Sirikit Ho
Dr. sc. hum.

Optimierung diagnostischer Kriterien des Neugeborenenscreenings durch Anwendung von Data Mining Techniken

Geboren am 23.09.1976 in Heidelberg
Diplom der Fachrichtung Medizinische Informatik am 11.01.2001 an der Universität Heidelberg

Promotionsfach: Medizinische Biometrie und Informatik
Doktorvater: Prof. Dr. rer. nat. Thomas Wetter

Das Neugeborenenscreening ist in Deutschland seit den 70er Jahren als medizinisch-diagnostische Maßnahme zur Früherkennnung von Neugeborenen mit angeborenen Stoffwechselerkrankungen und Endokrinopathien etabliert. Seit April 2005 ist auch das Erweiterte Neugeborenenscreening eine Regelleistung der Krankenkassen. Es erlaubt die Untersuchung von etwa 20 zusätzlichen Stoffwechselerkrankungen mit dem Multi-Analytverfahren Tandem-Massenspektrometrie.
Ziel dieser Arbeit war es, ein Prozessmodell zu entwickeln, mit dem die Sensitivität und Spezifität von Screeningtests für Erkrankungen des „Erweiterten Neugeborenenscreenings“ unter Anwendung von Techniken des Data Minings optimiert werden kann.
Die Praktikabilität des Prozessmodells sollte zudem anhand von Anwendungsbeispielen belegt werden. Zu beiden Zielen wurde in Kapitel 1 eine detaillierte Frage- und Aufgabenstellung formuliert. Im Kapitel 2 wurde der Kontext der Arbeit beschrieben und die Voraussetzungen für das Verständnis der Daten und der Anforderungen an die Datenmodellierung geschaffen.
Durch eine organisatorische Anforderungsanalyse und eine Recherche zu bestehenden diagnostischen Kriterien wurden die Anforderungen an diagnostische Kriterien ermittelt (vgl. Kapitel 3). Die besondere Herausforderung bei dieser Arbeit bestand in den hohen Anforderungen an die Sicherheit bei der Stellung von Verdachtsdiagnosen. Ein Übersehen von erkrankten Kindern musste trotz niedriger Inzidenzen der Erkrankungen unbedingt ausgeschlossen werden und dennoch sollten die Screeningtests auch eine hohe Spezifität erzielen. Des weiteren wurden Verständlichkeit und eine einfache Handhabung als Voraussetzung für einen Praxiseinsatz diagnostischer Kriterien im Neugeborenenscreening identifiziert.
Als Ausgangsbasis für die Auswahl von geeigneten Methoden und Arbeitsschritten diente das KDD-Prozessmodell nach Fayyad et al. (1996). In einer Spezialisierung dieses Prozessmodells wurde die Optimierung von diagnostischen Kriterien durch die Bearbeitung von Klassifikationsproblemen realisiert. Dazu wurden die einzelnen Prozessschritte nach Fayyad et al. (1996) für die spezielle Aufgabenstellung im Erweiterten Neugeborenenscreening konkretisiert (vgl. Kapitel 4):
Die vorbereitende Datenselektion beinhaltet die Datensichtung und -Analyse und die Erschließung von Datenquellen aus dem Neugeborenenscreening Heidelberg für die weiteren Prozessschritte. Die Datenvorverarbeitung umfasst die Datenbereinigung und Behebung von Abbildungsdefekten in den gesammelten Daten. Die Datenselektion für ein Klassifikationsproblem erfordert als ersten Prozessschritt eine Anpassung an die betrachtete Erkrankung und beinhaltet die Aufgaben Klassenbildung und Erzeugen von Stichproben sowie Partitionierung von Stichproben für die Nutzung in einer Kreuzvalidierung (vgl. Abschnitt 4.6).
Die Datentransformation stellt den Schwerpunkt in dieser Arbeit dar, da eine Analyse bestehender diagnostischer Kriterien Hinweise ergeben hatte, dass eine Verbesserung der diagnostischen Genauigkeit durch zusätzliche Analytkombinationen als Marker für die untersuchten Erkrankungen möglich ist. Ein Algorithmus zur Merkmalskonstruktion, also zur Auswahl geeigneter Analytkombinationen als Merkmale für die Stichproben, wurde entworfen und implementiert (vgl. Abschnitt 5.3). Zudem wurden zwei Strategien zur Binärisierung der Merkmale mit linearen Threshold-Funktionen erarbeitet. In den binären Merkmalen ist der Threshold integraler Bestandteil

und dient der Unterscheidung von Erkrankten und Kontrollen noch vor der Anwendung von Lernverfahren zur Klassifikation (vgl. Abschnitt 5.4).
Für das Data Mining wurde ein Algorithmus zur Regelinduktion ausgewählt, der numerische oder nominale Merkmale zur Vorhersage von Klassen (Klassifikation) nutzt. Die erzeugten Regeln sind einfach zu interpretieren und können direkt auf die Rohdaten im Neugeborenenscreening angewandt werden. Somit werden die Anforderungen an Verständlichkeit und einfache Handhabung erfüllt und die Klassifikationsregeln entsprechen den typischen in der Literatur beschriebenen diagnostischen Kriterien.
Die Evaluation der Ergebnisse der Anwendung des Prozessmodells entspricht größtenteils dem Standardverfahren beim KDD-Prozessmodell nach Fayyad et al. (1996). Sie umfasst die Evaluation und Interpretation der Ergebnisse und ein Review der vorangegangenen Prozessschritte und bildet den Abschluss des Prozssmodells.
Mit Hilfe dreier Anwendungsbeispiele wurde das spezialisierte Prozessmodell auf seine Praktikabilität getestet (vgl. Kapitel 6 bis 8). Dabei ergaben sich aus der Wahl möglichst unterschiedlicher Erkrankungen ganz unterschiedliche Ergebnisse betreffend die Anzahl und Güte der Merkmale aus der Merkmalskonstruktion und die Ergebnisse der Datenmodellierung. Alle drei Anwendungsbeispiele wurden erfolgreich bearbeitet und ihre Ergebnisse belegen, dass das spezialisierte Prozessmodell geeignet ist, um diagnostische Kriterien für das Erweiterte Neugeborenenscreening zu erarbeiten und mit diesen neuen Kriterien die Spezifität der Screeningtests zu verbessern.
Die ausgewählten Verfahren für das Prozessmodell wurden in Kapitel 9 zu den aufgestellten Aufgaben und Fragen aus Kapitel 1 in Bezug gesetzt und diskutiert. Die Arbeit schließt mit einem Ausblick auf mögliche Ansatzpunkte für Verbesserungen an den Datentransformationstechniken und weitergehende Arbeiten betreffend die Validierung und die inhaltlichen Interpretation der Ergebnisse (vgl. Abschnitt 9.3).