Ilmenauer Beiträge zur Wirtschaftsinformatik

Herausgegeben von U. Bankhofer, V. Nissen
D. Stelzer und S. Straßburger

Tobias Rockel, Udo Bankhofer, Dieter Joenssen

# Entscheidungsbäume zur Imputation kategorialer Daten

**Arbeitsbericht Nr. 2015-02, August 2015**

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften und Medien
Institut für Wirtschaftsinformatik

**Autor**: Tobias Rockel, Udo Bankhofer, Dieter Joenssen

**Titel:** Entscheidungsbäume zur Imputation kategorialer Daten

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2015-02, Technische Universität Ilmenau, 2015

**ISSN 1861-9223**

ISBN 978-3-938940-57-0

urn:nbn:de:gbv:ilm1-2015200187



Anschrift: Technische Universität Ilmenau, Fakultät für Wirtschaftswissenschaften und Medien, Institut für Wirtschaftsinformatik, PF 100565, D-98684 Ilmenau.
http://www.tu-ilmenau.de/wid/forschung/ilmenauer-beitraege-zur-wirtschaftsinformatik/

# Gliederung

*Zusammenfassung: Entscheidungsbäume sind als Prognoseverfahren im Bereich des maschinellen Lernens verbreitet. Jedoch fehlt bisher eine Untersuchung, in wieweit sich die Algorithmen auch zur Imputation eignen. Dies ist insbesonders vor dem Hintergrund der verschiedenen Zielstellungen einer Imputation und einer Klassifikation relevant. So liegt bei einer Imputation häufig ein stärkerer Fokus auf der Struktur des Gesamtdatensatzes, wohingegen eine Klassifikation auf eine möglichst genaue Vorhersage einzelner Objekte abzielt. Neben den klassischen deterministischen Entscheidungsbäumen mit Majority Rule werden daher auch Entscheidungsbäume mit Class Probability Rule mit einer Zufallskomponente in die Untersuchung mit einbezogen. In einer Simulationsstudie, in der als Vergleichsverfahren zusätzlich eine Modus-Imputation und ein Random Hot Deck eingesetzt werden, zeigt sich, dass kein Verfahren in allen Fällen zum besten Ergebnis führt. So führt die Imputation mittels Entscheidungsbaum und Class Probability Rule meist zur geringsten Verzerrung der Häufigkeitsverteilung, jedoch stellt der Entscheidungsbaum mit Majority Rule meist die ursprünglichen Werte am besten wieder her. Welches Verfahren zur Imputation verwendet werden sollte, ist also abhängig vom Ziel der Imputation.*

*Schlüsselworte: Entscheidungsbäume, fehlende Daten, Imputation, Simulationsstudie*

# 1 Problemstellung

Unvollständige Datensätze sind in allen Bereichen ein Problem, treten jedoch besonders häufig bei Befragungen in den Sozialwissenschaften auf. Indes setzen die meisten Datenanalyseverfahren vollständige Daten voraus und sind daher auf Datensätze mit fehlenden Werten nicht ohne weiteres anwendbar (vgl. Graham 2009, S. 550–551). Eine Lösung dieses Problems ist es, fehlende Werte zu ersetzen, um so einen vollständigen Datensatz zu erhalten. Dieser Ersetzungsvorgang wird auch als Imputation bezeichnet (vgl. Bankhofer 1995, S. 104). Nach der Imputation steht ein vollständiger Datensatz zur Verfügung, der mit den herkömmlichen Datenanalyseverfahren ausgewertet werden kann.

Welche Imputationsverfahren im konkreten Fall unverzerrte Ergebnisse liefern, hängt davon ab, welchem Ausfallmechanismus die fehlenden Werte unterliegen. Hierzu hat Rubin eine Klassifikation entwickelt. Diese unterscheidet die drei Ausfallmechanismen Missing Completely at Random (MCAR), Missing at Random (MAR) und Not Missing at Random (NMAR) (vgl. Rubin 1976; Little und Rubin 2002, S. 12). Diese drei Ausfallmechanismen unterscheiden sich darin, welche Werte für den Datenausfall verantwortlich sind. Wenn das Fehlen der Werte stochastisch unabhängig von den vorhanden sowie den fehlenden Werten ist, spricht man von MCAR. In diesem Fall lassen sich auch einfache Imputationsverfahren anwenden, wie etwa die Ersetzung fehlender Werte durch einen geeigneten Lageparameter der vorhandenen Werte. Sind die fehlenden Werte nicht stochastisch unabhängig von den vorhandenen, aber von den fehlenden Werten selbst, dann bezeichnet man den Ausfallmechanismus als MAR. In diesem Fall müssen Imputationsmethoden verwendet werden, die zusätzlich auf vorhandene Informationen im Datensatz zurückgreifen, z. B. multiple Regression. Wenn das Fehlen eines Wertes auch von den unbeobachteten Werten abhängt, wird der Ausfallmechanismus NMAR genannt. In diesem Fall können die fehlenden Werte nicht ohne Hinzuziehen von externen Informationen sinnvoll imputiert werden.

