DOI: 10.17185/duepublico/40721

Regel-basierte Kategorisierung mit dem Hierarchischen Dirichlet Prozess

Thomas Glassen*. Verena Nitsch**

**Institut für Arbeitswissenschaft, Universität der Bundeswehr München, Neubiberg, Deutschland (Tel: ++49(0)89-6004-2115; e-mail: thomas.glassen@unibw.de).*
***Institut für Arbeitswissenschaft, Universität der Bundeswehr München, Neubiberg, Deutschland (e-mail: verena.nitsch@unibw.de)*

Abstract: Der Hierarchische Dirichlet Prozess (HDP), ein rationales Kategorisierungsverfahren, welches auch als Klassifikator für autonome Roboter vorgeschlagen wurde (Nakamura, Nagai & Iwahashi, 2011; Nakamura, Yoshiki, Takayuki & Masahide, 2015), wird anhand eines klassischen Phänomens der menschlichen Kategorisierungsleistung evaluiert und mit alternativen Modellen aus der psychologischen Forschung verglichen. Ziel des Vorhabens ist es sowohl zu einer Reduzierung des Evaluationsdefizits prominenter Kategorisierungsmodelle beizutragen, zu welchen der HDP gerechnet werden kann, als auch Verbesserungspotential des Modells zu ermitteln (Pothos & Wills, 2011; Wills & Pothos, 2012). Gegenstand der Evaluation ist die experimentell ermittelte Lernschwierigkeit von sechs unterschiedlichen Kategorisierungsregeln aus Nosofsky, Gluck, Palmeri, McKinley und Glauthier (1994b). Die Gegenüberstellung der vom HDP vorhergesagten Lernschwierigkeit mit der tatsächlichen Schwierigkeit zeigen ein klassisches Defizit des Modells, wie es häufig auch in älteren Modellen zu menschlichen Kategorisierung festgestellt wurde. Es werden mögliche Ursachen für die ungenügende Modellierungsleistung des HDP benannt.

1. EINFÜHRUNG

Die vorwissensgesteuerte Zuordnung von neuen Objekten zu bestehenden Klassen, auch als Klassifikation bezeichnet, zählt nach Kotsiantis (2007) zu den am häufigsten ausgeführten Aufgaben sogenannter intelligenter Systeme. Die Automatisierung dieser Aufgabe stellt ein Teilgebiet des maschinellen Lernens dar, wobei gängige Verfahren nach ihrer Funktionsweise (z.B. logisch, perceptron-basiert oder statistisch) und ihrer Eignung für bestimmte Daten (diskrete/kontinuierliche bzw. in-/interdependente Attribute, vollständige/fehlende Werte, kein/vorhandenes Rauschen) unterschieden werden (Kotsiantis, 2007). Innerhalb dieses Forschungsfeldes ist bereits länger bekannt, dass kein Klassifikations-Algorithmus, über alle denkbaren Lernsituationen betrachtet, einem anderen hinsichtlich der durchschnittlichen Generalisierungsleistung überlegen sein kann (Schaffer, 1994; Lattimore & Hutter, 2013). Dieses als No Free Lunch (NFL) bezeichnete Faktum ist auch dann gültig, wenn Klassifikatoren mit unterschiedlichen Stärken als ein Ensemble kombiniert eingesetzt werden (Schaffer, 1994; Giraud-Carrier & Provost, 2005).

Dennoch kann es einen besten Klassifikator für einen abgegrenzten Aufgabenbereich geben. So haben beispielsweise Klassifikatoren von Service- und Rettungsrobotern eine große Schnittmenge an Lernsituationen zu bewältigen, welche typischerweise Menschen mit großer Effizienz meistern. Bisher entwickelte Roboter kommen jedoch nicht annähernd an diese Leistung heran, was sicherlich nicht nur an Problemen bei der Sensorverarbeitung liegen mag. Betrachtet man etwa den Menschen als ein funktionierendes Beispiel eines effizienten 'Roboters', müssen funktionelle Defizite bei aktuellen Robotern ebenso den bestehenden Klassifikationssystemen zugeschrieben werden. So sehen Cohen und Lefebvre (2005) die mentale Operation der Kategorisierung als Grundlage des menschlichen Wissenserwerbs und als elementarstes Phänomen der Kognition. Ohne sie wäre eine Generalisierung von früherem Wissen auf neue Beobachtungen nicht möglich, da jedes neue Objekt als einzigartig angesehen werden müsste (Waldmann, 2008). Es ist daher naheliegend, die Fähigkeit zur effizienten Kategorisierung auch als Basis der menschlichen Flexibilität und somit der gegenüber Robotern überlegenen Anpassungsfähigkeit anzunehmen.

Tatsächlich zeigen Giraud-Carrier und Provost (2005) dass ein solcher bester Algorithmus für einen eingegrenzten Aufgabenbereich existieren kann, wenn *meta-learning* als zugrundeliegendes Lernprinzip angenommen wird. Im Wesentlichen handelt es sich dabei um einen Lernalgorithmus, welcher einen Bias eines Meta-Selektors erlernt, der aus einer definierten Anzahl an Klassifikationsalgorithmen den geeignetsten für eine Lernsituation auswählt. Giraud-Carrier und Provost (2005) gehen davon aus, dass die relevantesten Lernsituationen strukturelle Ähnlichkeiten aufweisen, wodurch das NFL Theorem an Bedeutung verliere. Sie kommen deshalb zu dem Ergebnis: „Finding a ULA [Ultimate Learning Algorithm] thus consists of finding a learning algorithm whose induced models closely match our world's underlying distribution of functions“ (Giraud-Carrier & Provost, 2005, S. 15).

