universität wien

Data Mining: Alchemie oder Wissenschaft?

Bernd Brandl
Institut für Wirtschaftssoziologie
Universität Wien

Sechste Österreichische TA-Konferenz
„Vermessen, codiert, entschlüsselt?“
Institut für Technikfolgenabschätzung
Österreichische Akademie der Wissenschaften

Motivation der Arbeit und Bemerkungen zum Titel

„Data Mining“ wird in den Sozial- und Wirtschaftswissenschaften kontrovers diskutiert

- Verfügbarkeit von „Daten“ hat zugenommen
- „Data Mining“ ist Ausdruck und Mittel diesem Phänomen zu begegnen
- Jedoch wird Data Mining in der Wissenschaft unterschiedlich „wahrgenommen“
- Wie wird mit sich ergebenden Chancen und Problemen umgegangen bzw. woher kommt die unterschiedliche Wahrnehmung von Data Mining
- **Ziel der Arbeit ist die „Wissenschaftlichkeit“ von der Verwendung von Data Mining herauszuarbeiten**
- Analogie zum „Auftauchen“ der computerunterstützten Ökonometrie:

  David Hendry (1980), Econometrics: Alchemy or Science?, Economica 47, 287-406

universität wien

Inhalte der Präsentation

I. Was ist Data Mining?
II. Data Mining in den Sozial- und Wirtschaftswissenschaften
III. Beispiele für die Möglichkeiten und Probleme von Data Mining
IV. Data Mining: Alchemie oder Wissenschaft?

universität wien

I. Was ist Data Mining?

I. Was ist Data Mining?

Versuch einer Definition

Data Mining „Gesellschaften“ definieren üblicherweise prozessorientiert:

„*The nontrivial extraction of implicit, previously unknown, and potentially useful information from data*“. (Frawley et al 1992)

„*The **science** of extracting useful information from large data sets or databases*“. (Hand et al 2001)

Hinsichtlich der Methode ist Data Mining ein Sammelbegriff für Verfahren und Prozesse welche das systematische und (halb-) automatische Entdecken unbekannter Informationen aus (großen Mengen von) Daten ermöglichen.

I. Was ist Data Mining?

Versuch einer allgemeinen Definition

- Data Mining Verfahren sind üblicherweise der *explorativen Datenanalyse* zugeordnet
  - Ziel der explorativen Datenanalyse ist über die Darstellung der Daten hinaus die *Suche nach Strukturen und Besonderheiten*.
  - *Sie wird daher typischerweise eingesetzt, wenn die theoretischen Zusammenhänge nicht präzise genug sind oder die Wahl eines geeigneten (statistischen) Modells unklar ist*.

- Data Mining kann somit als „Ergänzung“ oder „empirische Präzisierung“ zu „theoretischen Methoden“ gesehen.
- Data Mining in der prozessorientierten Definition beschränkt die Anwendungsgebiete auf keine Wissenschaftsrichtung.
- **Entsprechend dieser Sichtweise ist der Begriff Data Mining positiv belegt.**

II. Data Mining in den Sozial- und Wirtschaftswissenschaften

II. Data Mining in den Sozial- und Wirtschaftswissenschaften

Abweichende Verwendung des Begriffs „Data Mining“ mit negativen Assoziationen!

Versuch einer Definition des Begriffs in den Sozial- und Wirtschaftswissenschaften:

- Data Mining beschreibt Aktivitäten welche ausserhalb einer „traditionellen“ Modellbildung stehen, d.h. nicht theoriegeleitet sind. (Hoover/Perez 2000)
- Data Mining beschreibt das Anpassen von mehr als einer Spezifikation der gleichen Hypothese. (Mayer 2000)

- Zusammenfassend wird Data Mining in den Sozial und Wirtschaftswissenschaften üblicherweise als die wiederholte Analyse von (gleichen) Daten bezeichnet.

- D.h. die Definition erfolgt auch prozessorientiert.

- Synonyme in den empirischen Sozial- und Wirtschaftswissenschaften sind: „*data fishing*“ „*data peeking*”, “*data snooping*”, “*hunting with a shotgun*” oder “*hunting without a licence*” und ähnliche.
  - Diese Synonyme sind Ausdruck für die „Ziellosigkeit“, „Theorielosigkeit) und „Blindheit“ von Data Mining

II. Data Mining in den Sozial- und Wirtschaftswissenschaften

Was haben die unterschiedlichen Definitionen gemein:

- Beide Definitionen von Data Mining sind prozessorientiert und
- beziehen sich auf eine wiederholte Suche nach (empirischen) Zusammenhängen.