Zur Imputation kann grundsätzlich jedes Prognoseverfahren herangezogen werden. So auch Entscheidungsbäume, die im Bereich des maschinellen Lernens als Prognoseverfahren verbreitet sind. Jedoch wurde die Verwendung von Entscheidungsbäumen zur Imputation im Bereich Missing Data noch nicht eingehend untersucht. Die Effektivität von Entscheidungsbäumen als Imputationsmethode lässt sich jedoch nicht direkt aus der Literatur des maschinellen Lernens ableiten, da das Ziel im maschinellen Lernen eine Prognose

möglichst exakter Werte für jedes einzelne Objekt ist. Bei statistischen Verfahren hingegen sind meist die Zusammenhänge und die Struktur des Gesamtdatensatzes entscheidend. Wie aber unter anderem van Buuren (2012, S. 45–46) anmerkt, sind diese Ziele nicht zwingend kongruent. Mit Blick auf diese unterschiedlichen Ziele soll im Rahmen dieser Arbeit untersucht werden, wie gut Entscheidungsbäume zur Imputation geeignet sind.

Erste Ideen, Entscheidungsbäume im Bereich fehlender Daten zu verwenden, finden sich bereits bei Kalton und Kasprzyk (1982, S. 28) und Kalton (1983, S. 84). Allerdings werden in diesen Arbeiten Entscheidungsbäume nur als ein Zwischenschritt genutzt. Hingegen nutzen Creel und Krotki (2006) Entscheidungsbäume direkt zur Konstruktion von Imputationsklassen, innerhalb derer sie aber auf andere Imputationsverfahren zurückgreifen. In dieser Arbeit hingegen sollen Entscheidungsbäume unmittelbar zur Imputation fehlender Werte verwendet werden. Die Möglichkeiten hierzu werden aufgezeigt und untersucht.

Dazu wird in Kapitel 2 eine Einführung in die Konstruktion von Entscheidungsbäumen und Möglichkeiten zur Auswahl von Imputationswerten bei Entscheidungsbäumen gegeben. Anschließend wird in Kapitel 3 eine Simulationsstudie beschrieben, welche zur Untersuchung der Imputationsgüte von Entscheidungsbäumen herangezogen wird. Im 4. Kapitel werden dann die Ergebnisse der Simulationsstudie dargestellt. Abschließend werden die Ergebnisse dieser Arbeit noch einmal zusammengefasst und ein Ausblick gegeben.

# 2 Entscheidungsbäume zur Imputation fehlender Daten

Bei der Verwendung von Entscheidungsbäumen zur Imputation können die vollständig beobachteten Objekte als Trainingsset zur Konstruktion eines Baumes verwendet werden. Dieses Vorgehen setzt streng genommen einen MCAR-Ausfallmechanismus voraus. Der mithilfe der vollständigen Objekte erstellte Baum wird dann zur Bestimmung der fehlenden Werte in den übrigen Objekten verwendet. Im Rahmen dieser Arbeit wird sich dabei auf die Verwendung von Entscheidungsbäumen zur Imputation kategorialer Daten beschränkt. Hierbei ist die Imputation eines Wertes im Missing Data Bereich gleichbedeutend mit der Klassifikation eines Objektes im maschinellen Lernen.

In Abbildung 1 ist ein Beispiel für einen einfachen Entscheidungsbaum dargestellt. Ein Objekt, das anhand dieses Entscheidungsbaums klassifiziert werden soll, beginnt im obersten Knoten des Baums, der Wurzel. Wenn der beobachtete Wert im Merkmal $X_1$ des Objektes größer als 0 ist, gelangt man in den linken Tochterknoten, andernfalls in den rechten.

Da es sich bei dem linken Tochterknoten der Wurzel um ein Endknoten, auch Blatt genannt, handelt, wird ein Objekt, das in diesem Knoten landet, mit $Y = 2$ klassifiziert, wobei $Y$ eine fiktive Klassenvariable ist. Falls ein Objekt in den rechten Tochterknoten der Wurzel gelangt, wird es in Abhängigkeit der beobachteten Ausprägung von $X_2$ in das linke oder rechte Blatt weitergeleitet und durch das entsprechende Blatt klassifiziert (vgl. z. B. Witten et al. 2011, S. 64).

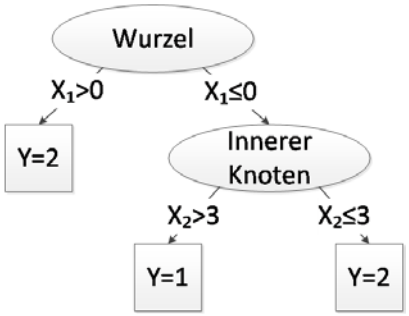


**Abbildung 1: Ein Entscheidungsbaum**

## 2.1 Konstruktion von Entscheidungsbäumen

Während die oben beschriebene Klassifikation eines Objektes durch einen bestehenden Entscheidungsbaum simpel ist, ist die Konstruktion eines Entscheidungsbaums ungleich komplexer. Zur Konstruktion von Entscheidungsbäumen werden häufig Algorithmen, wie CART (Breiman et al. 1984), CHAID (Kass 1980) oder C4.5 (Quinlan 1993) verwendet. Die Entscheidungsbaumalgorithmen unterscheiden sich im Detail, ihre generelle Struktur ist jedoch sehr ähnlich. Sie basieren meist auf einer Divide and Conquer-Strategie, mit der ein Datensatz anhand jeweils eines Merkmals in immer homogenere Teildatensätze zerlegt wird.

