Effizientes Lernen aus Vergleichen

Lucas Maystre[1]

1 Motivation

In Auswahlentscheidungen manifestieren sich unsere Meinungen und Präferenzen. Wir treffen eine Auswahl der Musik, die wir hören, und der Filme, die wir sehen. Wir wählen den Ort aus, an dem wir leben, und den politischen Kandidaten, dem wir unsere Stimme geben. Laufend vergleichen wir Alternativen miteinander, um diejenige zu erkennen, die für uns die richtige ist. So überrascht es nicht, dass sich durch Observieren der Ergebnisse solcher Vergleiche ein umfassendes Verständnis für kollektive und persönliche Meinungen erlangen lässt.

Die Idee, menschliche Auswahlentscheidungen zu analysieren, zieht schon seit geraumer Zeit Forscher und Praktiker aus zahlreichen Disziplinen, wie der Psychologie, der Soziologie und den Wirtschaftswissenschaften, in ihren Bann. Um nur ein Beispiel von vielen zu nennen: In der Ökonometrie zählt die Discrete-Choice-Analyse (DCA) inzwischen zum Standardrepertoire. Die DCA hat wichtige Anwendungen. So konnten mit ihrer Hilfe beispielsweise die Auswirkungen einer neuen Stadtbahnlinie in der San Franscisco Bay Area auf die Nutzung unterschiedlicher Verkehrsmittel genau vorhergesagt werden [Mc77]. Die in diesem Zusammenhang entwickelten Theorien und Methoden brachten ihrem hauptsächlichen Erfinder einen Nobelpreis ein [Mc01].

Die vorliegende Arbeit reiht sich ein in die Bemühungen um bessere Methoden zur Analyse von menschlichen Auswahlentscheidungen. Konkret interessieren wir uns für das Problem, wie aus rohen *Auswahldaten* knappe und aussagekräftige Informationen (etwa über unsere Präferenzen) gewonnen werden können. Unter Auswahldaten verstehen wir dabei empirische Daten, die eine von mehreren Alternativen auszeichnen. Konkret könnte eine typische Aufgabe lauten, alle Alternativen in eine Rangfolge von der am meisten zu der am wenigsten bevorzugten Alternative zu gliedern. Oft erfolgt dies anhand von numerischen Wertungen, die den Nutzen der einzelnen Alternativen beschreiben und denen Vorhersagekraft für künftige Entscheidungen zukommt. Obwohl die Forschung zu Auswahlmodellen bereits eine Reihe bewährter Methoden hervorgebracht hat, machen moderne Online-Anwendungen (für die im Weiteren Beispiele angeführt werden) neue Ansätze zum Handhaben großer Datenmengen erforderlich. Tatsächlich werden neue Herausforderungen sowohl durch die

[1] École Polytechnique Fédérale de Lausanne, Schweiz. Korrespondenz: lucas.maystre@epfl.ch. Englischer Originaltitel: *Efficient Learning from Comparisons*.

große Anzahl an *Observationen* als auch durch die große Anzahl an *Alternativen* geschaffen, die für moderne Anwendungen typisch sind. Es wird wichtig, Methoden zu entwickeln, die effizient sind – nicht nur, um schnell alle Observationen zu verarbeiten, sondern auch, um ausreichende Informationen über jede einzelne Alternative zu erhalten. Der *Effizienzgedanke* zieht sich als roter Faden durch die vorliegende Arbeit und wird im Abschnitt 3 weiter ausgeführt.

Wozu Auswahldaten studieren? Wenn es also darum geht, eine numerische Wertung des Nutzens einer jeweiligen Alternative zu erhalten, so stellt sich die berechtigte Frage: Warum nicht einfach *direkt* nach einer solchen Wertung fragen? Dafür gibt es zwei wichtige Gründe:

1. Für Menschen ist es besonders einfach und natürlich, Vergleiche anzustellen. Eine weit verbreitete Theorie in der sozialen Psychologie sagt sogar aus, dass wir uns selbst, unsere Überzeugungen und unsere Meinungen dadurch erkennen und definieren, dass wir uns mit anderen vergleichen [Fe54]. Demgegenüber fällt uns das Abgeben angemessener und konsistenter numerischer Wertungen eher schwerer. Was bedeutet eine Wertung von „3,5 Sternen“ bei einem Restaurant wirklich? In einer Welt, in der alles relativ ist, ist eine absolute Wertung vielleicht einfach die falsche Abstraktion.