Wie kann nun ein solcher Lernalgorithmus gefunden werden? Zwei kognitionspsychologische Herangehensweisen können hierfür eine wichtige Orientierung geben. Dies sind die rationale und die mechanistische Herangehensweise (Sakamoto, Jones & Love, 2008). Ersterer wurde von

DOI: 10.17185/duepublico/40721

Anderson und Milson (1989) in den Kognitionswissenschaften eingeführt und beschreibt einen sechsstufigen Prozess um kognitive Prozesse zu analysieren (Chater & Oaksford, 1999). Dabei werden zunächst (1) Ziele des kognitiven Systems definiert, (2) ein formales Modell der Umweltbedingungen entwickelt und (3) computationale Einschränkungen festgelegt um (4) anschließend das optimale Verhalten abzuleiten. In den darauffolgenden zwei Schritten werden (5) empirische Evaluationen durchgeführt und (6) der Prozess wiederholt um die Theorie zu verbessern (Chater & Oaksford, 1999). In der mechanistischen Herangehensweise hingegen wird Verhalten über Prozesse und Repräsentationen simuliert, von denen man annimmt, dass sie in ähnlicher Weise im menschlichen Gehirn vorkommen (Sakamoto et al, 2008). Wie auch im rationalen Ansatz sind dabei eine anschließende Evaluation und Iteration des Vorgehens unerlässlich.

Beide Ansätze haben in den letzten Jahrzehnten einflussreiche formale Modelle der menschlichen Kategorisierung hervorgebracht (Pothos & Wills, 2011), beispielsweise das Generalized Context Model (Nosofsky, 1986), das Rational Model of Categorization (RMC; Anderson, 1991), das Attention Learning Covering Map Modell (ALCOVE; Kruschke, 1992) oder das Supervised and Unsupervised Stratified Adaptive Incremental Network Modell (SUSTAIN; Love, Medin & Gureckis, 2004). Dass die hierbei gewonnenen Erkenntnisse auch für das maschinelle Lernen von Bedeutung sein können, zeigt ein Leistungsvergleich von SUSTAIN mit einer linearen Support Vector Machine (SVM; Boser, Guyon & Vapnik, 1992; Cortes & Vapnik, 1995) im Bereich der visuellen Objekterkennung (Carmantini, Cangelosi & Wills, 2014). Hierbei erreichte SUSTAIN mindestens die Genauigkeit einer SVM und übertraf diese sogar partiell.

Bisher fehlt jedoch noch ein einheitliches Modell, welches möglichst alle bekannten Phänomene der menschlichen Kategorisierungsleistung beschreiben kann. Dies ist eine Voraussetzung um die Mechanismen des zugrundeliegenden Klassifikationsystems verstehen und beispielsweise für technische Anwendungen nutzen zu können. Ein Hindernis auf dem Weg zur Vereinheitlichung stellt der bestehende Mangel an ausführlichen Vergleichsstudien dar, durch welche vielversprechende Modelle systematisch identifiziert werden könnten (Wills & Pothos, 2012). So gibt es zwar vereinzelte Gegenüberstellungen von Modellen, doch beziehen sich diese oft nur auf wenige Kategorisierungsphänomene.

Um zur Schließung dieser Lücke beizutragen, wird in dem vorliegenden Artikel ein wichtiges aber bisher kaum evaluiertes Modell des Rationalen Ansatzes mit alternativen Kategorisierungsmodellen verglichen. Es handelt sich dabei um den Hierarchischen Dirichlet Prozess (HDP) welcher von Teh (2004, 2006) als statistisches Clusterverfahren vorgestellt und von Griffiths, Canini, Sanborn und Navarro (2007) wenig später als Kategorisierungsmodell vorgeschlagen wurde. Die Wichtigkeit des Modells resultiert daraus, dass es ein Vereinigungsmodell zweier kontroverser Repräsentationsparadigmen, der Exemplar- und der Prototypensichtweise, darstellen soll. In Ersterer wird davon ausgegangen, dass jedes beobachtete Mitglied einer Kategorie gesondert gespeichert wird, während man in Letzterer die ausschließliche Existenz eines kategorienspezifischen Prototypen annimmt, welcher ein typisches Exemplar dieser Kategorie beschreibt (Ross & Makin, 1999). Kontrovers dabei ist, dass beispielsweise die Exemplar- gegenüber der Prototypensichtweise deutlich mehr Kategorisierungsbefunde erklären kann, jedoch speichertechnisch unplausibel ist (Waldmann, 2008). Griffiths et al. (2007) schlagen deshalb mit dem HDP einen Mittelweg vor, mit welchem je nach Datenlage nicht nur ein sondern eine beliebige Anzahl an Prototypen für dieselbe Kategorie gebildet werden können. Dabei zeigen sie, dass bisherige rationale Kategorisierungsmodelle lediglich Spezialfälle des HDP darstellen. Die Erwägung des HDP als geeignetes Modell der menschlichen Kategorisierung reiht sich zudem ein in eine Serie an Arbeiten, welche unter dem Forschungsthema „Bayes'sche Kognition“ zusammengefasst werden können (Griffiths, Kemp & Tenenbaum, 2008; Griffiths, Chater, Kemp, Perfors & Tenenbaum, 2010; Holyoak & Cheng, 2010; Gopnik & Wellman, 2012; Pouget, Beck, Ma & Latham, 2013). Die zugrunde liegende These dieser Arbeiten unterstellt der menschlichen Kognition, Entscheidungen auf der Basis (approximierter) bayes'scher Statistik zu treffen (Kruschke, 2010; Gelman et al., 2013). Ob dieser Ansatz vielversprechend ist, wird noch ausführlich diskutiert (Marcus, 2010; Jones & Love, 2011; Bowers & Davis, 2012; Griffiths, Chater, Norris & Pouget, 2012; Endress, 2013; Marcus & Davis, 2013). Eine Evaluation des HDP könnte weitere Argumente für oder gegen diesen Ansatz aufdecken.