Für die Unterteilung wird ein Maß zur Bestimmung der Heterogenität (engl. impurity measure) bzw. Homogenität eines (Teil-) Datensatzes bzw. Knotens benötigt. Ein solches Heterogenitätsmaß sollte umso höher sein, je mehr verschiedenartige Objekte im Datensatz vorhanden sind. Zwei Objekte werden dabei als verschieden angesehen, wenn sie nicht dieselbe Ausprägung in der Klassenvariable $Y$ besitzen (vgl. Rokach und Maimon 2008, S. 53). Als Maß werden dabei unter anderem der Gini-Index bei CART (vgl. Breiman et al. 1984, S. 109), der Chi-Quadrat-Wert bei CHAID (vgl. Kass 1980, S. 121) und Information Gain bei C4.5 (vgl. Quinlan 1993, S. 23–24) verwendet.

Ein Algorithmus bewertet mit dem Heterogenitätsmaß jede mögliche Unterteilung des Datensatzes. Die Unterteilung, die zum größten Homogenitätsgewinn führt, wird als vorteil-

hafteste angesehen und der Datensatz wird entsprechend geteilt. Für jeden der so entstandenen Teildatensätze wird erneut die beste Unterteilung ermittelt und durchgeführt (vgl. Han et al. 2012, S. 333).

Diese Unterteilung des Datensatzes erfolgt rekursiv, bis in einem Teildatensatz nur noch Objekte einer Klasse vorhanden sind, oder ein anderes Abbruchkriterium (z. B. Unterschreitung der minimalen Anzahl an Objekten in einem Teildatensatz) zum Stopp der Unterteilung führt. Jede Unterteilung eines Datensatzes wird im Baum durch einen Knoten und die daran anschließenden Äste symbolisiert. Ein Datensatz, der nicht weiter unterteilt wird, stellt im Entscheidungsbaum ein Blatt dar.

Ein häufiges Problem der so erzeugten Bäume ist, dass sie die Besonderheiten des Datensatzes zu stark wiederspiegeln und nicht – wie eigentlich gewünscht – ein allgemeingültiges Modell darstellen. Daher wird ein Baum häufig noch „beschnitten“. Dabei werden Teile des Baums, die als zu speziell angesehen werden, wieder entfernt (Han et al. 2012, S. 344).

## 2.2 Auswahl des Imputationswertes

Nach der Konstruktion eines Baumes steht das Gerüst. Jedoch muss noch festgelegt werden, welche Klasse einem Objekt in einem Blatt zugewiesen wird, respektive welcher Imputationswert durch den Baum bestimmt wird. Hierfür gibt es zwei verschiedenen Verfahren – die Entscheidung nach Majority Rule (MR) oder Class Probability Rule (CPR).

Die MR wird vor allem im Bereich des maschinellen Lernens verwendet. Bei ihr wird jedes Objekt mit dem Modus der vorhandenen, während der Konstruktion des Baumes in den Knoten gelangten Objekte klassifiziert. Anstelle der MR kann zur Klassifikation eines Objektes auch die CPR verwendet werden. Bei ihr wird zufällig aus den beobachteten Werten eines Blattes gezogen. Ein Baum, der diese Regel zur Klassifikation verwendet, wird auch als Class Probabilty Tree bezeichnet. Die CPR besitzt demzufolge – im Gegensatz zur MR – eine Zufallskomponente (vgl. Breiman et al. 1984, S. 121–125). Quinlan (1986, S. 152–153) hat gezeigt, dass die MR eine höhere erwartete Genauigkeit besitzt als die CPR.

Die Unterschiede zwischen CPR und MR werden am einfachsten an einem Beispiel deutlich. Wenn in einem Blatt 60% der Objekte zu Klasse 1 gehören und 40% zu Klasse 2, dann weist die MR bei der Imputation allen Objekten die Klasse 1 zu. Die CPR hingegen weist einem Objekt mit 60%-Wahrscheinlichkeit die Klasse 1 und mit 40%-Wahrscheinlichkeit die Klasse 2 zu. Die MR führt in dem Beispiel zu einer erwarteten Ge-

nauigkeit von 60%, wohingegen die CPR nur eine erwartete Genauigkeit von 52% besitzt.[1] Jedoch ist auch offensichtlich, dass die erwarteten Anteilswerte der CPR exakt mit den beobachteten relativen Häufigkeiten übereinstimmen, während die MR zu einer – in diesem Beispiel erheblichen – Verzerrung der Anteile von Klasse 1 und 2 führt, da sie allen Objekten Klasse 1 zuweist.