2. Manchmal ist es möglich, Auswahlentscheidungen *implizit* zu beobachten, indem wir einfach Handlungen protokollieren sowie den Kontext, in dem diese stattfinden. So lassen sich Auswahldaten auf eine viel unaufdringlichere Art und Weise erlangen als durch *explizites* Fragen nach Feedback. In der Praxis kann so oft Zugriff auf viel größere Datensätze erlangt werden, was potentiell zu genaueren Modellen führt.

Umgang mit inkonsistenten Daten Auf den ersten Blick mag es einfach erscheinen, aus Vergleichsdaten ein Verständnis für die zugrunde liegenden Meinungen zu entwickeln. Tatsächlich wäre das auch so, wenn die beobachteten Auswahlentscheidungen auf perfekte Weise einen einzige Menge von Meinungen widerspiegeln würden. Betrachten wir jedoch „in freier Wildbahn“ gesammelte Daten, so erkennen wir schnell, dass die Ergebnisse von Vergleichen nicht immer miteinander konsistent sind: Anscheinend treffen wir auch bei identischen Alternativen manchmal unterschiedliche Auswahlentscheidungen. Hierfür sind zahlreiche Faktoren verantwortlich, beispielsweise: (*a*) Ein Teil des Kontexts, in dem die Auswahl getroffen wird, wurde eventuell nicht beobachtet, hat jedoch möglicherweise einen wesentlichen Einfluss auf das Ergebnis. (*b*) Bei dem Versuch, kollektive Präferenzen auf der Grundlage individueller Auswahlentscheidungen zusammenzufassen, ist offensichtlich ein gewisses Maß an Nichtübereinstimmung zwischen den Individuen zu erwarten, selbst wenn es auch gemeinsame Trends gibt. (*c*) Manchmal schleichen sich auch Fehler in die Daten ein, die auf fehlerhafte Messungen oder unvollkommene Interpretationen zurückzuführen sind. Die vorliegende Arbeit geht davon aus, dass solche Inkonsistenzen unvermeidbar sind,

es jedoch möglich ist, mit Hilfe eines Wahrscheinlichkeitsmodells systematisch mit ihnen umzugehen. Kurz gefasst beruht der Ansatz darauf, dass bei einem gegebenen Satz von Alternativen *jedes* Vergleichsergebnis möglich ist, aber manche Ergebnisse wahrscheinlicher sind als andere – abhängig von den zugrunde liegenden Präferenzen. Die Aufgabe reduziert sich dann darauf, diejenigen Präferenzen zu ermitteln, die die Observationen gut erklären. Dieser Ansatz dominiert in der Fachwelt und wurde auch für die vorliegende Arbeit gewählt. Er wird im Abschnitt 2 näher erläutert.

Moderne Anwendungen Auswahlmodelle haben eine lange und reichhaltige Tradition, doch im Zusammenhang mit massenhafter Online-Datenerfassung ist das Interesse an ihnen neu aufgelebt. Das Web macht es für Unternehmen einfach, Kunden auf der ganzen Welt zu erreichen und dabei ihre Interaktionen mit dem Leistungsangebot des Unternehmens zu protokollieren. Betrachten wir hierzu drei Beispiele.

- Anbieter kommerzieller Online-Dienste verlassen sich in zunehmendem Maße auf Empfehlungssysteme (d. h. Systeme, die die Präferenzen von Kunden erlernen), um die Kundenbindung zu erhöhen und den Absatz zu steigern. Spotify und Netflix, zwei beliebte Musik- und Videostreaming-Dienste, erlernen Präferenzen anhand von impliziten Observationen der Auswahlentscheidungen der Benutzer (d. h., welche Titel sie hören bzw. welche Filme sie sehen). Der E-Commerce-Riese Amazon unterbreitet seinen Kunden personalisierte Kaufempfehlungen, die auf früheren Einkäufen basieren.