2. DER HIERARCHISCHE DIRICHLET PROZESS

Beim HDP handelt es sich um eine Priorverteilung aus der bayes'schen nichtparametrischen Statistik, welches üblicherweise in einem Clustermodell, einem sogenannten Mixture Model, Anwendung findet. Er unterscheidet sich von parametrischen Priorverteilungen darin, dass dessen Komplexität nicht vorab festgelegt ist, sondern mit den Daten wächst (Hjort, Holmes, Müller & Walker, 2010). Wie auch der Dirichlet Prozess (DP; Ferguson, 1973) ist der HDP ein Prior über potentielle Partitionierungen jeder beliebigen Menge an Ereignissen. Abweichend vom DP ermöglicht der HDP jedoch eine gemeinsame Clusternutzung über Kategorien hinweg. Dabei können Veränderungen einer Kategorienstruktur Auswirkungen auf die Repräsentation einer anderen Kategorie haben. Um diese Funktionalität zu ermöglichen, ist der HDP hierarchisch aufgebaut. Jede Kategorie wird dabei durch einen eigenen DP modelliert, während ein übergeordneter DP einen Pool an Clusterprototypen zur Instanziierung der konkreten Kategoriencluster vorhält. Drei Parameter bestimmen dabei die Wahrscheinlichkeit kategorienspezifischer Partitionen. Dies sind die Konzentrationsparameter α und γ und die Basisverteilung H. Ist α niedrig werden kategorienspezifische Partitionen mit wenigen Clustern wahrscheinlicher, ist α hoch trifft dies stattdessen auf Partitionen mit vielen Clustern zu. γ hingegen beeinflusst die Wahrscheinlichkeit der gemeinsamen Nutzung von Kategorieclustern. Dabei wird bei einem niedrigeren γ ein stärkeres Teilen von Clustern wahrscheinlicher. Technisch gesehen handelt es sich bei den Kon-

DOI: 10.17185/duepublico/40721

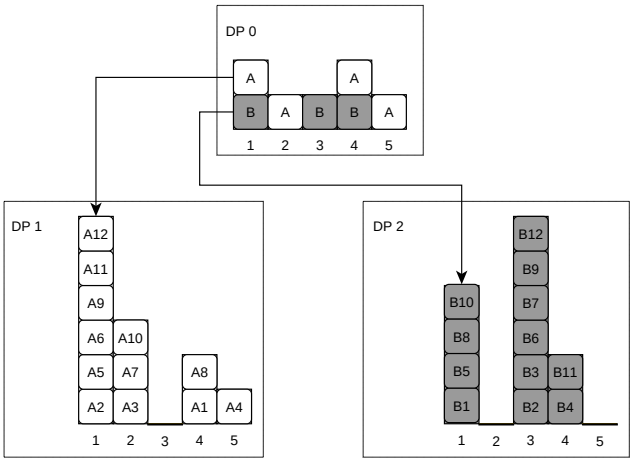

Abb. 1. Schematische Darstellung des Hierarchischen Dirichlet Prozesses (Griffiths, Sanborn, Canini, Navarro & Tenenbaum, 2011, S.181).

zentrationsparametern um Streumaße, welche bestimmen wie sehr eine potentielle Partition von der Durchschnittspartition eines DPs, der zugehörigen Basisverteilung, abweichen kann (Teh, 2010). Die Basisverteilungen der Kategorien-DPs stammen beim HDP aus dem übergeordneten DP, während dessen Basisverteilung konkret festgelegt werden muss. Dies ist die Basisverteilung H.

Eine Momentaufnahme des HDP ist in Abbildung 1 dargestellt. DP 1 repräsentiert in diesem Beispiel Kategorie A, DP 2 Kategorie B. Buchstaben mit Zahlen stellen Objekte der jeweiligen Kategorie dar. Die Einsortierreihenfolge erfolgte den Zahlen nach aufsteigend. Jede Spalte in DP 0 stellt einen Prototyp dar, welcher in mindestens einer Kategorie einem Cluster zugeordnet werden kann. Quadratische Elemente mit Buchstaben benennen die jeweilige Kategorie in welcher der Prototyp vorkommt.

Um den HDP-Prior in einem Mixture Model (HDPMM) verwenden zu können, muss zusätzlich eine Likelihoodverteilung gewählt werden, welche die Wahrscheinlichkeit eines Ereignisses in Abhängigkeit von den Mitgliedern eines ausgewählten Clusters modelliert. Hierbei kann es sich beispielsweise um eine Multinomialverteilung wie bei Teh et al. (2006) handeln. Entsprechend dem Bayes-Theorem kann dann bei Vorliegen konkreter Beobachtungsdaten eine Bestimmung der Posteriorverteilung erfolgen, deren Modalwert (= globaler Hochpunkt) einen Rückschluss auf die optimale Clusterung der Ereignisse erlaubt. Dies wird üblicherweise über Gibbs Sampling erreicht (Geman & Geman, 1984; Blunsom, Cohn, Goldwater & Johnson, 2009), ein Markov Chain Monte Carlo Verfahren (MCMC; Gilks, 2005) welches es ermöglicht, anhand der Konditionalverteilungen aller Parameter eines Modells, die Posteriorverteilung zu approximieren.

Für eine detailliertere technische Beschreibung des HDP werden Interessierte auf einführende und weiterführende Literatur zum Thema DP (Neal, 2000; Navarro, Griffiths, Steyvers & Lee, 2006; Teh, 2010; Gershman & Blei, 2012) und HDP (Teh et al., 2006; Blunsom et al., 2009; Teh & Jordan, 2010) verwiesen.

3. METHODIK

Für die Evaluation des Modells wurde eine von Griffiths et al. (2007) erweiterte Variante der Software npBayes2.1 von Teh et al. (2006) verwendet (im folgenden npBayes2.1G genannt). npBayes2.1G ermöglicht es gegenüber npBayes2.1 mehrdimensionale binäre Stimuli zu kategorisieren, indem anstelle einer Multinomialverteilung unabhängige Bernoulliverteilungen für jede Dimension zur Bestimmung der Likelihood verwendet werden. Eine grafische Repräsentation des verwendeten Modells findet sich in Abbildung 2 (links). Abgerundete Rechtecke in der Repräsentation symbolisieren definierbare Parameter des Modells. Kreise stellen Variablen dar, deren Verteilung in Abbildung 2 (rechts) angegeben sind. Der grau hinterlegte Kreis steht für die bereits beobachteten Ereignisse (= Trainingsdaten). Umrahmte Knoten kennzeichnen Replikationen des Modells innerhalb des Rahmens. Die Gammaverteilung(en) für $\beta 0_k$, $\beta 1_k$ werden durch die Parameter Shape und Scale, für γ und α durch Shape und Inverse Scale (= Rate) spezifiziert. Die Likelihoodverteilung F stellt eine multivariate Verteilung aus d Bernoulliverteilungen dar, welche durch das Tupel $\boldsymbol{\varphi_{ji}} = (\psi_{ji1},...,\psi_{jid})$ parametrisiert ist.