Diese Beobachtungen über das Verhalten der beiden Methoden gelten nicht nur für das Beispiel, sondern generell im Fall von MCAR. So führt die MR immer zu einer höheren erwarteten Genauigkeit als die CPR. Jedoch reproduziert die CPR die Verteilung der Werte im Mittel besser. Dies gilt jedoch nur unter der MCAR-Annahme und ist beim Vorliegen eines MAR- oder NMAR-Ausfallmechanismus fragwürdig. Außerdem werden im Bereich der Missing Data Literatur häufig die Vorzüge einer stochastischen Imputation, wie bei Verwendung der CPR, betont (vgl. Little und Rubin 2002, S. 66). Da also die Empfehlung der beiden Literaturzweige sich wiedersprechen, sollte die Eignung von Entscheidungsbäumen zur Imputation empirisch untersucht werden.

## 3 Aufbau der Simulationsstudie

Um die Imputationsgüte von Entscheidungsbäumen in Abhängigkeit von verschiedenen Einflussfaktoren zu untersuchen, wurde eine Simulationsstudie mittels der Statistiksoftware R (R Core Team 2014) durchgeführt. Abbildung 2 stellt den Ablauf der Studie und die variierten Faktoren grafisch dar.

- Für jeden Datensatz (Brustkrebs, SPECT, Schach, Alex5)
  - Für die Ausfallmechanismen MCAR, MAR1, MAR2
    - Für die Ausfallwahrscheinlichkeiten von 0,1 bis 0,5
      - 1000 Iterationen
        - Erzeugen einen Datensatz mit fehlenden Werten
        - Für jedes Imputationsverfahren (CPR, MR, Modus, RHD)
          - Imputieren der fehlenden Werte
          - Analysieren der Imputation

**Abbildung 2: Ablaufplan der Simulationsstudie**

Um einen möglichen Einfluss der Objekt- und Merkmalsanzahl zu untersuchen, wurden vier reale Datensätze mit unterschiedlichen Kombinationen, wie in Tabelle 1 dargestellt,

[1] Die erwarteten Genauigkeiten wurden unter der Annahme berechnet, dass die beobachteten relativen Häufigkeiten mit den wahren Anteilswerten übereinstimmen.

ausgewählt. Die Datensätze Brustkrebs[2], SPECT und Schach stammen vom UCI Machine Learning Repository (Lichman 2013) und Alex5 sind die ersten 5000 Objekte des Trainingsdatensatz eines Wettbewerbs aus dem Bereich des maschinellen Lernens (Causality Workbench 2014). Bis auf den Datensatz Brustkrebs enthalten die Datensätze keine fehlenden Werte. Bei Brustkrebs wurden nur die vollständigen Objekte in der Simulationsstudie verwendet. Die fehlenden Werte wurden in folgenden dichotomen Merkmalen erzeugt: „Wiedertauftreten einer Krebserkrankung“ (Brustkrebs), „Gesamtdiagnose“ (SPECT), Merkmal 11 (Alex5) und „Gewinn für…“ (Schach).

| | wenige Merkmale | viele Merkmale |
|---|---|---|
| wenige Objekte | Brustkrebs (10 Merkmale, 277 Objekte) | SPECT (23 Merkmale, 267 Objekte) |
| viele Objekte | Alex5 (11 Merkmale, 5000 Objekte) | Schach (37 Merkmale, 3196 Objekte) |

**Tabelle 1: Verwendete Datensätze**

Ferner wurde der Anteil fehlender Werte von 10% bis 50% in den Stufen 10%, 20%, 30%, 40% und 50% variiert. Die fehlenden Werte wurden dabei mit drei unterschiedlichen Ausfallmechanismen erzeugt. Zum einen wurde ein MCAR-Mechanismus (zufällige Löschung innerhalb des Merkmals) und zum anderen wurden zwei MAR-Mechanismen verwendet. Bei den MAR-Mechanismen ist die Wahrscheinlichkeit für das Fehlen eines Wertes in den oben genannten Merkmalen abhängig von der Ausprägung des Objektes in einem anderen dichotomen Merkmal. Die Wahrscheinlichkeit für das Fehlen ist bei MAR 1 doppelt so hoch, wenn das Objekt im „ausfallerzeugenden“ Merkmal den häufigeren Wert besitzt, als wenn es den weiniger häufigen aufweist. Bei MAR 2 ist die Wahrscheinlichkeit vierfach so hoch. Als „ausfallerzeugende“ Merkmale wurden Merkmale mit einem mäßigen Zusammenhang[3] zum Merkmal mit fehlenden Werten verwendet: Merkmal 5 (Brustkrebs), Merkmal 22 (SPECT), Merkmal 1 (Alex5), Merkmal 33 (Schach).

Zur Imputation wurden die Entscheidungsbaumalgorithmen CART (mittels rpart-Paket Version 4.1-8, Therneau et al. 2014) und C4.5 (mittels RWeka Version 0.4-23, Witten et al. 2011, Hornik et al. 2009) jeweils einmal mit der CPR und mit der MR verwendet.[4] Zum

[2] Der Datensatz stammt ursprünglich von M. Zwitter und M. Soklic, University Medical Centre, Institut für Onkologie, Ljubljana, Jugoslawien.

