



Ruprecht-Karls-Universität Heidelberg
Medizinische Fakultät Mannheim
Dissertations-Kurzfassung

**Multiple Retrieval Case-based Reasoning – Klinisches
Entscheidungsunterstützungssystem auf unvollständigen
Datenbanken in Anwendung für das Tumorboard**

Autor: Nikolas Immanuel Löw
Institut / Klinik: Klinik für Strahlentherapie und Radioonkologie
Doktorvater: Prof. Dr. J. Hesser

Case-based Reasoning (CBR) ist ein Verfahren des maschinellen Lernens, das aufgrund des Wissens vergangener ähnlicher Fälle und ihren Lösungen Prognosen und Hilfestellungen zu einem neuen Fall liefert. Es besteht aus vier eigenständigen Phasen, die einen Kreislauf bilden. Der neue Fall und seine gewonnene Lösung werden im nächsten Durchlauf Teil dieses Kreislaufes und führen zu einer Erweiterung des bisherigen Wissens. Die Leistung des CBR hängt zum einen von der Größe und Informationsdichte der Datenbank ab und zum anderen maßgeblich von der korrekten Rangfolge der ähnlichen Fälle in der Retrieve-Phase, da alle folgenden Phasen von deren Ergebnissen betroffen sind. Mit den rasant wachsenden Datenbanken der heutigen Zeit steigert sich allerdings aus den verschiedensten Gründen auch die Menge an fehlenden Daten innerhalb dieser Datenbanken.

Diese Tatsache hat zur Folge, dass das gesamte CBR auf solchen Datenbanken destabilisiert wird und besonders die Rangfolge der Retrieve-Phase darunter leidet, weil unvollständige Fälle weniger verlässlich gewertet werden können als vollständige Fälle. Überraschenderweise existieren für diese Problematik bisher kaum Arbeiten, welche den Einfluss der fehlenden Daten auf die Retrieve-Phase des CBR untersuchen und eine zuverlässige Lösung basierend auf modernen Verfahren bieten. Insbesondere fehlt es an einer umfassenden Lösung, welche unterschiedliche Arten von Variablen betroffen von unterschiedlichen Typen von fehlenden Daten verarbeiten kann.

Als Antwort auf diese Fragestellung wird in dieser Arbeit das Multiple Retrieval Case-based Reasoning (MRCBR) vorgestellt und evaluiert. MRCBR ist ein Framework für CBR auf unvollständigen Datenbanken, das eine möglichst korrekte Rangfolge der ähnlichen Fälle mit Hilfe von modernen Methoden des maschinellen Lernens und der Statistik gewährleistet. Es bezieht die Verteilung der vollständigen Daten und die mögliche Verteilung der unvollständigen Daten in seine Berechnungen mit ein, indem es die Vorteile der Multiple Imputation und CBR in einem Verfahren effizient vereint.

Das Verfahren wurde als Erweiterung des klinischen Entscheidungsunterstützungssystems für das Tumorboard (CBR-TDS) entworfen, damit dieses auf der unvollständigen Datenbank der Klinik für Strahlentherapie und Radioonkologie (MOSAIQ) fehlerfrei und vertrauenswürdig arbeiten kann. In diesem Hinblick wurde es optimiert und getestet. Des Weiteren ist es jedoch ein allgemeines Verfahren, das nicht auf diesen Anwendungsbereich allein beschränkt ist und für jedes CBR System angepasst werden kann. Es erlaubt die Verarbeitung der gängigen Arten von Variablen in medizinischen Datenbanken, numerische und kategoriale Variable, und aller Typen von fehlenden Daten.

Die Methodik des MRCBR wurde mit acht konkurrierenden Methoden des letzten Standes der Technik verglichen, welche in der Lage sind CBR im Kontext von fehlenden Daten auszuführen. Die Ergebnisse auf der wahren vollständigen Datenbank bildeten die Referenz für die Einstufung der Ergebnisse der unterschiedlichen Verfahren mit Hilfe zweier verschiedener Fehlermaße. In vier repräsentativen Experimenten bestehend aus mehreren eigenständigen Versuchen wurden verschiedene Umgebungen und Bedingungen der fehlenden Daten realistisch simuliert und untersucht. Auch der Einfluss der Größe der Datenbank und andere Parameter des CBR wurden in Betracht gezogen. Für eine korrekte statistische Auswertung entsteht das Ergebnis jeder Methode in jedem Versuch aus der Mittelung von 200 einzelnen Ergebnissen.

MRCBR hat in so gut wie allen Versuchen die bestehenden Methoden übertroffen und zeigte verlässliche stabile Ergebnisse in fast jedem der Experimente. Besonders in großen Datenbanken und einer großen Anzahl von unvollständigen Variablen konnte es seinen Abstand zu den anderen Methoden noch vergrößern. Die Analyse des Verhaltens der Verfahren zeigte, dass es keine Möglichkeit gibt fehlende Daten zu ignorieren ohne die Leistung des CBR drastisch zu reduzieren.