Die Ausführung der Software npBayes2.1G erfolgte in Matlab 2012a unter OpenSuse 13.2. Zur Suche der optimalen Parameter wurde der genetische Algorithmus *ga* der Global Optimization Toolbox mit einer Populationsgröße von 80 bei 100 Generationen ausgeführt. Der Algorithmus ruft iterativ eine Fitnessfunktion mit einem Parametersatz auf, in welcher das entsprechende Experiment unter Verwendung der übergebenen Parameter simuliert wird. Als Rückgabe erhält *ga* die Summe der Fehlerquadrate (SSE), welche sich aus den vorhergesagten Werte des Modells und den vorliegenden Probandendaten berechnet und die Güte des Parametersatzes repräsentiert. In jedem Schritt (= 1 Generation) evaluiert *ga* 80 Parametersätze. Eine Generation entsteht dabei aus der vorherigen durch Mutation der Parametersätze, unter

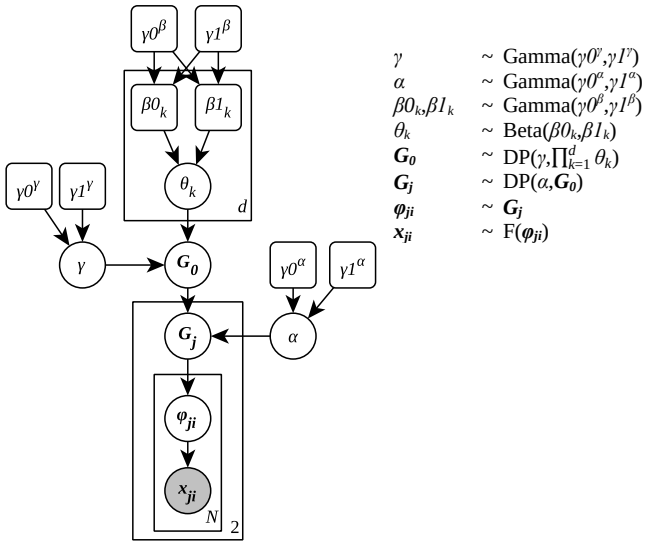

Abb. 2. Darstellung des verwendeten HDPMM aus Griffiths et al. (2007).

DOI: 10.17185/duepublico/40721

Berücksichtigung der jeweiligen SSEs, mit dem Ziel den Rückgabewert der Fitnessfunktion weiter zu minimieren.
Die MCMC Simulation der npBayes2.1G Software erfolgte mit einem Burn-In von 200 und 30 Samples in einem Abstand von jeweils 10 Iterationen. Der Burn-In bezeichnet dabei die Anzahl an Iterationen ohne Sampling ab Beginn der Simulation. Er dient dazu den Einfluss einer ungünstigen Startposition zu minimieren, welche bei MCMC Verfahren üblicherweise zufällig gewählt wird (Kruschke, 2010). Der Abstand zwischen den tatsächlich gespeicherten Samples wird auch Thinning genannt. Eine Erhöhung dieser kann zu einer geringeren Autokorrelation führen, da die Sampleschritte durch mehrfache Wiederholung einer zufälligen Parameterwahl unabhängiger von einander werden; dadurch erhöht sich bei gleicher Anzahl an Samples die Präzision. Dies ist insbesondere dann von Vorteil, wenn eine Erhöhung der Sampleanzahl aufgrund eines aufwendigen post-processings zu einem deutlich erhöhten Zeitaufwand führt und somit keine bessere Option darstellt (Link & Eaton, 2012).

4. EXPERIMENT

Im Folgenden soll die Evaluation des HDP an Daten der Studie von Nosofsky et al. (1994b) vorgestellt werden, welche mit dem Ziel erhoben wurden, Kategorisierungsmodelle miteinander vergleichen zu können. Nach Love et al. (2004) stellen sie einen schweren Test für jedes Kategorisierungsmodell dar. Es handelt sich bei dieser Studie um eine Teilreplikation und Erweiterung des Experiments 1 von Shepard, Hovland & Jenkins (1961), in welchem sechs Probanden die Aufgabe hatten, die Kategorienzugehörigkeit von acht Stimuli bezüglich sechs unterschiedlicher Zuordnungsregeln zu erlernen. Nosofsky et al. (1994b) replizierten die Schwierigkeitsreihenfolge der Zuordnungsregeln bei 120 Versuchspersonen und zeichneten zusätzlich den regelspezifischen Lernverlauf auf. In Abbildung 3 sind die verwendeten Regeln (A) und Stimuli (B) dargestellt. Jeder Kreis in A repräsentiert ein Item, gleichfarbige Kreise symbolisieren die Zugehörigkeit zur selben Kategorie innerhalb der jeweiligen Regel. Das Aussehen der Stimuli variierte auf den drei Dimensionen Größe (groß, klein), Farbe (schwarz, weiss), und Form (Quadrat, Dreieck). Eine beispielhafte Anordnung der Items für eine Versuchsperson und Zuordnungsregel findet sich in B. Die Bedeutung der Dimensionen und der Dimensionsausprägungen wurden für jede Versuchsperson (VP) und jede Zuordnungsregel randomisiert. Dies entspricht einer randomisierten Rotation des Würfels in B. Jede der 120 Versuchspersonen hatte nacheinander zwei Zuordnungsregeln zu erlernen, womit jede dieser Regeln über 40 Probanden getestet wurde. Eine Lernaufgabe startete zunächst mit zwei Blöcken à acht Stimuli. Hierbei wurden alle Stimuli genau einmal präsentiert und sollten von der Versuchsperson entweder Kategorie A oder B zugeordnet werden. Nach jeder Kategorisierung erfolgte ein Feedback, welches die VP über die Korrektheit der Zuordnung informierte. Anschließend folgten Blöcke à 16 Stimuli, in welchen jeder Stimulus zweimal präsentiert wurde. Die Lernaufgabe endete sobald die VP vier zusammenhängende

