手法中心のUAIなどの国際会議もデーマイニング研究を扱い出した。
  - テキストマイニング研究への分化
    - 自然言語処理研究にデータマイニングが取り入れられ，テキストマイニング研究分野が確立し，独自の発展を遂げる。
  - Webマイニング研究への分化
    - Webページランキングなど Web検索研究にデータマイニングが取り入れられ，Webマイニング研究分野が確立し，独自の発展を遂げる。
  - バイオインフォマティクス研究への分化
    - 遺伝子配列情報やたんぱく質合成過程の分析研究にデータマイニングが取り入れられ，独自の発展を遂げる。

データマイニングとは？（その2）

- 現実の**不均一な（ムラのある）**膨大なデータ
  - 企業やネットワークのデータなどは、様々な条件や対象のデータ。
    - 世論調査：都市と地方、関東と関西では意見も違う。。。
    - 商品売上：顧客年齢や性別、地域で購入商品や購入金額が違う。。。
  - 多くの統計手法のデータ分布均一性の仮定が成立しない。
    - 標本抽出して解析、全データを1つの統計分布で分析など。。。
    - 一様に砂鉄が含まれる砂浜。

- 実際に必要なデータ傾向分析
  - 異なる傾向のデータに分け、それぞれ傾向分析。
  - 場所によって鉄鉱脈があったりなかったり。あたかも**発掘現場**。
- 1996年 U. Fayadd: Data mining
  - **膨大で不均一なデータから有用な知識を掘り出す方法の研究**
    - データをある程度均質な分布の**セグメントに分解**
    - 各セグメントについて適切な**モデリングや解析**

9

データマイニングとは？（その3）

- ただし、分解されたセグメントから分析によって分かる傾向は、見掛け上のものかも知れない。
  - 統計誤差：セグメントデータが少な過ぎ。
  - 系統誤差（バイアス）：セグメントデータが特定の性質に偏り過ぎ。

データマイニング
- データから最終結論を知識として取り出すとは限らない。
- 多くの場合、**可能性のある仮説を知識**として発掘する。

- データマイニングは、多くの場合、他の仮説検証方法と組み合わせて用いられる。
  - 対象セグメントに関してもっとデータを集めて統計精度を高める。
  - 実験や実地調査を行って、本当かどうかを直接確かめる。

10

データマイニングと情報検索，統計との関係

情報検索
要素技術：データベースやネットワーク，検索（サーチ）
機能内容：膨大なデータから欲しい**個別知識**を探し出す

データマイニング
要素技術：機械学習，データマイニング固有技術，統計
機能内容：膨大なデータに隠されているが役立ちそうな**部分的傾向知識**を発見する

統計解析
要素技術：確率・統計学，統計的手法
機能内容：データが表す**全体傾向知識**を把握する

11

知識発見とデータマイニング
データマイニングの多様性

- 知識発見：
  - あくまでも導出結果から，人間が解釈して知識として重要なものを選び取ることが目的．
    通常，統計的有意性検定を課さないで，粗いフィルタリングに留める．
    - 統計的有意性と人間の解釈可能性が必ずしも一致するとは限らない．
    - 人間がデータを追加収集し検討・解析したり，他の実験や専門的知見によって検証する余地を残す．
    - 人間が必要に応じて統計的有意性検定を実施．

12

知識発見とデータマイニング
データマイニングの多様性

- 知識発見：
  - データのセグメント化の妥当性と発見すべき部分的知識の妥当性はニワトリと卵の関係．人間の解釈も含めた探索的な検証が必要．
  - データマイニングが人間に提供するのは，人間が処理しきれない膨大・複雑なデータからの網羅的な知識の候補や仮説の導出．最終的な知識発見は人間に委ねられる．
  - 解析対象データや導出規則を人間に分かりやすく表示するヒューマンインターフェイスも，知識発見の重要な研究分野の1つである．

13

知識発見とデータマイニング
データマイニングの多様性

- 最適予測：
  - 統計的手法による予測と同様，予測結果について，各種精度指標による性能評価やクロスヴァリデーションによる予測の妥当性，安定性検証などが行われる．
  - 予測手法の性質上，過学習（オーバーフィッティング）を生じない場合には，クロスヴァリデーションを省略する場合もある．