[3] korrigierter Kontingenzkoeffizient ca. 0,4

[4] Für den C4.5 wurden die von Quinlan (1993) vorgeschlagenen Einstellungen (mindestens zwei Tochterknoten müssen zwei Elemente enthalten, 25% Konfidenzintervall beim Post-Pruning) verwendet. Bei rpart wurden alle Bonsai-Techniken bis auf die minimale Blattgröße (Einstellung: 2 Elemente) deaktiviert und die 0-SE-Regel zum Post-Pruning verwendet.

Vergleich wurden außerdem ein Random Hot Deck (RHD) und eine Modus-Imputation durchgeführt.

Zur Beurteilung der Ergebnisse wurde zum einen die Genauigkeit der Imputation als der Anteil der imputierten Werte, die mit dem Originalwert übereinstimmen, gemessen. Zum anderen wurden die Abweichungen der relativen Häufigkeiten einer Merkmalsausprägung *a* im imputierten Merkmal berechnet. Die absolute Differenz zwischen der wahren relativen Häufigkeit dieser ursprünglichen Ausprägung $f_{Y^{orig}}(a_1)$ und den Werten nach der Imputation $f_{Y^{imp}}(a_1)$ wurden wie folgt berechnet

$$\Delta f(Y^{orig}, Y^{imp}) = \left| f_{Y^{orig}}(a_1) - f_{Y^{imp}}(a_1) \right|.$$

Die Werte beider Beurteilungskriterien wurden für jede Faktorstufenkombination in 1000 unabhängigen Simulationsläufen berechnet und gemittelt.

# 4 Ergebnisse der Simulationsstudie

Im Folgenden werden die Ergebnisse der Simulationsstudie getrennt nach den beiden Gütekriterien dargestellt. Dabei werden die Ergebnisse des CART-Algorithmus nicht dargestellt, da sie mit den Ergebnissen des C4.5-Algorithmus nahezu identisch sind.

## 4.1 Genauigkeit der Imputation

In der Abbildung 3 ist die Genauigkeit der Imputationssverfahren für alle Datensätze gegenübergestellt. Allgemein lässt sich feststellen, dass der Entscheidungsbaumalgorithmus in Verbindung mit der MR – bis auf sehr wenige Ausnahmen – zu den genausten Imputationsergebnissen führt. Hingegen liefert das RHD tendenziell die ungenauste Imputation.

Die im Kapitel 2 beschriebene höhere erwartete Genauigkeit der MR zeigt sich in den Ergebnissen nicht nur für MCAR sondern für alle Ausfallmechanismen. So ist der Entscheidungsbaum mit der MR immer genauer als mit der CPR. Diese Ordnungsrelation ist auch zwischen den Verfahren Modus- und RHD-Imputation zu finden, bei denen die Modus-Imputation fast immer genauere Ergebnisse liefert. Ferner ist der Entscheidungsbaum mit CPR stets genauer als das RHD.

In den meisten Fällen hat weder der verwendete Ausfallmechanismus noch der Anteil fehlender Werte einen starken Einfluss auf die Imputationsergebnisse der Verfahren. Nur beim Datensatz Schach verschlechtert sich der Modus, wenn viele Daten fehlen und ein MAR-Ausfallmechanismus vorliegt. Hingegen zeigt sich, dass die Entscheidungsbäume

sowohl vom Übergang von wenigen Merkmalen zu vielen Merkmalen, als auch von der Erhöhung der Objektanzahl profitieren.