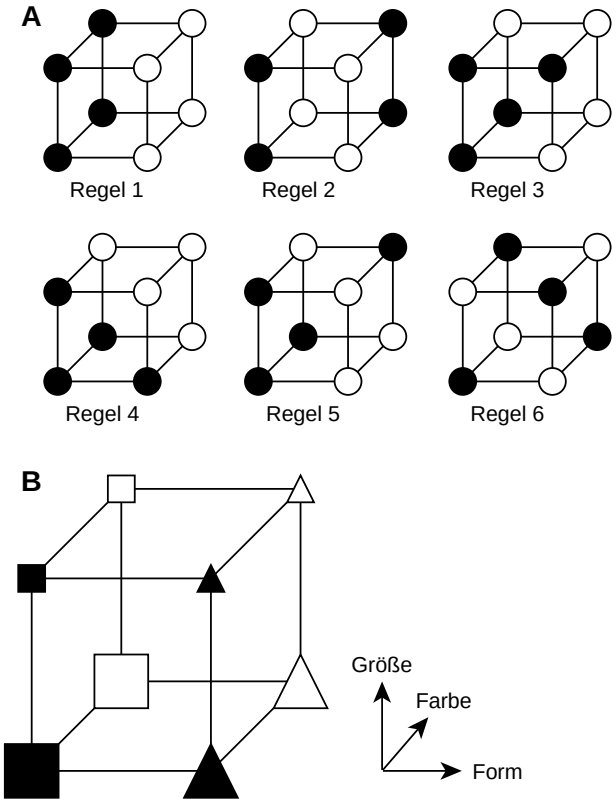

Abb. 3. Zuordnungsregeln und Stimuli aus Nosofsky et al. (1994b). (A) Darstellung der sechs Zuordnungsregeln.

Subblöcke à acht Stimuli fehlerfrei bewältigte oder alternativ nach insgesamt 400 Trials. In Abbildung 4 oben ist der mittlere Lernverlauf bei jeder Kategorisierungsregel über die Blöcke 1 bis 16 dargestellt. Die X-Achse bezeichnet den jeweiligen Block, die Y-Achse die zugehörige mittlere Präzision der VP. Da sich das Kategorisierungsphänomen ausschließlich auf das abgebildete Muster der Präzisionskurven bezieht und die individuellen Itemschwierigkeiten durch viele Modelle vorhergesagt werden (Nosofsky et al., 1994b), wurden im Weiteren die falsch-negativ und falsch-positiv Raten nicht berücksichtigt.

5. MODELLIERUNG

Um die Daten erfolgreich zu reproduzieren, wurden verschiedene Prior-Varianten ausprobiert. Hierbei wurde das HDPMM mit einem Gamma-Prior auf einen gemeinsamen Konzentrationsparameter α_0 (= α = γ) und einem gemeinsamen Gamma-Prior auf jedem $\beta 0_k$, $\beta 1_k$ (Variante 1) bzw. direkt festgelegtem β_0 (= $\beta 0_k$ = $\beta 1_k$) (Variante 2) getestet. Zudem wurde überprüft, wie gut das HDPMM mit einem Gamma-Prior auf jeweils α und γ und einem gemeinsamen Gamma-Prior auf jedem $\beta 0_k$, $\beta 1_k$ (Variante 3) bzw. direkt festgelegtem β_0 (= $\beta 0_k$ = $\beta 1_k$) (Variante 4) die Daten modelliert. Je nach Modellierungsvariante wurden entweder die direkt festgelegten Parameter und / oder die Hyperparameter der jeweiligen Gamma-Prior mittels *ga* gesucht (siehe Abb. 2). Wurden die $\beta 0_k$ und $\beta 1_k$ direkt gewählt, beschränkte sich die Suche auf nur einen Parameter

DOI: 10.17185/duepublico/40721

β_0 mit θ_k ~ Beta(β_0, β_0). Hierdurch wurde ein Dimensionsbias (= Vorwissen) vor dem ersten zu kategorisierenden Objekt vermieden. Die Wahrscheinlichkeit des HDPMM einen Stimulus S falsch zu kategorisieren, wurde über die Wahrscheinlichkeit der Generierung von S durch jede Kategorie bestimmt. Bei einer korrekten Kategorie A für S galt:

$$P_{Error}(S) = 1 - \frac{P_{Gen}(A_S)}{P_{Gen}(A_S)+P_{Gen}(B_S)} \quad (1)$$

$P_{Gen}(A_S)$ bzw. $P_{Gen}(B_S)$ wurden über das arithmetische Mittel der aus der Posteriorverteilung gesampelten Generierungswahrscheinlichkeiten von S berechnet. Sofern noch keine Mitglieder einer Kategorie bekannt waren, wurde die initiale Generierungswahrscheinlichkeit eines Stimulus S bei einer leeren Kategorie (beispielsweise A) wie folgt festgelegt:

$$P_{Init}(A_S) = \frac{1}{2^{N_{Dim}}} \quad \text{mit } N_{Dim} = 3 \quad (2)$$

Dies entspricht der Wahrscheinlichkeit den Erwartungswert aus der verwendeten Basisverteilung zu ziehen.

