TECHNISCHE UNIVERSITÄT DRESDEN

Institut für Angewandte Informatik
Professur für Technische Informationssysteme

Data-Mining und Knowledge Discovery in Databases (KDD) Ein Überblick

Pham Tien Hung

Dresden, 15.01.2009

Gliederung

1. Eigenschaften von KDD und Data-Mining
2. Modelle eines KDD-Prozesses
3. Schritte eines KDD-Prozesses
4. Anwendungsbeispiel: Fehlerdiagnose an einem Stellventil
5. Zusammenfassung

Eigenschaften von KDD und Data-Mining

Definition

- Knowledge Discovery in Databases (KDD) ist ein nichttrivialer mehrstufiger Prozess der Wissensfindung aus vorhandenen Informationen.
- KDD-Prozess umfasst alle Schritte, von woher die Daten abgeholt werden, über Vorverarbeitung und eigentliche Verarbeitung zur Informationsgewinnung (Data-Mining-Schritt), bis hin wie die Endinformation interpretiert und dargestellt wird.
- Data-Mining ist nur ein Schritt im KDD-Prozess, oft aber als Synonym verwendet.

Die Entwicklung

- KDD ist Ergebnis von verschiedenen Forschungsgebieten aus Informatik und Mathematik wie Datenbanken, Maschinellem Lernen, Soft Computing, Statistik, Datenvisualisierung und High-Performance-Computing.
- Durch technologischen Fortschritt Möglichkeit zur Speicherung großer Datenmengen, historische sowie gegenwärtig vorhandene.
- Interessen, aus diesen Daten nützliche, kompakte Informationen oder Erkenntnisse zu gewinnen.
- 1989: KDD erstmals erwähnt.
- 1996: mit Fayyad-Modell erstmals die einzelnen Schritte konkretisiert.

Herausforderung für KDD-Algorithmen

- Gigantische Datenbanken.
- Komplexe Beziehungen zwischen Feldern/Attributen.
- Fehlende Daten, falsche Daten und Rauschen.
- Über- oder Unteranpassung.
- Verständlichkeit der gewonnenen Muster/Wissen.
- Integration in andere Systeme.