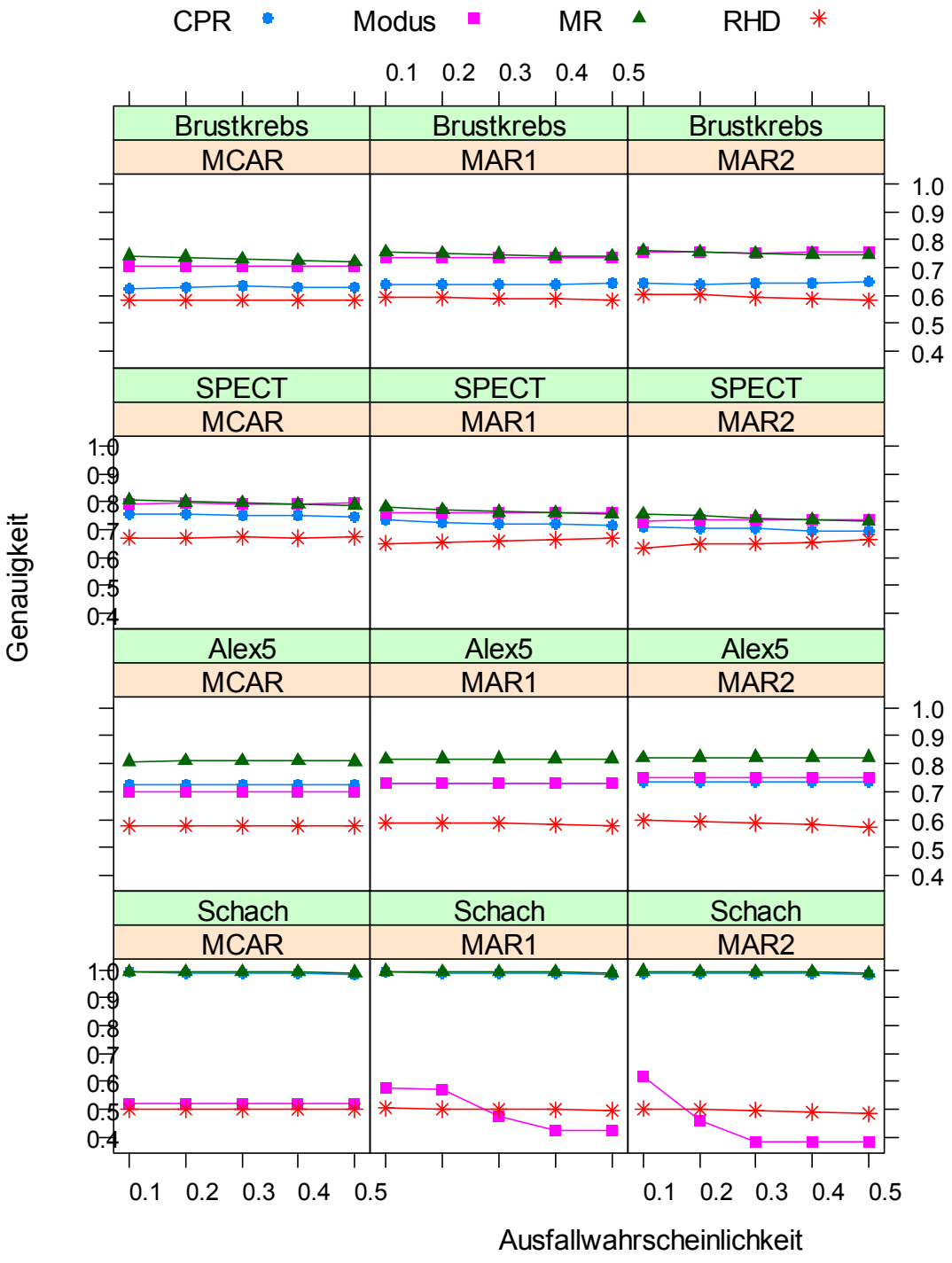


**Abbildung 3: Genauigkeit der Imputationsverfahren**

So sind die Ergebnisse des Entscheidungsbaums mit CPR beim kleinen Brustkrebsdatensatz dem ungenauen RHD noch wesentlich ähnlicher als der genaueren Modus-Imputation. Hingegen nähert sich der Entscheidungsbaum mit CPR bei SPECT (Erhöhung der Merk-

malsanzahl) dem Modus deutlich an und bei Alex5 (Erhöhung der Objektanzahl gegenüber Brustkrebs) ist er bei MCAR sogar besser als die Modus-Imputation. Auch kann sich der Entscheidungsbaum mit der MR bei Alex5 zum ersten Mal deutlich von der Modus-Imputation absetzen. Besonders deutlich wird der Gewinn, den Entscheidungsbäume aus zusätzlichen Informationen ziehen, beim Schachdatensatz. Hier liefern beide Entscheidungsbäume eine Genauigkeit von nahezu 100%, während die beiden anderen Verfahren meist unter 60% liegen.

RHD und Modus hingegen profitieren auf Grund ihres univariaten Charakters nicht von den zusätzlichen Informationen durch die Erhöhung der Merkmals- oder Objektanzahl. Ihr Ergebnis hängt vielmehr davon ab, wie die Häufigkeitsverteilung im Merkmal mit fehlenden Werten aussieht. So stimmt bei MCAR die Genauigkeit des Modus sehr gut mit der relativen Häufigkeit der häufigsten Ausprägung überein.

## 4.2 Abweichung der relativen Häufigkeit

Die Abweichung der relativen Häufigkeit der Imputationsverfahren ist in Abbildung 4 dargestellt. Besonders auffällig ist, dass die Modus-Imputation immer zur – meist mit Abstand – größten Abweichung führt und damit die schlechtesten Ergebnisse liefert. Die Ergebnisse der anderen Verfahren liegen im Gegensatz dazu relativ nahe beieinander.

Allgemein lässt sich feststellen, dass alle Verfahren – unabhängig vom Ausfallmechanismus – mit steigender Anzahl fehlender Daten schlechtere Ergebnisse liefern. Beim MCAR-Ausfallmechanismus führen meist das RHD und der Entscheidungsbaum mit CPR zu den besten Ergebnissen. Jedoch verschlechtern sich die Ergebnisse des RHD beim Übergang zu den MAR-Mechanismen. Die Ergebnisse des Entscheidungsbaumalgorithmus bleiben hingegen, vor allem bei den Datensätzen mit vielen Objekten, vergleichsweise konstant. Bei der Modus-Imputation ist kein eindeutiges Muster beim Übergang von MCAR zu MAR erkennbar.

Ähnlich wie bei der Genauigkeit der Imputation profitieren die Entscheidungsbäume von dem zusätzlichen Informationsgehalt, den ein größerer Datensatz bereitstellt. So nähern sich die Ergebnisse des Entscheidungsbaums mit MR den Ergebnissen des RHD und des Entscheidungsbaums mit CPR sowohl beim Übergang vom Brustkrebsdatensatz zu SPECT, als auch beim Übergang von Brustkrebs zu Alex5 deutlich an. Beim Schachdatensatz ist der Entscheidungsbaum mit MR sogar in der Lage das RHD zu überholen.