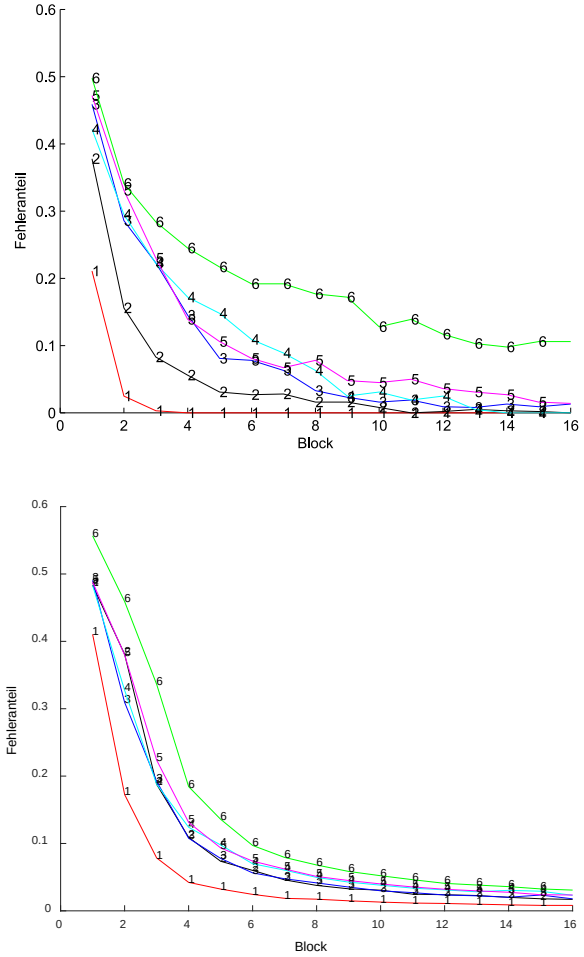

Abb. 4. Gegenüberstellung der sechs mittleren Lernkurven der Versuchspersonen aus Nosofsky et al. (1994b) (oben) und des HDPMM (Variante 4) (unten).

6. ERGEBNISSE

Für jede der genannten Prior-Varianten wurde ein Parametersatz $\boldsymbol{p_{best}}$ über *ga* gesucht, mit welchem die Summe der Fehlerquadrate zwischen den vorhergesagten Daten des Modells und der Experimentaldaten aus Nosofsky et al. (1994b) minimal ist. Zusätzlich wurde das beste Ergebnis jeder Prior-Variante optisch inspiziert und die zugehörigen vorhergesagten Lernkurven mit denen der Versuchspersonen verglichen.

Bei allen genannten Prior-Varianten zeigte sich dabei das qualitativ gleiche Muster in den vorhergesagten Daten. Variante 4 hatte die geringste SSE (= 0.297), danach folgten Variante 1 (SSE = 0.318), 2 (SSE = 0.324) und 3 (SSE = 0.333). Für Variante 4 wurde ein β_0 = 0.086709 mit $\gamma 0^{\alpha}$ = 0.53899, $\gamma 1^{\alpha}$ = 0.4214, $\gamma 0^{\gamma}$ = 0.6763 und $\gamma 1^{\gamma}$ = 0.68906 als bester Parametersatz identifiziert. Die zugehörigen Lernkurven dieser Variante sind in Abbildung 4 dargestellt.

Wie zu sehen ist, kann die Zuordnungsregel 1 auch vom HDPMM am besten gelernt werden. Danach folgen Regel 2, 3, 4 und 5 dicht auf einander. Zuordnungsregel 6 wird vom Modell tendenziell am Schlechtesten gelernt, was qualitativ dem Muster der VP-Daten entspricht. Besonders hervorzuheben ist jedoch die Lernkurve der Zuordnungsregel 2. Hier zeigt sich keine Unterscheidbarkeit zur Regel 3, welche aus den VP-Daten deutlich hervorgeht.

7. DISKUSSION

Nur die wenigsten Modelle sind in der Lage eine leichtere Lernbarkeit der Zuordnungsregel 2 gegenüber den Regeln 3,4 und 5 vorherzusagen (Nosofsky et al., 1994b; Love et al., 2004). Wie die Modellierungsergebnisse verdeutlichen, scheitert auch das Kategorisierungsmodell von Griffiths et al. (2007) an diesem Punkt. Gemessen am SSE stellt der HDP ebenfalls kein gutes Modell für die Daten dar. Er modelliert diese schlechter als das technisch verwandte bayes'sche RMC (SSE = 0.182) und bleibt damit weit hinter Modellen wie ALCOVE (SSE = 0.061) oder dem Rule-Plus-Exception Model (RULEX; SSE = 0.077) (Nosofsky et al., 1994b; Nosofsky, Palmeri & McKinley, 1994a). Wie die zwei letztgenannten Modelle kann auch SUSTAIN die Daten von Nosofsky et al. qualitativ besser reproduzieren als der HDP. Zwar ist kein SSE bei Love et al. (2004) angegeben, doch deutet die optische Darstellung der Kurven auf eine vergleichbare Vorhersageleistung wie die von ALCOVE hin.

Die Stärken des HDP liegen darin, gegenüber traditionellen Prototypen- und Exemplarmodellen einen Kompromiss zu wählen, indem nicht alle bekannten Exemplare der fraglichen Kategorien aber auch nicht nur deren singuläre repräsentative Prototypen für die Kategorisierungsentscheidung herangezogen werden. Zudem besitzt der HDP die Fähigkeit, Cluster während des Lernprozesses adaptiv zu reorganisieren und somit je nach Datenlage geeignetere Kategorienstrukturen herauszubilden. Letzeres ist eine Fähigkeit, die aktuell nur die wenigsten Kategorisierungsmodelle beherrschen (McDonnell & Gureckis, 2011). Die Ergebnisse der Modellierung zeigen jedoch, dass für das vorliegende Problem des Regellernens die genannten Eigenschaften nicht hinreichend sind.

DOI: 10.17185/duepublico/40721

Es wird angenommen, dass das HDPMM aus Griffiths et al. (2007) ähnliche Modelldefizite aufweist wie das ebenfalls bayes'sche RMC von Anderson (1991), welches technisch betrachtet mit einem DP Mixture Model vergleichbar ist (Sanborn, Griffiths & Navarro, 2010). Bei letztgenanntem vermutet Nosofsky et al. (1994b) ein fehlendes dimensionsspezifisches Aufmerksamkeitslernen (= Gewichtung von Dimensionen) als (Teil-)Ursache der ungenügenden Modellierungsleistung. Darüber hinaus sieht er auch in alternativen Konvertierungsformeln für die Berechnung von beobachteten Antwort- aus internen Modellwahrscheinlichkeiten eine Option zur Steigerung der Modellierungsgüte. Das Potential dieser Vorschläge ist in einem weiteren Schritt zu prüfen.