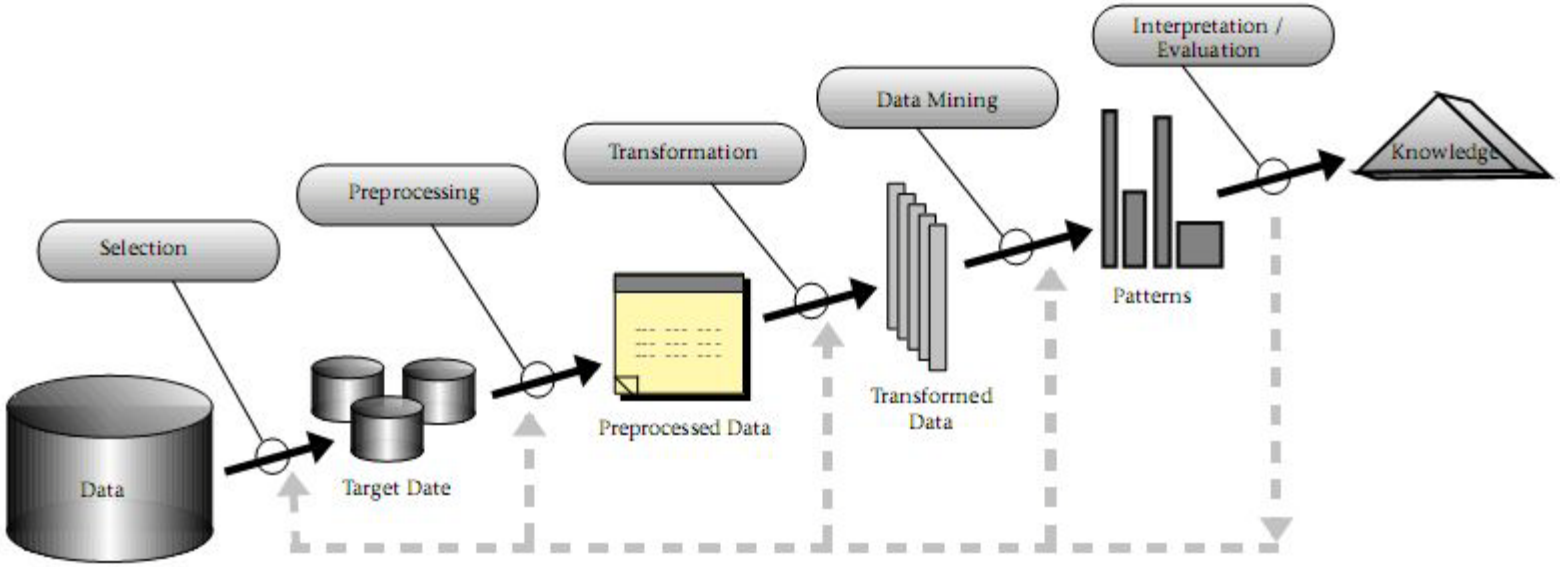
Einige Beispiele für Anwendungen

- In Wissenschaft: Erkentnisgewinnung als Basis zum Theorieaufbau:
  - Medizin: Patienten-Gesundheitsverlauf, und Diagnose.
  - Biologie: DNA- und Protein-Sequenzierung.
- In Wirtschaft: zum Vorhersagen, Entscheidungshilfe wie:
  - Marketing: Verhalten verschiedener Benutzergruppen.
  - Betrugsentdeckung: Kreditkartenbetrug, Geldwäsche...
- In Produktion: zur Diagnose und Vorhersage von Produktdefekten.

Gliederung

1. Eigenschaften von KDD und Data-Mining
2. Modelle eines KDD-Prozesses
3. Schritte eines KDD-Prozesses
4. Anwendungsbeispiel: Fehlerdiagnose an einem Stellventil
5. Zusammenfassung

Modelle eines KDD-Prozesses



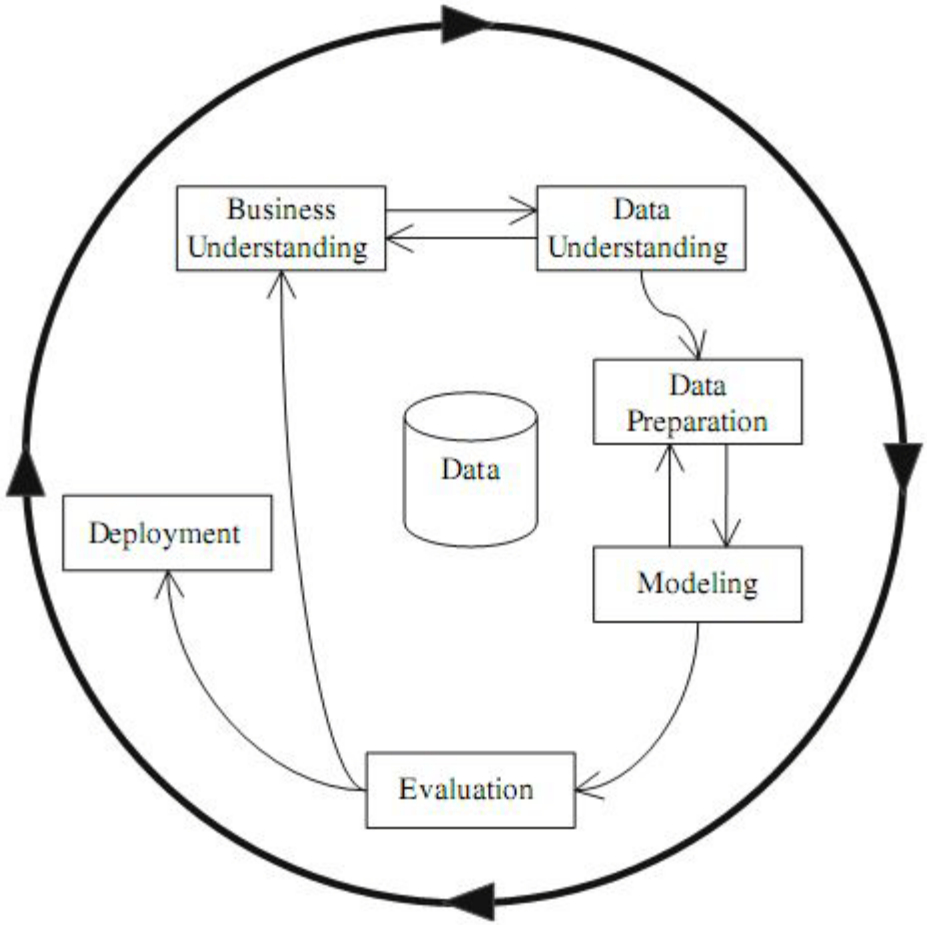
Fayyad Modell (Quelle: [1])

Erstes akademisches Modell (Fayyad et al.)