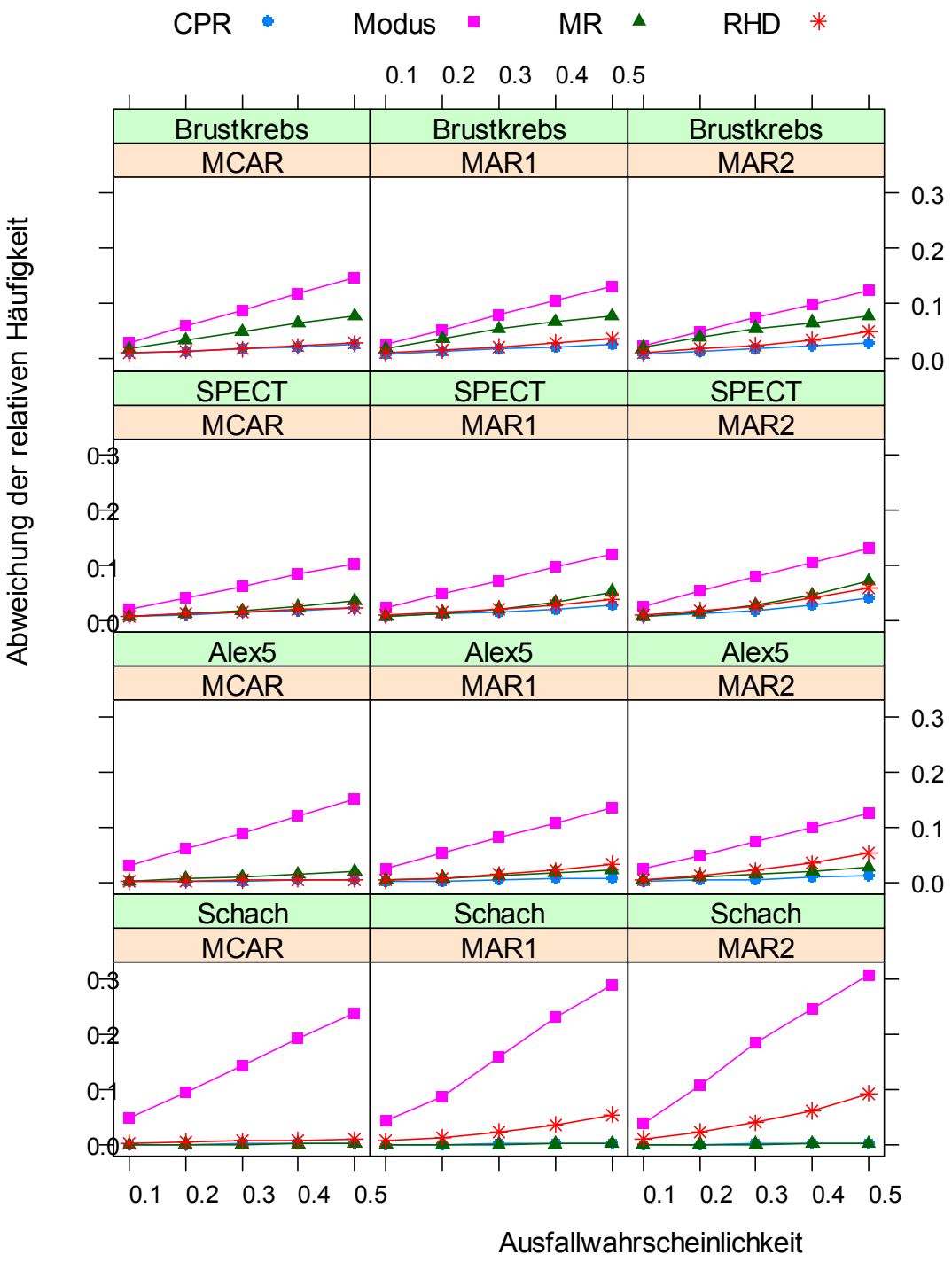


**Abbildung 4: Abweichung der relativen Häufigkeit**

## 5 Zusammenfassung und Ausblick

Im Rahmen der Arbeit werden Entscheidungsbäume auf Ihre Eignung zur Imputation untersucht. Dabei wird zwischen Entscheidungsbäumen mit der deterministischen MR und Entscheidungsbäumen mit der stochastischen CPR unterschieden. Hierbei entspricht ein

Baum ohne Verzweigungen mit der MR bzw. der CPR einer Modus- bzw. RHD-Imputation. Diese vier Verfahren wurden mittels einer Simulationsstudie verglichen. Dabei führen Entscheidungsbäume mit MR nie zu schlechteren Ergebnissen als die Modus-Imputation und Entscheidungsbäume mit CPR nie zu schlechteren Ergebnissen als das RHD. Diese Ergebnisse zeigen, dass der mit den Entscheidungsbäumen verbundene Aufwand sich auf die Imputationsqualität positiv auswirkt. Welcher Entscheidungsbaumalgorithmus verwendet wird, spielt eine untergeordnete Rolle, da die Ergebnisse von CART und C4.5 nahezu identisch sind. Vielmehr sind die Größe des Datensatzes und die verwendete Regel (CPR oder MR) für die Imputationsergebnisse entscheidend.