Love et al. (2004) nennen neben den Daten von Nosofsky et al. (1994b) noch weitere Herausforderungen für aktuelle formale Modelle der Kategorisierung. Sie stellen, wie Nosofskys charakteristische Lernkurven, Phänomene der menschlichen Kategorisierungsleistung dar, welche ein gutes Modell der menschlichen Kategorisierung beziehungsweise ein menschenähnliches kognitives System vorhersagen bzw. abbilden sollte. Hierbei handelt es sich um den Befund, dass bei komplexen Stimuli, wie beispielsweise Fotografien von menschlichen Gesichtern, das Identifizieren von Exemplaren (z.B. Benennen des Vornamens) schneller gelingt als das Erlernen von Kategorien (z.B. Benennen des Nachnamens) (Medin, Dewey & Murphy, 1983). Des Weiteren hat sich gezeigt, dass das Erlernen einer Attributsausprägung eines Objekts (z.B. hat Flügel) bei linearen Kategorien schneller funktioniert als das Erlernen der richtigen Kategorie eines Objekts (Yamauchi & Markman, 1998) während es sich bei nicht-linearen Kategorien umgekehrt verhält (Yamauchi, Love & Markman, 2002). Ebenso zeigte sich, dass nicht-supervidiertes Lernen leichter gelingt, wenn Stimulusdimensionen interkorrelieren (z.B. Flügel, Federn und Schnabel als interkorrelierende Eigenschaften bei Tieren) (Billman & Knutson, 1996) und Menschen dazu neigen, Stimuli entlang einer einzigen Dimension zu sortieren (Medin, Wattenmaker & Hampson, 1987). In einem nächsten Schritt soll die Evaluation des HDP an diesen Befunden fortgesetzt werden. Hierzu sind zunächst einige Modifikationen am HDPMM und an der Software vorzunehmen. Drei der durchzuführenden Experimente benötigen beispielsweise ein HDPMM, welches drei Werte pro Attributsdimension modellieren kann. Bisher unterstützt npBayes2.1G durch ihre Independent Bernoulli HDP nur binär codierbare Attribute. Da Versuchspersonen im letzterwähnten Experiment instruiert wurden, eine Ansammlung von Stimuli unsupervidiert in zwei gleichgroße Gruppen einzuteilen, ist es zudem notwendig, dem Modell eine feste Anzahl zu bildenden Cluster vorschreiben zu können.

Eine Bewertung der Stärken und Schwächen des Modells einschließlich der Bennung von Verbesserungsmöglichkeiten trägt abschließend nicht nur dem von Wills und Pothos (2012) angesprochenen Evaluationsdefizit aktueller Kategorisierungsmodelle Rechnung, sondern fördert auch die Bewertbarkeit der Bayes'schen Kognition als potentiell vielversprechende Forschungsrichtung. Beides kann langfristig zur Identifikation und Weiterentwicklung geeigneter Kategorisierungsmechanismen beitragen, welche, implementiert in künstlichen Kategorisierungssystemen, menschenähnliches Verhalten in kognitiven Systemen ermöglichen.

LITERATURVERZEICHNIS

Anderson, J. R. (1991) The adaptive nature of human categorization. *Psychological Review*, **98**, 409–429.

Anderson, J. R. & Milson, R. (1989) Human memory: An adaptive perspective. *Psychological Review*, **96**, 703–719.

Billman, D. & Knutson, J. (1996) Unsupervised concept learning and value systematicitiy: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 458–475.

Blunsom, P., Cohn, T., Goldwater, S. & Johnson, M. (2009) A Note on the Implementation of Hierarchical Dirichlet Processes In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. (Su, K.-Y., Su, J., Wiebe, J. & Li, H. (Ed)), 337–340.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992) A Training Algorithm for Optimal Margin Classifiers In: *Proceedings of the fifth annual workshop on Computational learning theory*. (Haussler, D. (Ed)), 144–152. Association for Computing Machinery, Pittsburgh, PA, USA.

Bowers, J. S. & Davis, C. J. (2012) Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, **138**, 389–414.

Carmantini, G. S., Cangelosi, A. & Wills, A. (2014) Machine learning of visual object categorization: an application of the SUSTAIN model In: *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. (Bello, P., Guarini, M., McShane, M. & Scassellati, B. (Ed)), 290–295.

Chater, N. & Oaksford, M. (1999) Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, **3**, 57–65.

Cohen, H. & Lefebvre, C. (2005) *Handbook of categorization in cognitive science*. Elsevier, Amsterdam [etc.].

Cortes, C. & Vapnik, V. (1995) Support-vector networks. *Machine Learning*, **20**, 273-297.

Endress, A. D. (2013) Bayesian learning and the psychology of rule induction. *Cognition*, **127**, 159–176.

Ferguson, T. S. (1973) A Bayesian Analysis of Some Nonparametric Problems. *The annals of statistics*, **1**, 209–230.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013) *Bayesian Data Analysis, Third Edition*. CRC press, Boca Raton, FL.

DOI: 10.17185/duepublico/40721

Geman, S. & Geman, D. (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **PAMI-6**, 721–741.

Gershman, S. J. & Blei, D. M. (2012) A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, **56**, 1–12.

Gilks, W. R. (2005) Markov Chain Monte Carlo In: *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd.

Giraud-Carrier, C. & Provost, F. (2005) Toward a Justification of Meta-learning: Is the No Free Lunch Theorem a Show-stopper? In: *Proceedings of the ICML-2005 Workshop on Meta-learning*. 12–19.

Gopnik, A. & Wellman, H. M. (2012) Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, **138**, 1085–1108.

Griffiths, T. L., Canini, K. R., Sanborn, A. N. & Navarro, D. (2007) Unifying rational models of categorization via the hierarchical Dirichlet process In: *Proceedings of the 29th Annual conference of the Cognitive Science Society*. (McNamara, D. S. & Trafton, J. G. (Ed)), 323–328. Erlbaum, Hillsdale, NJ.

Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. (2008) Bayesian Models of Cognition In: *The Cambridge Handbook of Computational Psychology*. (Sun, R. (Ed)), 59–100. Cambridge University Press, Cambridge, UK.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. (2010) Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, **14**, 357–364.

