PG 377
MedMiner - Maschinelles Lernen in der Intensivmedizin

Lehrstuhl für Künstliche Intelligenz

`http://www-ai.cs.uni-dortmund.de/LEHRE/PG/PG377/`

1. **PG-Thema:** *MedMiner* - Maschinelles Lernen in der Intensivmedizin
2. **PG-Zeitraum:** WS 2000/2001 und SS 2001
3. **PG-Umfang:** jeweils 8 SWS
4. **PG-Veranstalter:**
   - Ralf Klinkenberg, Lehrstuhl Informatik VIII, Klinkenberg@ls8, GB IV/R117, Tel. 5103
   - Stefan Rüping, Lehrstuhl Informatik VIII, Rueping@ls8, GB IV/R121, Tel. 5104
5. **PG-Aufgabe:** Die Behandlung von Patienten auf modernen Intensivstationen ist geprägt von einem enormen Einsatz von Überwachungsgeräten und Messverfahren für die verschiedensten medizinischen Daten des Patienten. In klinischen Informationssystemen können leicht 2000 verschiedene Variablen überwacht und im Minutenrythmus gespeichert werden. Dadurch kommen schnell mehrere Gigabyte an Daten zusammen. Zusätzlich zur reinen Masse an Daten erschweren die sehr unterschiedlichen Arten der Variablen die Analyse: Die erhobenen Messwerte reichen von kontinuierlich erhobenen, stetigen Daten wie Sauerstoffkonzentration und Blutdrücke über kontinuierlich erhobene, diskrete Daten wie Auffälligkeiten im EKG, unregelmässig erhobene Daten wie Laboruntersuchungen, berechnete Daten wie verschiedenen medizinische Indizes bis hin zu unstrukturierten Daten wie der medizinischen Vorgeschichte des Patienten.

   Im krassen Gegensatz zu der Masse an Informationen, die dem Arzt an der Bettseite des Patienten zur Verfügung steht, steht die in der Psychologie schon lange bekannte Erkenntnis, dass ein Mensch in der Regel nur mit sieben Faktoren gleichzeitig bewusst umgehen kann. Zusätzlich muss ein

Arzt am Intensivbett seine Entscheidung unter großem Zeitdruck treffen. Viele wichtige Informationen, die sich vielleicht erst aus der Kombination einer größeren Menge von Merkmalen erschließen lassen, bleiben daher ungenutzt.

Konkret ergeben sich für die PG folgende Aufgabengebiete:

Alarmfunktionen: Ist der Zustand des Patienten stabil oder ist eine medizinische Intervention nötig?

Interventionsvorhersage: Welche Intervention sollte vorgenommen werden (z.B. Dosierung von Medikamenten)?

Gruppierung: Lassen sich die Patienten anhand ihrer Reaktion auf Interventionen gruppieren?

Validierung: Reagiert der Patient auf die Intervention wie vorhergesagt?

Zur Untersuchung dieser Aufgabenstellung stehen in der PG reale Daten aus der chirurgischen Intensivstation der städtischen Kliniken Dortmund zur Verfügung. Dazu existiert am Lehrstuhl bereits eine Kooperation mit dem Leiter dieser Station, der auch für medizinische Rückfragen und zur Evaluation der Ergebnisse aus Anwendungssicht zur Verfügung steht.

Wissensentdeckung: Die Herausforderung, in der Intensivmedizin wichtige Informationen aus einer großen Menge von Daten herauszusuchen, ist ein typisches Beispiel für einen Wissensentdeckungsprozess (*Knowledge Discovery in Databases* - KDD). In der Praxis wird KDD in steigendem Umfang auf vielen Anwendungsgebieten betrieben, etwa um Informationen über das Käuferverhalten in Supermärkten zu gewinnen, Werbemaßnahmen zu planen oder industrielle Fertigungsabläufe zu optimieren. KDD ist damit ein stetig wachsendes Betätigungsfeld für Informatiker.

KDD ist als ein Prozess zu sehen, der in mehreren ineinander verzahnten Phasen abläuft, die eventuell auch iterativ wiederholt werden müssen. Er beginnt damit, zusammen mit dem Auftraggeber (in diesem Fall: mit den Ärzten) die Ziele und Anforderungen an das KDD-Projekt zu definieren und Einblick in das Anwendungsgebiet zu erhalten. In der zweiten Phase des KDD-Prozesses werden die vorhandenen Rohdaten betrachtet und auf Qualitätsprobleme (Ausreißer im EKG, fehlende Laboruntersuchungen, widersprüchliche Zeitangaben) untersucht und es wird versucht, die Daten zu bereinigen oder ergänzende Daten zu erhalten. In der dritten Phase werden die Rohdaten so transformiert, dass sie in der nächsten Phase leichter benutzt werden können (Berechung weiterer Attribute, z.B. medizinische Indizes, Umwandeln in das Eingabeformat spezieller Software). Der nächste Schritt wird als *Data Mining* bezeichnet. Hierbei werden verschiedene Verfahren (maschinelle Lernverfahren, statistische Verfahren, sonstige

Modellierungstechniken) angewandt, um aus den Daten Hypothesen (z.B. über das Auftreten von Auffälligkeiten im EKG oder die beste Medikamentendosierung) zu generieren und die besten Hypothesen zu suchen. Die Ergebnisse des Data Mining werden in der nächsten Phase evaluiert, indem sie auf neuen Daten getestet oder den Auftraggebern vorgestellt und von diesen aus der Anwendungssicht beurteilt werden. Ergebnis eines KDD-Prozesses kann ein schriftlicher Bericht sein, eine Implementation der Ergebnisse in ein größeres Computersystem oder eine erweiterte Fragestellung für einen weiteren KDD-Prozess.