1. Anwendungsdomäne verstehen.
2. Daten für formulierte Ziele auswählen.
3. Daten bereinigen und vorverarbeiten.
4. Daten reduzieren und transformieren.
5. Geeignete Data-Mining-Methode auswählen.
6. Data-Mining-Schritt.
7. Gefundene Ergebnisse interpretieren.
8. Gefundenes Wissen anwenden.

Der Prozess ist iterativ und es können beliebige Sprünge zwischen den Schritten erfolgen.

Modelle eines KDD-Prozesses



CRISP-DM Modell (Quelle: [2])

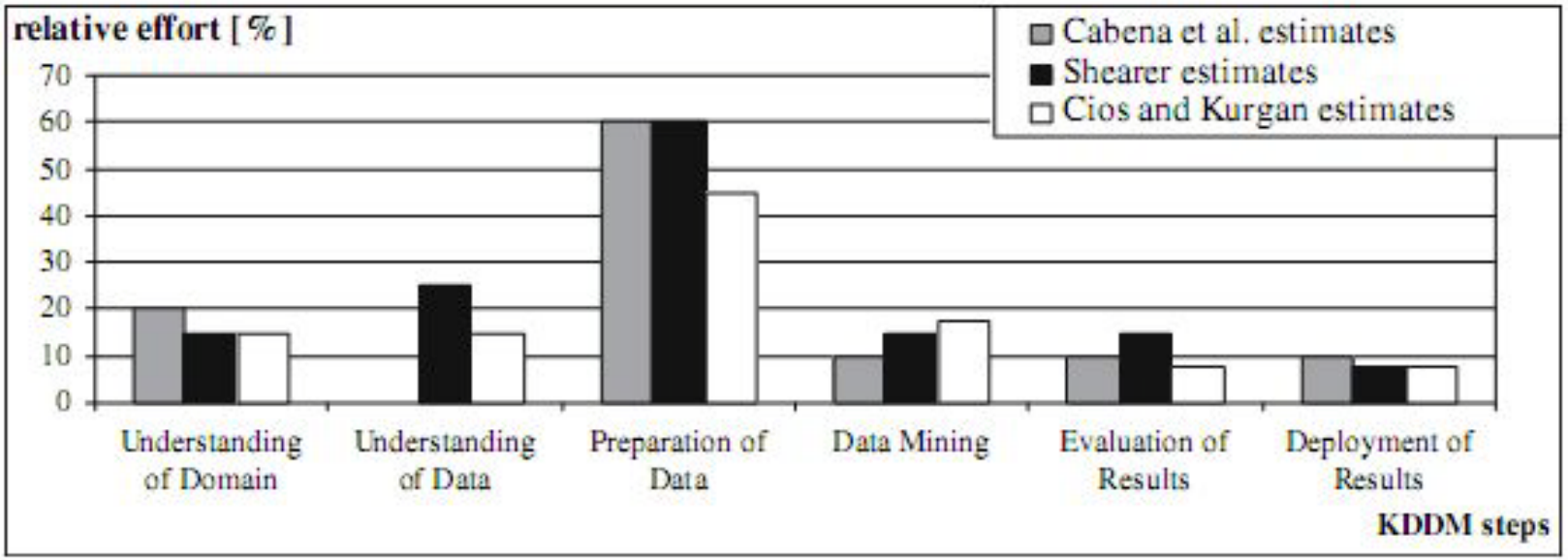
CRISP-DM Modell (Cross Industry Standard Process for Data Mining)

1. Anwendungsdomäne verstehen.
2. Daten verstehen.
3. Daten vorbereiten.
4. Data-Mining.
5. Modelle evaluieren.
6. Ergebnisse anwenden.

Schritte 3 und 4 umfassen die Schritte 3, 4 bzw. 5, 6 im Fayyad-Modell

Modelle eines KDD-Prozesses

Relativer Aufwand für Schritte eines KDD-Prozesses (Quelle: [2])



Gliederung

1. Eigenschaften von KDD und Data-Mining
2. Modelle eines KDD-Prozesses
3. Schritte eines KDD-Prozesses
4. Anwendungsbeispiel: Fehlerdiagnose an einem Stellventil
5. Zusammenfassung

Schritte eines KDD-Prozesses

A. Daten verstehen

- **Datentyp**

+ Statisch: nominal, ordinal, Set, Objekt.