Welche Unterschiede gibt es:

- In den Sozial- und Wirtschaftswissenschaften konzentriert sich die Verwendung von Data Mining hauptsächlich auf die (Regressions-)modell bzw. Variablenselektion
- wohingehend die allgemeine Definition breiter ist und mehrere Methoden beinhaltet.

Der wesentliche Unterschied liegt jedoch nur in der Tatsache begründet wie die Sinnhaftigkeit der Ergebnisse gesehen wird!

Kann Data Mining auch sinnvoll sein in den Sozial- und Wirtschaftswissenschaften?

- Wenn die Suche nicht „blind“ und „ziellos“ ist und
- wenn die theoretischen Zusammenhänge nicht genau definiert sind oder die Wahl eines geeigneten (statistischen) Modells unklar ist.

Kann Data Mining als „Ergänzung“ oder „empirische Präzisierung“ zu „theoretischen Methoden“ gesehen werden?

II. Data Mining in den Sozial- und Wirtschaftswissenschaften

Ist Data Mining vermeidbar in der empirischen Forschung?

- Data Mining bzw. die wiederholte Analyse von Daten ist unvermeidbar!
  - Im Gegensatz zu anderen Wissenschaften ist es in den Sozial- und Wirtschaftswissenschaften unmöglich „Experimente“ zu wiederholen.
  - Als Folge werden „Daten“ wiederholt analysiert.
  - Üblicherweise werden mehrere „Überlegungen“ am gleichen Datensatz analysiert **um zu entdecken welche „Überlegung“ am besten ist.**

Jedoch was ist die „beste Überlegung“ bzw. woran wird der Erfolg gemessen?

III. Beispiele für die Möglichkeiten und Probleme von Data Mining

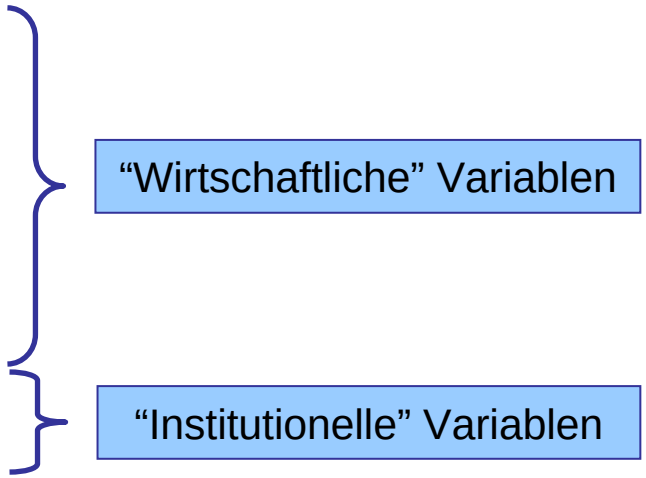
III. Ein Beispiel

- **Eine Fragestellung bei der die Theorie nicht präzis genug ist und die Wahl eines geeigneten (statistischen) Modells unklar ist.**
  - Kann Data Mining als „Ergänzung“ oder „empirische Präzisierung“ zu „theoretischen Methoden“ gesehen werden?

- **Was bestimmt (empirisch) das österreichische Wirtschaftswachstum und den Euro/Dollar Wechselkurs?**
  - Theoretische Zusammenhänge sind sehr allgemein formuliert
  - “Angewandte” Literatur bietet Fülle (sich teilweise widersprechende) von Einflussfaktoren die zu berücksichtigen sind
  - Empirische Literatur zeigt sehr unterschiedliche Ergebnisse

III. Ein Beispiel

- **Datengrundlage Österreichisches Wirtschaftswachstum**
  - Abhängige Variable: Österreichisches Wirtschaftswachstum (jährlich)
  - Länge der Zeitreihen: 32 Beobachtungen (1971-2002)
  - **49 unabhängige Variablen**
    - Unit labour costs
    - Wage rates
    - Exports
    - Imports
    - Gross financial liabilities (government)
    - Nominal and real exchange rate
    - Investment (private, government)
    - Employment
    - Inflation rate
    - Interest rates (long and short term)
    - Purchasing power parity, exchange rates
    - Number of strike days
    - Lefties in government
    - Union concentration
    - And more!