Außerdem liefert in der Simulationsstudie der Entscheidungsbaum mit MR die genausten Imputationsergebnisse, während der Entscheidungsbaum mit CPR die relative Häufigkeit des ursprünglichen Datensatzes am zuverlässigsten rekonstruiert. Folglich ist in Anwendungen, in denen die genaue Prognose einzelner Objekte eine übergeordnete Rolle spielt, eine Entscheidungsbaumimputation mit MR zu empfehlen. Hingegen sollte in Bereichen, in denen die Struktur des Datensatzes von primärem Interesse ist, eine Entscheidungsbaumimputation mit CPR verwendet werden.

Diese Ergebnisse sind insbesondere deshalb interessant, weil in der Missing Data Literatur, unter anderem von Little und Rubin (2002, S. 66), empfohlen wird, mit Zufallskomponente zu imputieren. Die durchgeführte Simulationsstudie zeigt, dass dieses Vorgehen zwar zu einer geringeren Verzerrung der Häufigkeitsverteilung führt, jedoch die einzelnen Werte nicht so gut rekonstruiert. In der klassischen Statistik ist meist die Verteilung der Merkmalsausprägungen für die Schätzung von Parametern entscheidend und daher das Vorgehen von Little und Rubin (2002) zu empfehlen. Falls jedoch die Struktur der einzelnen Objekte maßgeblich für ein Analyseverfahren ist, was zum Beispiel im Bereich des maschinellen Lernens der Fall sein kann, erscheint eine Imputation der Lageparameter sinnvoller. Bei der Entscheidung zwischen deterministischen und stochastischen Imputationsverfahren ist folglich das Ziel der späteren Analyse entscheidend.

## Literaturverzeichnis

Bankhofer, Udo (1995): Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse. Bergisch Gladbach, Köln: Eul.

Breiman, Leo; Friedman, Jerome H.; Olshen, Richard A.; Stone, Charles J. (1984): Classification and Regression Trees. New York et al.: Chapman & Hall.

van Buuren, Stef (2012): Flexible Imputation of Missing Data. Boca Raton: CRC Press.

Causality Workbench (2014): Active Learning Challenge. Online verfügbar unter http://www.causality.inf.ethz.ch/activelearning.php?page=datasets.

Creel, Darryl V.; Krotki, Karol (2006): Creating Imputation Classes Using Classification Tree Methodology. In: American Statistical Association (Hg.): Proceedings of the Section on Survey Research Methods, S. 2884–2887.

Graham, John W. (2009): Missing Data Analysis: Making it Work in the Real World. In: *Annual Review of Psychology* 60, S. 549–576.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2012): Data Mining. Concepts and Techniques. 3. Aufl. Amsterdam et al.: Morgan Kaufmann.

Hornik, Kurt; Buchta, Christian; Zeileis, Achim (2009): Open-Source Machine Learning: R Meets Weka. In: *Comput Stat* 24 (2), S. 225–232.

Kalton, Graham (1983): Compensating for Missing Survey Data. The University of Michigan. Ann Arbor, Michigan (Research Report Series, Institute for Social Research).

Kalton, Graham; Kasprzyk, Daniel (1982): Imputing for Missing Survey Responses. In: American Statistical Association (Hg.): Proceedings of the Survey Research Methods Section, S. 22–31.

Kass, G. V. (1980): An Exploratory Technique for Investigating Large Quantities of Categorical Data. In: *Journal of the Royal Statistical Society* 29 (2), S. 119–127.

Lichman, M. (2013): UCI Machine Learning Repository. Online verfügbar unter http://archive.ics.uci.edu/ml.

Little, Roderick J. A.; Rubin, Donald B. (2002): Statistical Analysis with Missing Data. 2. Aufl. Hoboken: Wiley.

Quinlan, John Ross (1986): The Effect of Noise on Concept Learning. In: Ryszard S. Michalski, Jaime G. Carbonell und Tom M. Mitchell (Hg.): Machine Learning. An Artificial Intelligence Approach. Volume 2. Los Altos: Morgan Kaufmann, S. 149–166.

Quinlan, John Ross (1993): C4.5. Programs for Machine Learning. San Mateo: Morgan Kaufmann.

R Core Team (2014): R: A Language and Environment for Statistical Computing. Version. Vienna, Austria. Online verfügbar unter http://www.R-project.org/.

Rokach, Lior; Maimon, Oded (2008): Data Mining with Decision Trees. Theroy and Applications. Singapore et al.: World Scientific.

Rubin, Donald B. (1976): Inference and Missing Data. In: *Biometrika* 63 (3), S. 581–592.

Therneau, Terry; Atkinson, Beth; Ripley, Brian D. (2014): rpart: Recursive Partitioning and Regression Trees. R package version 4.1-8. Online verfügbar unter http://CRAN.R-project.org/package=rpart.

Witten, I. H.; Frank, Eibe; Hall, Mark A. (2011): Data Mining. Practical Machine Learning Tools and Techniques. 3. Aufl. Burlington: Morgan Kaufmann.