+ Dynamisch: (unendliche) Stream-Daten, Zeitreihen...

- **Datenquellen**
  Datenbanken, Data Warehouse, WWW, Sensoren...

Schritte eines KDD-Prozesses

A. Daten verstehen

• **Datenmenge**
- Anzahl der Datensätze, Anzahl der Attribute, Anzahl der zulässigen Werte eines Attributs.

• **Datenqualität**
- Nichtvollständige, nicht exakte, überflüssige, verrauschte Daten.

• **Methoden zum Datenverstehen, z.B.**
- Statistische Auswertung der Daten (Minimum, Maximum, Mittelwert, Varianz, Ausreißerdetektion.
- Verschiedene Visualisierungen für Daten, Zusammenhänge (Scatter-, Verteilungs-, Korrelationsgraph...).

B. Daten vorbereiten

- Formatanpassung und Datensätze zusammenführen.
- Behandlung der fehlenden Daten, Extremwerte: weglassen, durch Mittelwert, alle möglichen, häufigsten Wert ersetzen...
- Daten normalisieren.
- Kontinuierliche Daten diskretisieren und/oder neu gruppieren, Histogramm, Clustering.
- Sampling-Methoden für Stichproben aus den Daten.
- Merkmalsextraktion und -selektion.

Schritte eines KDD-Prozesses

Klassifikation der Data-Mining-Methoden (Quelle: [5])

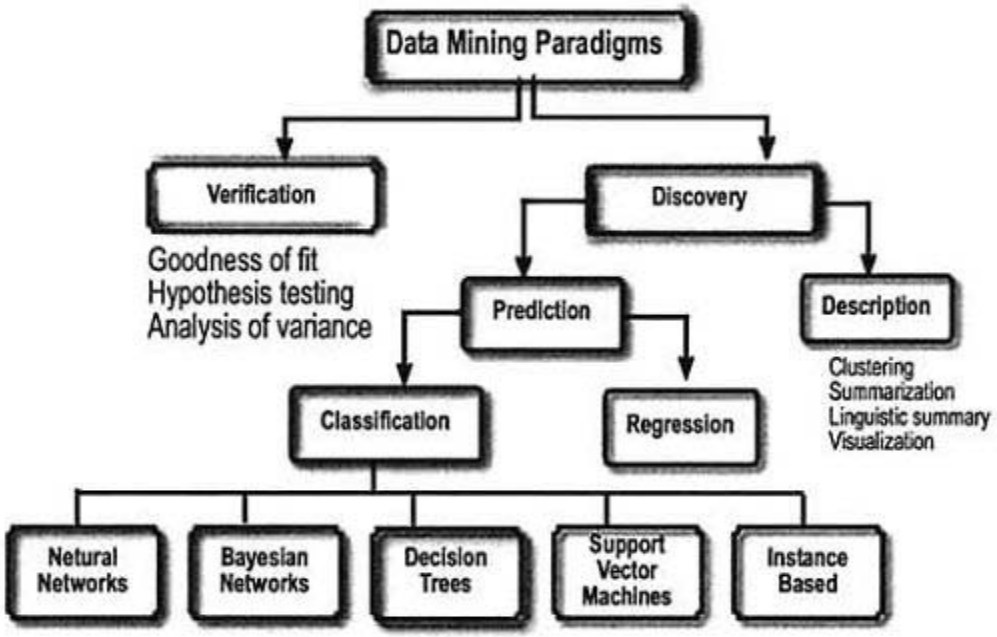


Figure 1.2. Data Mining Taxonomy.

- Verifikationsorientierte Methoden
- Entdeckungsorientierte Methoden (nach O. Maimon und L. Rokach)

C. Data-Mining

Clustering

Gruppierung ähnlicher Datenpunkte, z.B.:

- Hierarchisch: Bottom-Up und Top-Down.
- Partitionierung: K-Means, K-Medoids.

Statistische Methoden

- Bayessche Methoden:
  - Naive Bayessche Klassifikation.
  - Bayessche Glaubensnetzwerke.
- Regression: findet beste Anpassungskurven zu Datenpunkten z.B. Einfache, mehrfache lineare, nichtlineare Regression, logistische Regression...

C. Data-Mining

Neuronale Netzwerke

Nachbildung der Funktionsweise des menschl. Gehirns.
Typische Anwendungen z.B.:
- Klassifikation.
- Funktionsapproximation.
- Clustering.