III. Ein Beispiel

- **Datengrundlage zukünftiger Euro/Dollar Wechselkurs (täglich)**

Data Mining zur Suche von:	Variablen
• Kombination von technischen und ökonomischen Variablen • Modellgröße • Die Länge der Zeitreihen	• 125 “ökonomische” Variablen • 12 “technische” Variablen • 110 verzögerte Variablen • = 247 Variablen

Größe des Suchraums

2^{86x247} mögliche Lösungen

III. Ein Beispiel

- **“Größe” des Data Mining Problems bei Österreichischem Wirtschaftswachstum**
  - 49 unabhängige Variablen
    - ~ 556.000.000.000.000 mögliche Lösungen
    - Unter der Annahme, dass die Berechnung einer einzigen Lösung eine Mikrosekunde dauert (d.h. ein Millionstel einer Sekunde) würde eine vollständige Enumeration

- **18 Jahre benötigen!**

III. Ein Beispiel

- **Was wird gesucht?**

  Was wird als „gutes“ Ergebnis gesehen?

  Wie wird die Güte ausgedrückt bzw. operationalisiert?

  Welche Kriterien sollen erfüllt werden?

  Womit können die Ergebnisse verglichen werden?

- **Worin wird gesucht?**

  Was ist die Datengrundlage?

  Wie ist die Qualität der Daten?

  Ist die Antwort in den verwendeten überhaupt Daten „versteckt“?

- **Wie wird gesucht?**

  Welche Data Mining Methode soll verwendet werden?

III. Ein Beispiel

Was wird als „gutes“ Ergebnis gesehen?

– **Beispiel Wirtschaftswachstum:**

Eine Spezifikation welche auf Basis verlässlicher Schätzergebnisse eine hohe Erklärung aufweist und dem Prinzip der „sparsamen Modellierung“ gerecht wird.

Ausgedrückt durch das BIC: $BIC = -2\hat{l}_\gamma / n + q_\gamma \ln(n)/n$

– **Beispiel Wechselkurse:**

Da multivariate Ansätze in der Beschreibung zukünftigen Wechselkursverhaltens eine geringe Erklärung haben wird versucht die vergangene und prognostizierte Anpassung zu erhöhen.

Ausgedrückt durch die mittlere absolute Abweichung:

$$MAE = \frac{1}{h+1}\sum_{t=S}^{S+h}\left|\hat{y}_t - y_t\right|$$

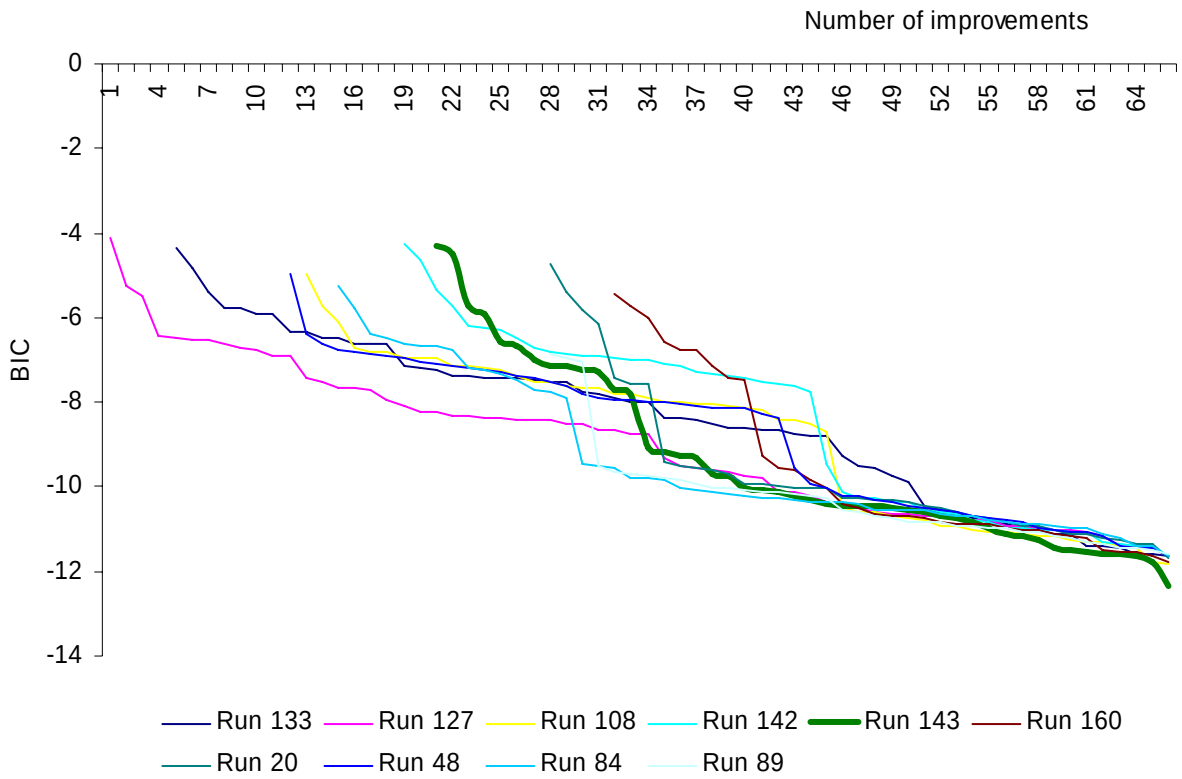
III. Ein Beispiel (österreichisches Wirtschaftswachstum)