14

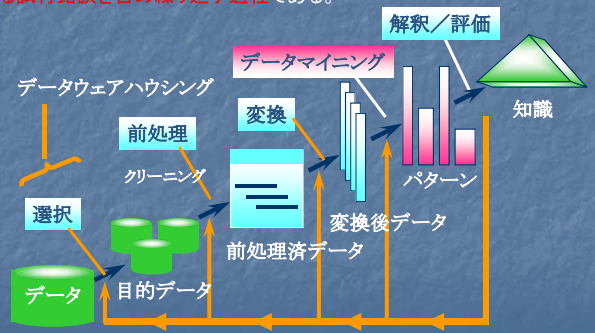
知識発見とデータマイニング
データマイニングの多様性

- 対象データ：
  - データ数：
    - 数千から数千万個までという広い範囲にまたがる．
  - データの形式：
    - 数値，ノミナルな記号，天体写真のような画像，ホームページのような自然文やタグ・リンク付きハイパーテキスト，遺伝子配列のような系列，半順序を表す木構造，化学式のようなグラフなど
  - データ品質：
    - ノイズの多寡，欠測値がの多寡，実験や調査の制約上の不可避的に強いデータバイアス，解析目的が遂行可能な品質や内容のデータかどうかは事前に分からないことが多い．
  - データ前処理：
    - 型通りの手法に適用できるように多大な労力を要する．形式変換，データ項目の取捨選択，欠測値の補完（人間が介在する場合もある），他データからの新たなデータ項目生成など．

15

データマイニングとは？（その4）

- 多くのデータマイニング技術はデータ解析者を支援のためのものであり、解析を自動化するためのものではない。
- 目的に応じた知識が得られるまで、様々なデータ処理や解析を人手による試行錯誤を含め繰り返す過程である。



データマイニングの解析過程 [U.Fayyad(1996)]

16

データマイニングとは？（その5）

- データマイニングは一種の情報総合工学
  - 数千種類を超える前処理、変換、解析、解釈・評価手法の総称
  - 機械学習（人工知能）、統計数理、データベース、情報検索など、種々の情報処理技術から成り立つ。
- 最近では、技術内容を益々膨らませている。
  - テキストマイニング
  - Webマイニング
  - バイオインフォマティクス
  - パターン認識
  - 信号処理関連など

17

データマイニング技術の俯瞰（その1）

データ解析技術（殆どが完全自動解析ではなく、人間の介在が必要。）

- **分類技術**：データ中の各事例を分類する規則や数式を見出す。
  決定木系技術、ニューラルネット系技術、ベイジアンネット系技術、統計的判別分析系技術、 サポートベクターマシン系技術、エマージングパターンルール系技術、アンサンブル学習系技術など
- **クラスタリング技術**：データを似た事例のグループに仕分けする。
  k-means系技術、密度ベース系技術、樹状図系技術、ニューラルネット系技術、カーネル関数系など
- **バスケット分析技術**：事例に共起する条件を見出す。
  Apriori系技術、Pattern Growth系、LCM系技術、その他多数
- **構造パターン解析技術**：系列や木構造、グラフ構造データから頻出パターンを見出す。
  系列マイニング系、木構造マイニング系、グラフマイニング系、カーネル関数系など
- **統計的関数分析技術**：データ項目間の関係やその中で主要な関係を見出す。
  重回帰分析系、ロジスティック回帰系主成分・因子分析系、独立成分分析系など

各々の技術について、代表的なものだけで数百種類ある。

18

データマイニング技術の俯瞰（その2）

- **データ前処理技術**（殆どが完全自動解析ではなく、人間の介在が必要。）
  - 属性選択技術
    - データから解析目的に必要性の高い項目属性を選択する。
  - 属性生成技術
    - データから解析目的に必要性の高い項目属性を合成する。
  - 事例選択技術
    - データから解析目的に必要性の高い事例を選択する。
  - 事例生成技術
    - データから解析目的に必要性の高い事例を合成する。
  - 数値データ離散化技術
    - 数値データを解析目的に必要な記号データに離散化する。
  - その他（細かな処理技術は多数）
- **データ可視化技術**：複雑な発掘結果を人間に分かりやすく表示。