Griffiths, T. L., Chater, N., Norris, D. & Pouget, A. (2012) How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, **138**, 415–422.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J. & Tenenbaum, J. B. (2011) Nonparametric Bayesian models of categorization In: *Formal Approaches in Categorization*. (Pothos, E. M. & Wills, A. J. (Ed)), 173–198. Cambridge University Press, Cambridge.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G. (eds.) (2010) *Bayesian Nonparametrics*. Cambridge University Press, Cambridge, UK.

Holyoak, K. J. & Cheng, P. W. (2010) Causal Learning and Inference as a Rational Process: The New Synthesis. *Annu. Rev. Psychol.*, **62**, 135–163.

Jones, M. & Love, B. C. (2011) Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, **34**, 169–188.

Kotsiantis, S. B. (2007) Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, **31**, 249–268.

Kruschke, J. K. (1992) ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, **99**, 22–44.

Kruschke, J. K. (2010) *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press, Burlington, MA.

Lattimore, T. & Hutter, M. (2013) No Free Lunch versus Occam's Razor in Supervised Learning In: *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*. (Dowe, D. L. (Ed)): *Papers from the Ray Solomonoff 85th Memorial Conference, Melbourne, VIC, Australia, November 30 – December 2, 2011*, 223–235. Springer, Heidelberg.

Link, W. A. & Eaton, M. J. (2012) On thinning of chains in MCMC. *Methods in Ecology and Evolution*, **3**, 112–115.

Love, B. C., Medin, D. L. & Gureckis, T. M. (2004) SUSTAIN: a network model of category learning. *Psychological review*, **111**, 309–332.

Marcus, G. F. (2010) Neither size fits all: comment on McClelland et al. and Griffiths et al. *Trends in cognitive sciences*, **14**, 346–347.

Marcus, G. F. & Davis, E. (2013) How Robust Are Probabilistic Models of Higher-Level Cognition? *Psychological Science*, **24**, 2351–2360.

McDonnell, J. V. & Gureckis, T. M. (2011) Adaptive clustering models of categorization In: *Formal Approaches in Categorization*. (Pothos, E. M. & Wills, A. J. (Ed)), 220–252. Cambridge University Press, Cambridge.

Medin, D. L., Dewey, G. I. & Murphy, T. D. (1983) Relationships between item and category learning: Evidence that abstraction is not automatic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **9**, 607–625.

Medin, D. L., Wattenmaker, W. D. & Hampson, S. E. (1987) Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, **19**, 242–279.

Nakamura, Tomoaki, Nagai, T. & Iwahashi, N. (2011) Multimodal Categorization by Hierarchical Dirichlet Process In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. (Chen, I.-M., Luca, A. de & Brown, C. J. et al. (Ed)), 1520–1525. IEEE, New York, NY, USA.

Nakamura, T., Yoshiki, A., Takayuki, N. & Masahide, K. (2015) Concept Formation by Robots Using an Infinite Mixture of Models In: *IROS 2015 PROCEEDINGS*. (Zhang, J. (Ed)), 4593–4599.

DOI: 10.17185/duepublico/40721

Navarro, D. J., Griffiths, T. L., Steyvers, M. & Lee, M. D. (2006) Modeling individual differences using Dirichlet processes. *Special Issue on Model Selection: Theoretical Developments and Applications Special Issue on Model Selection: Theoretical Developments and Applications*, **50**, 101–122.

Neal, R. M. (2000) Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Nosofsky, R. M. (1986) Attention, similarity, and the identification-categorization relationship. *Journal of experimental psychology. General*, **115**, 39–61.

Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. (1994a) Rule-plus-exception model of classification learning. *Psychological Review*, **101**, 53–79.

Nosofsky, R., Gluck, M., Palmeri, T., McKinley, S. & Glauthier, P. (1994b) Comparing modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, **22**, 352-369.

Pothos, E. M. & Wills, A. J. (eds.) (2011) *Formal Approaches in Categorization*. Cambridge University Press, Cambridge.

Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. (2013) Probabilistic brains: knowns and unknowns. *Nat Neurosci*, **16**, 1170–1178.

Ross, B. H. & Makin, V. S. (1999) Prototype versus Exemplar Models in Cognition In: *The Nature of Cognition*. (Sternberg, R. J. (Ed)), 205–241. MIT Press, Cambridge, MA.

Sakamoto, Y., Jones, M. & Love, B. (2008) Putting the psychology back into psychological models: Mechanistic versus rational approaches. *Memory & Cognition*, **36**, 1057-1065.

Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. (2010) Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, **117**, 1144–1167.

Schaffer, C. (1994) A Conservation Law for Generalization Performance In: *Machine Learning, Proceedings of the Eleventh International Conference*. (Cohen, W. W. & Hirsh, H. (Ed)), 259–265. Morgan Kaufmann, San Francisco, CA.

Shepard, R. N., Hovland, C. I. & Jenkins, H. M. (1961) Learning and memorization of classifications. *Psychological Monographs: General and Applied*, **75**, 1–42.

Teh, Y. W. (2010) Dirichlet Process In: *Encyclopedia of Machine Learning*. (Sammut, C. & Webb, G. (Ed)), 280-287. Springer US.

Teh, Y. W. & Jordan, M. I. (2010) Hierarchical Bayesian nonparametric models with applications In: *Bayesian Nonparametrics*. (Hjort, N. L., Holmes, C., Müller, P. & Walker, S. G. (Ed)), 158–206. Cambridge University Press, Cambridge, UK.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2004) Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes In: *Advances in Neural Information Processing Systems 17*. (Saul, L. K., Weiss, Y. & Bottou, L. (Ed)): *Proceedings of the 2004 Conference*, 1385–1392. MIT Press, Cambridge, MA.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Waldmann, M. R. (2008) Kategorisierung und Wissenserwerb In: *Allgemeine Psychologie*. (Müsseler, J. (Ed)), 2nd edn., 377–427. Spektrum Akademischer Verlag, Heidelberg.

Wills, A. J. & Pothos, E. M. (2012) On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, **138**, 102–125.

Yamauchi, T., Love, B. C. & Markman, A. B. (2002) Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**, 585–593.

Yamauchi, T. & Markman, A. B. (1998) Category learning by inference and classification. *Journal of Memory and Language*, **39**, 124–148.