Suche nach einer BIC optimalen Lösung:

- Variablen in der optimierten Regression
  - ULC: Unit Labour Costs
  - CPEI: Nominal Labour Cost Index
  - NEER: Nominal Effective Exchange Rate (Index)
  - EXPORT: Exports (Prices)
  - IMPORT: Imports (Prices)
  - OPEN: Foreign Trade Penetration (Prices)
  - MGSV: Imports (Volume)
  - XGSV: Exports (Volume)
  - OPENV: Foreign Trade Penetration (Volume)
  - WR: Wage Rate
  - WRMAN: Wage Rate Manufacturing
  - HOUR: Hours worked in Manufacturing
  - UC: Union Concentration
  - CURRBAL: Current Balances as percentage of GDP
  - PPP: Purchasing Power Parity

III. Ein Beispiel (österreichisches Wirtschaftswachstum)

Robustheit und Konvergenz: Prozessverlauf BIC (BIC Kriterium, 10 von 200 Läufen)



III. Ein Beispiel (Wechselkurs)

Variablen im Regressionsmodell sind nicht interpretierbar und sind zumeist nicht signifikant!

Der Anspruch einer Erklärung kann nicht gestellt werden

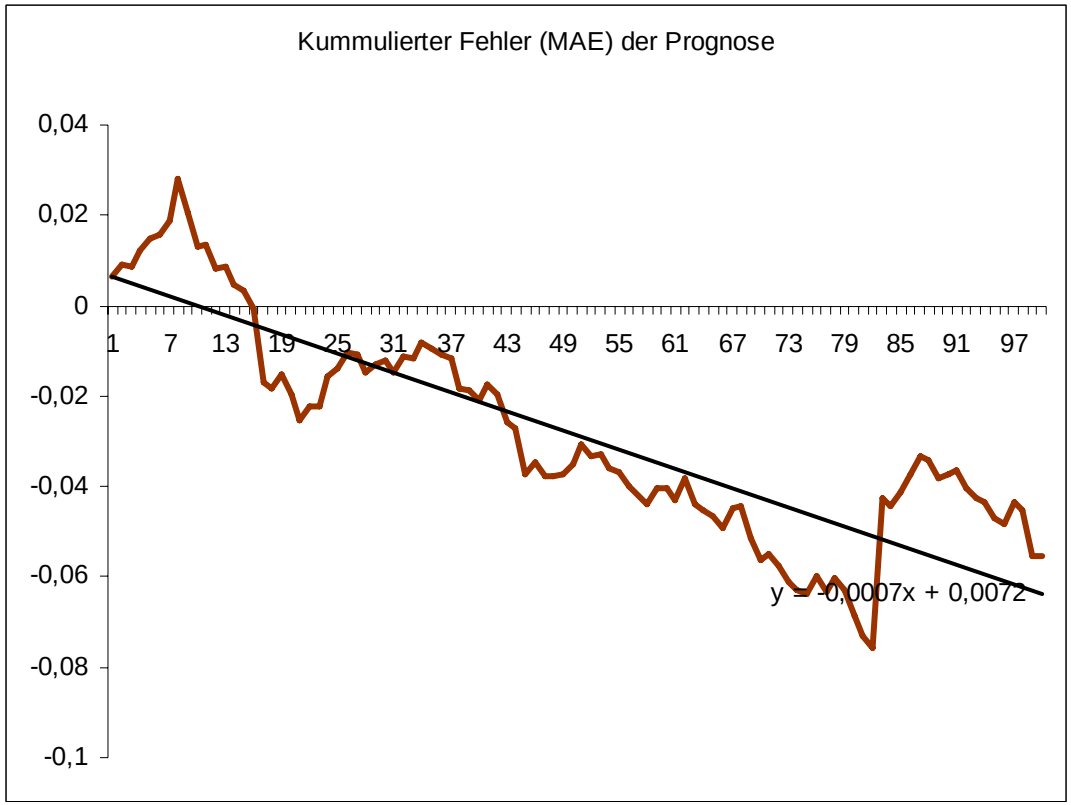
Erklärungsgehalt wird „künstlich“ erhöht