上記各技術について、代表的なものだけで数十～数百種類ある。

19

データマイニング技術の俯瞰（その3）

- データマイニング技術があまりに多岐なため、百科事典ともいうべきハンドブックが多数刊行されている。
  - Handbook of Data Mining and Knowledge Discovery, by Willi Klosgen, Jan M. Zytkow, and Jan Zytkow, Morgan Kaufmann, New York (2002)
    - 世界初のハンドブック（私も記事を寄稿）
    - これ以降、多数のハンドブックが出版。
  - Data Mining and Knowledge Discovery Handbook, by Oded Maimon and Lior Rokach, Morgan Kaufmann, New York (2005)
  - Handbook of Statistics, Volume 24: Data Mining and Data Visualization (Handbook of Statistics) by C.R. Rao, E. J. Wegman, and J. L. Solka, Morgan Kaufmann, New York (2005)
  - Handbook of Data Mining, by Sanjay Ranka, Chapman & Hall/Crc Computer & Information Science Series (2007)
  - The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, by Ronen Feldman and James Sanger (2006)

これら分厚いハンドブックでも一部しかカバーできない。
データマイニング手法のWikipedia作成が国際的に進みつつある。

20

用語による混乱の問題

- 各学問領域固有の用語の問題
  データマイニングにも固有の用語が多数.
  - 属性　　:特徴量や説明変数
  - クラス　:目的変数
  - 支持度　:出現頻度(確率と頻度の違いは微妙),　. . . . . .
- 外来語の和訳の問題
  - AssociationとCorrelation:相関
    - Correlation:物事の間に存在する関係
      物理的関係:Aが起こればBが起こる, Aが起きなければBも起きない.
    - Association:記憶や想像の中で何らかの繋がりをもつこと
      共起の観察関係:
      AとBは共に起き易い. Aが起きなくてもBが起きる可能性はある.
- 結果の取り扱いの問題
  - 一般にデータマイニングで得られる規則や傾向はあくまで仮説.
    客観的事実かどうか, 現実に当たって検証(統計的検証も含む)
    する必要がある.

21

概論の内容

- データマイニングとは?
  - データマイニング研究の歴史的変遷
  - データマイニングとは何をする技術か
  - データマイニングと他技術との関係
  - データマイニング技術の俯瞰
- 世界の基礎研究コミュニティ
- データマイニングツールと実用の現状

22

世界的基礎研究の中心コミュニティ(その1)

- データマイニング基礎研究の中心は主要国際会議
- データマイニング中心
  - SIG-KDD:米国ACM主催, データマイニング分野で頂点の国際会議
  - ICDM:米国IEEE主催, SIG-KDDに並びデータマイニング分野の頂点
  - SDM:米国SIAM主催, 上記2つほどではないが, 発表論文は広く読まれる。
  - (PAKDD:アジア系, 比較的高レベルの論文が集まる。)
- 機械学習中心
  - ICML:米国系, 機械学習分野で頂点の国際会議
  - PKDD/ECML:欧州系, 上記ほどではないが非常に高レベルの論文が集まる。
  - (ACML:アジア系, 今年から開催, 比較的高レベルの論文が集まる。)
- 統計・数理モデル中心
  - UAI:米国系, 確率統計ベースの人工知能分野で頂点の国際会議
  - NIPS:米国系, ニューラルネットや数理モデル分野で頂点の国際会議
  - AISTATS :米国・欧州系, 確率統計ベースの非常に高レベルの論文が集まる。

23

世界的基礎研究の中心コミュニティ(その2)

- データベース中心
  - SIG-MOD:米国ACM主催, データベース分野で頂点の国際会議
  - VLDB:米国系, SIG-KDDに並びデータベース分野で頂点の国際会議
- 情報検索中心
  - SIG-IR:米国ACM主催, 情報検索分野で頂点の国際会議
  - CIKM:米国ACM主催, 情報検索やデータマイニング分野で非常に高レベルの論文が集まる。
- 人工知能中心
  - IJCAI:米国系だが世界持ち回り開催, 人工知能分野で頂点の国際会議
  - AAAI:米国AAAI主催, IJCAIに並んで人工知能分野で頂点の国際会議