Weitere Methoden

- Assoziationsregel.
- Entscheidungsbaum.
- Support Vector Machine.
- Lazy Learners.
- Text Mining.

Schritte eines KDD-Prozesses

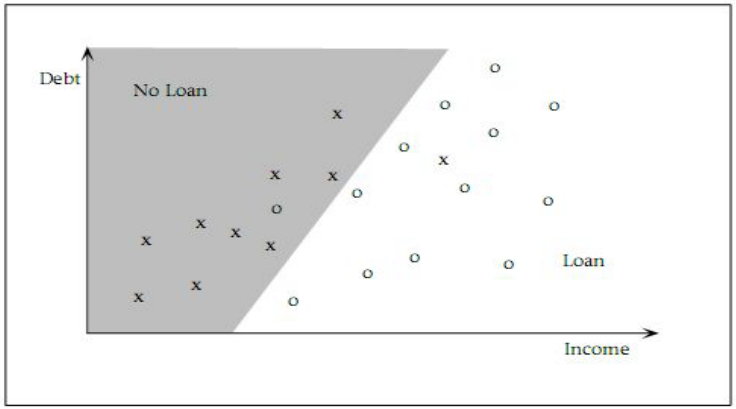


Figure 3. A Simple Linear Classification Boundary for the Loan Data Set.
The shaded region denotes class no loan

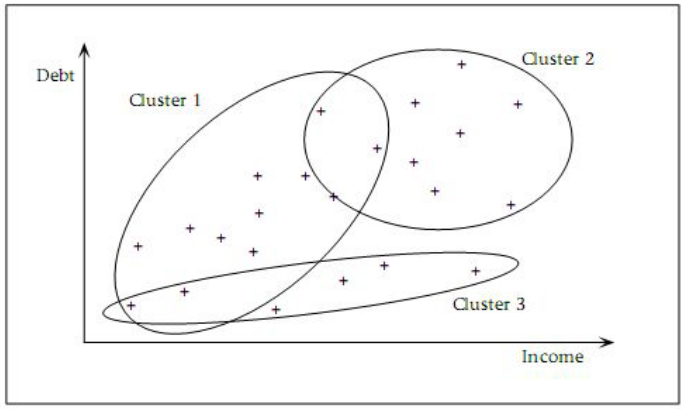


Figure 5. A Simple Clustering of the Loan Data Set into Three Clusters.

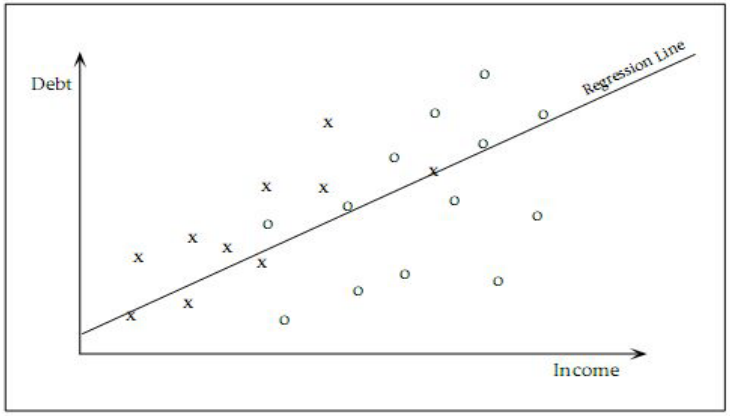


Figure 4. A Simple Linear Regression for the Loan Data Set.

- Klassifikation
- Clustering
- Regression

(nach Fayyad et al. Quelle: [1])

D. Datenmodell evaluieren

Methoden

• Resampling-Methoden: einfache Zweiteilung, Kreuzvalidierung, Bootstrap für überwachte Methoden.
• Sensitivität, Spezifität und ROC-Analyse (True Positiv/False Positiv).
• Kriterien für Interessantheits- und Abstandsmaße [2].
• Einfachste Modelle auswählen.

E. Unterstützende Methoden

Data-Mining-Abfragesprachen
Standardisiert DM-Schritte, unterstützt High-Level-Interaktion mit Endbenutzer, vereinfacht Integration mit anderen Systemen, z.B. CRISP-DM, OLE DB for DM, DMX.

Meta-Lernen zur automat. Modellevaluierung
Konzentriert auf Beziehung zwischen Aufgabe und Lernstrategien.
Struktur: 2 Modi
- Lernmodus: Datencharakteristik, Lernstrategien und Modellbewertung als Metawissen.
- Arbeitsmodus: Anwendung von Metawissen zur Auswahl der besten Modelle.