Verwaltung der Daten: Angesichts der Menge der Daten ist es notwendig, der Verwaltung der Daten besondere Aufmerksamkeit zukommen zu lassen. Die meisten Lernverfahren erwarten die Daten in einem speziellen Eingabeformat. Zudem gibt es beliebig viele Möglichkeiten, aus den gegebenen Rohdaten durch Selektion und Konstruktion von Attributen, Beispiele für einen Lernalgorithmus zu generieren. Die richtige Auswahl von Lernattributen für jedes Lernverfahren beeinflusst dabei entscheidend die Qualität der Ergebnisse.

Da es nicht sinnvoll ist, die Daten im Eingabeformat jedes Lernalgorithmus vorrätig zu halten oder die Daten jedesmal von Hand zu generieren, sollen die Daten nur einmal in einer Datenbank gespeichert werden und möglichst allgemeine Operatoren implementiert werden, die die Konvertierung zwischen Datenbankformat und einem definierbaren Format für das Lernverfahren übernehmen. Durch den automatischen Aufruf von Lern- und Vorverarbeitungsverfahren über eine einheitliche Schnittstelle in einer geeigneten Reihenfolge ensteht damit aus den Einzelkomponenten ein übergreifendes Gesamtsystem.

Ziele der PG: Ziel der Projektgruppe ist eine Wissensentdeckung auf intensivmedizinischen Daten, um dadurch z.B. nötige Medikamenteninterventionen vorherzusagen oder eine Alarmfunktion zur Überwachung des Patientenzustands zu generieren. Dabei soll die Balance gehalten werden zwischen der konkreten Arbeit an der medizinischen Problemstellung und einer allgemein anwendbaren Vorgehensweise.

Um dieses allgemeine Ziel zu erreichen, sollen in der PG alle Phasen des KDD-Prozesses durchgeführt werden. Hierzu sind einige Verfahren des maschinellen Lernens oder aus der Statistik für die Datenbereinigungsphase, die Datentransformationsphase und die Data Mining-Phase auszuwählen, auf den Daten zu testen und die Tauglichkeit der Verfahren im Hinblick auf das Anwendungsfeld Medizin zu vergleichen. Insbesondere soll untersucht werden, wie eine geeignete Kombination von Verfahren zur Vorverarbeitung und zum Data Mining die Qualität der Ergebnisse verbessern kann.

6. **PG-Teilnahmevoraussetzungen:**

   - Nebenfach theoretische Medizin *oder* Statistik *oder* Statistik-Grundkenntnisse (aus der Grundstudiumsvorlesung für Informatiker) *(wünschenswert)*
   - Vorlesung "Maschinelles Lernen" *oder* "Künstliche Intelligenz" *oder* Vorlesung "Informationssysteme" *oder* "Information Retrieval" *oder* eine Spezialvorlesung oder ein Seminar aus einem dieser Gebiete *oder* Datenbankkenntnisse *(erforderlich)*
   - Kenntnis einer objekt-orientierten Programmiersprache *(wünschenswert)*

7. **Minimalziel:**

   - Entwurf und Implementierung von ineinandergreifenden Operatoren für die Datenvorverarbeitung und das maschinelle Lernen auf intensivmedizinischen Zeitreihendaten (Wissensentdeckung in Datenbanken)
   - Untersuchung der oben beschriebenen Anwendungsfragestellungen mit Hilfe dieser Operatoren

8. **Literatur:**

[1] M. J. Tobin, Hrsg. *Principles and Practice of Intensive Care Monitoring.* McGraw-Hill, New York, 1998.

[2] Special Issue on Data Mining Techniques and Applications in Medicine. *Artificial Intelligence in Medicine*, 16(1), Mai 1999.

[3] Katharina Morik, Michael Imhoff, Peter Brockhausen, Thorsten Joachims, und Ursula Gather. Knowledge Discovery and Knowledge Validation in Intensive Care. *Artificial Intelligence in Medicine*, 2000. Accepted for publication.

[4] *Feature Extraction, Construction, and Selection: A Data Mining Perspective.* Kluwer, 1998.

[5] Tom M. Mitchell. *Machine Learning.* McGraw Hill, New York, 1997.

[6] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

Weitere Literatur wird im Verlauf der PG bekannt gegeben.