	WS	VAR#	R2ADJ	DW	MAE IS	MAE OS
Mean	156	21	0,2092	1,9982	0,0047	0,0047
Std Dev.	58	4	0,0290	0,0089	0,0002	0,0002
Max	296	28	0,2766	2,0235	0,0052	0,0052
Min	81	13	0,1379	1,9681	0,0045	0,0044

Wenn die Erklärung bzw. Interpretation nicht möglich ist, kann eine Prognose damit gemacht werden?

III. Ein Beispiel (Wechselkurs)

Prognosefähigkeiten der Data Mining Ergebnisse



IV. Data Mining: Alchemie oder Wissenschaft?

IV. Data Mining: Alchemie oder Wissenschaft?

Was ist Alchemie?

- Alter Zweig der Naturphilosophie.
  - Praktiziert beispielsweise im antiken Ägypten, Indien, China, Griechenland, mittelalterlichen Europa
  - Häufig wird Alchemie mit der Erzeugung von Gold und der Astrologie assoziiert
- Heute wird der Begriff als Synonym für **Pseudowissenschaft** und **Scharlatanerie** verwendet

IV. Data Mining: Alchemie oder Wissenschaft?

Was ist Wissenschaft?

„Wissenschaft ist eine Methode zum Wissenserwerb. Ziel der wissenschaftlichen Methode ist es, ausgehend von einer oder mehreren Hypothesen eine tragfähige Erklärung (Theorie) zu entwickeln.“

„*Wissenschaft besteht im Kern darin, auf methodisch kontrollierte Weise neue Kenntnisse und Erkenntnisse zu gewinnen […] die in prinzipiell allen Einzelheiten nachvollziehbar und überprüfbar sind*.“

IV. Data Mining: Alchemie oder Wissenschaft?

Erfüllt Data Mining die Kriterien der Wissenschaftlichkeit?

- Ist es möglich ausgehend von einer oder mehreren Hypothesen eine tragfähige Erklärung mit Hilfe von Data Mining zu entwickeln?
  - Wenn Hypothesen festgelegt werden – dies subsumiert die Definition von „gut“ und was gesucht wird
- Ist es möglich mit Data Mining auf methodisch kontrollierte Weise neue Kenntnisse und Erkenntnisse zu gewinnen die in prinzipiell allen Einzelheiten nachvollziehbar und überprüfbar sind
  - Natürlich möglich und betrifft insbesondere auch die Veröffentlichung aller Ergebnisse (im Prinzip auch der schlechten)

IV. Data Mining: Alchemie oder Wissenschaft?

Nachsatz

„Sehr häufig sucht man etwas und findet etwas anderes. Es genüge das Beispiel der Alchimisten, die Gold machen wollten und eine große Zahl chemischer Verbindungen entdeckten. […] *Aus diesem seltsamen und komplizierten Verfahren folgt die Entdeckung des Phosphors. In vielen anderen ähnlichen Fällen haben Kombinationen nichts Nützliches ergeben. Man tastet wie ein Blinder: Manchmal findet man, meistens findet man nicht.“*

Vilfredo Pareto, Trattato di sociologia generale, 1916; (dt. Allgemeine Soziologie, 2006)