Gliederung

1. Eigenschaften von KDD und Data-Mining
2. Modelle eines KDD-Prozesses
3. Schritte eines KDD-Prozesses
4. Anwendungsbeispiel: Fehlerdiagnose an einem Stellventil
5. Zusammenfassung

Fehlerdiagnose an einem Stellventil [6]

1. Messsignalbestimmung und Klassifikation der Systemzustände

Art der Messwerterfassung:

- Mit zusätzlichem sprungförmigen Anregungssignal s.
- Gemessen: aktuelle Position x, Druck des Stellantriebes p, Regeldifferenz dx, Abtastrate r.

Die Dimension des Merkmalsvektors ist 23.

Anwendungsbeispiel

Bestimmung aller möglichen Zustände (Zustand = Klasse):

- Normalzustand 1.
- Fehlerzustand 2: Verstopfung der Stellluftzufuhr.
- Fehlerzustand 3: erhöhte Reibung.
- Fehlerzustand 4: reduzierter Zuluftdruck.
- Fehlerzustand 5: pneumatische Leckage.
- Fehlerzustand 6: Bruch einer Feder.

2. Bereitstellung eines Trainings-Datensatzes

Für jede Klasse wird ein Trainingsdatensatz von 8 Merkmalsvektoren zusammengestellt.

Insgesamt 6x8 = 48 Merkmalsvektoren.

echte Klasse	*vorhergesagte Klasse*					
	1	2	3	4	5	6
1	8	0	0	0	0	0
2	0	8	0	0	0	0
3	0	0	8	0	0	0
4	0	0	0	8	0	0
5	0	0	0	0	8	0
6	0	0	0	0	0	8

3. Auswahl der Klassifikatoren

4 Klassifikatoren werden hinsichtlich der Klassifikationsrate (größer oder 100%) und Umfang der Parameterdatei verglichen (je kleiner desto besser):

- SVM
- HSVM
- Bayes
- Entscheidungsbaum.

Die besten werden ausgewählt.

E. Merkmalsselektion

- Merkmalszahl sollen reduziert werden, so dass die Parameterzahl der Klassifikatoren kleiner wird, unter der Bedingung: die Klassifikationsrate konstant zu halten.
- Unter Verwendung des Quotientenkriteriums wird die Zahl der Merkmale von 23 auf 10 reduziert.
- Mit der reduzierten Merkmalszahl werden die Klassifikatoren erneut trainiert und verglichen.
- Aus 10 Merkmalen werden 5 ausgewählt und alle Kombinationen getestet.
- Ergebnis: HSVM und Bayes sind die besten Klassifikatoren mit Klassifikationsrate von 83,3 %.

Gliederung

1. Eigenschaften von KDD und Data-Mining
2. Modelle eines KDD-Prozesses
3. Schritte eines KDD-Prozesses
4. Anwendungsbeispiel: Fehlerdiagnose an einem Stellventil
5. Zusammenfassung

Zusammenfassung

- Knowledge Discovery in Databases (KDD) ist ein nichttrivialer Prozess der Wissensfindung aus Daten.
- Die Schritte des KDDs wurden in Modellen wie Fayyad, CRISP-DM... konkretisiert.
- Es gibt für KDD Software-Systeme wie KNIME, Rapid Miner, SPSS Clementine...
- Auswahl geeigneter Data-Mining-Methoden ist abhängig von Eigenschaften der Daten und Aufgabe.

Literatur

1. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From Data-Mining to Knowledge Discovery in Databases. AI Magazine Volume 1, Number 3 (1996).
2. K. Cios, W. Pedrycz, R. Swiniarski, L. Kurgan. Data-Mining: A Knowledge Discovery Approach. Springer Verlag 2007.
3. Jiawei Han, M. Kamber. Data-Mining, Concepts and Techniques. Elsevier Inc. 2006.
4. D. Olson, D. Delen. Advanced Data-Mining Techniques. Springer Verlag 2008.
5. O. Maimon, L. Rokach. Data-Mining and Knowledge Discovery Handbuch. Springer Verlag 2005.
6. M. Feid, M. Pandit. Systematischer Entwurf eines signalbasierten Fehlerdiagnosesystems, Teil 1 und 2. at Automatisierungstechnik 54 (2006). Oldenburg Wissenschaftsverlag.
7. Dokumentationen von Knime, Rapid Miner, SPSS Clementine.