- その他、テキストマイニング、Webマイニング、バイオインフォマティクス、パターン認識など、各分野の頂点国際会議でも高レベルのデータマイニング論文が発表されている。

24

データマイニングツールの現状（1）

現状の市販ツールベンダー（OSでいうマイクロソフトのような商用版）

SAS, SPSS, 富士通，日立，日本IBM，日本SGI，
日本ユニシス，東芝，数理システム，スポットファイア

現状の市販ツールのサポート手法

決定木：ID3, C4.5, C5.0, CART, CHAID, QUEST,
Pseudo Decision Tree, Option Tree, ファジイ決定木，
2次元領域の抽出決定木，機能拡張決定木
ニューロ：BP, MLP, RBF, ベイジアン
クラスタリング：コホーネン，K-means，Ward法，
コンドルセ手法，概念クラスタリング
相関ルール：Apriori, Generalized Rule Induction,
順序アソシエーション，複数時系列
統計的手法：重回帰分析，ロジスティック回帰分析，
判別分析，主成分・因子分析
テキストマイニング，Concept Base Search，記憶ベース推論

25

データマイニングツールの現状（2）

オープンソースフリーウエア（OSでいうLinux）

- MUSASHI, KGMOD
  日本の大学関係者がコンソーシアムを組んで作成。情報系のみならず業務基幹系での使用も視野にいれ、膨大データの高速処理を行う。Linux OSのアークテクチャ上で構築。
- Weka
  ニュージーランドワイカト大学のチームが開発。業務用ではなく、データマイニング研究者やデータマイニング試用時の小規模検証向き。
- R
  元商用のS-Plusというデータマイニングソフトをベースにフリーウエア版にしたもの。中規模データまで扱える。データマイニングの教育向き。

オープンソースフリーウエアのサポート手法
若干インターフェースが落ちるが機能は商用版とほぼ同じ

- 決定木：ID3, C4.5, C5.0, CART,
- ニューロ：BP, MLP, RBF, ベイジアン
- クラスタリング：コホーネン，K-means法，概念クラスタリング
- 相関ルール：Apriori, Generalized Rule Induction, 複数時系列, グラフマイニング
- 統計的手法：重回帰分析，ロジスティック回帰分析，判別分析，主成分・因子分析
- テキストマイニング，Concept Base Search，記憶ベース推論

26

データマイニングの適用事例

金融分野

・マーケティング分野
潜在的な住宅ローン申し込み顧客の推定，
顧客に応じた銀行商品の適切な組み合わせ
（クロスセールス）の設計提示支援，
生命保険の潜在的解約候補顧客の発掘，
効果的なダイレクトメール宛先候補顧客の発掘
・業務特化分野
消費者ローン与信審査の半無人化ルールの発掘，
リスク細分型の自動車保険の設計提示支援，
証券顧客と営業マンとのトラブル予測，
社債格付け推測，クレジット・カードの不正利用パターン推定

流通・小売分野

薬局チェーン販売データからの優良顧客の発掘，
投入時立ち上がり売れ行きデータに基づく新製品販売予測，
新製品のヒット要因分析，品物の売れ行き要因分析，
牛乳販売量の予測，消費者購買行動パターンの分析，
種々の販促条件下における併売パターンの分析

27

データマイニングの適用事例（つづき）

製造分野

ホームページでの顧客意見収集による
次世代新製品開発（カスタマーリレーションマーケティング），
顧客の製品クレーム情報と製造情報の突き合わせによる
設計・製造現場への品質管理要求発掘
製造現場の製造条件と製品検査結果の突き合わせによる
製造工程の改善

通信分野

ホームページ閲覧情報からの個別顧客の
プロファイリングと顧客傾向分析，
電話回線網管理のための負荷状況把握や障害診断，
電話網使用需要マーケティングのための通信トラフィックデータ分析，
顧客の通話パターンによる通話回線不正使用検出，
計算機システムへのアクセスログに基づく不正アクセス検出

製薬・化学

化学物質の生理活性分析
薬品物質の副作用可能性分析

28