- Wissenschaftler haben Online-Plattformen erstellt, mit denen sie große Mengen an Vergleichsdaten sammeln können, um schwierige Fragen aus der psychologischen und soziologischen Forschung zu beantworten. So hat sich zum Beispiel das GIFGIF-Projekt[2] zum Ziel gesetzt, den emotionalen Inhalt animierter GIF-Bilder zu verstehen, indem Benutzern Bildpaare gezeigt und die Benutzer gefragt werden, welches Bild eine Emotion wie Freude, Scham usw. besser ausdrückt. Das Place Pulse-Projekt[3] möchte verstehen, wie Stadtviertel wahrgenommen werden, und benutzt dazu ebenfalls paarweise Vergleichsfragen. In beiden Fällen sind Vergleiche ein natürliches Mittel, um Feedback von Benutzern zu erhalten. Beide Projekte haben Millionen von Datenpunkten über Tausende von Objekte gesammelt und faszinierende Erkenntnisse geliefert, die früher mit herkömmlichen Methoden nicht zu erreichen waren.

- Paarweise Vergleiche bilden die Grundlage von *Wiki-Surveys*, einer neuartigen Befragungsmethodik, die von Salganik; Levy [SL15] entwickelt wurde. Wiki-Surveys sind der Versuch, die Lücke zwischen Fragebögen auf der einen und mündlichen Befragungen auf der anderen Seite zu schließen – Fragebögen skalieren gut, bieten aber keinen Raum für das Hervortreten neuer Informationen. Mündliche Befragungen

[2] Siehe: `http://www.gif.gf/`.
[3] Siehe: `http://pulse.media.mit.edu/`.

können unverhoffte neue Erkenntnisse liefern, sind aber kostspielig in der Durchführung. Beispielsweise hat die Stadtverwaltung von New York City mit Wiki-Surveys Feedback zu einem Nachhaltigkeitsplan eingeholt. Die Respondenten konnten entweder neue Ideen vorschlagen oder eine Vergleichsfrage der Art „Welche der folgenden beiden Ideen ist Ihrer Meinung nach besser geeignet, eine grünere und bessere New York City zu schaffen?“ beantworten. Der Wiki-Survey-Dienst macht es möglich, gleichzeitig sowohl neue Vorschläge zu erhalten als auch bekannte Vorschläge nach ihrer Priorität zu ordnen. Zum Zeitpunkt der Entstehung der vorliegenden Arbeit gab es auf `http://www.allourideas.org/` bereits 11 739 Wiki-Surveys mit insgesamt 17.8 Millionen Stimmen zu 631 682 Ideen.

Mehr als Präferenzen: Anwendungen im Sport Abschließend sei darauf hingewiesen, dass dieselben Methoden, mit denen menschliche Auswahlentscheidungen modelliert werden, auch für Probleme benutzt werden können, die auf den ersten Blick konzeptuell völlig andersartig zu sein scheinen. So behandelt die vorliegende Arbeit auch das Problem der Vorhersage von Fußballergebnissen auf der Grundlage historischer Daten. Bei einem Fußballspiel werden zwei Mannschaften miteinander verglichen, und am Ende gewinnt eine davon. In unserer Terminologie können wir die Mannschaften als Alternativen betrachten, die miteinander verglichen werden, und den Sieger als das Ergebnis des Vergleichs. Interessanterweise haben sich die wesentlichen Modelle und Ideen, die in der vorliegenden Arbeit verwendet werden, gleichzeitig sowohl im Kontext der Analyse menschlicher Auswahlentscheidungen als auch im Kontext der Vorhersage von Wettkampfergebnissen entwickelt, wie wir im nächsten Abschnitt sehen werden.

2 Ausgewählte Wahrscheinlichkeitsmodelle

In diesem Abschnitt werden die statistischen Modelle und zugehörigen Methoden vorgestellt, die im Rahmen der vorliegenden Arbeit benutzt werden oder auf die Bezug genommen wird. Wir nehmen einen historischen Blickwinkel ein: Der Kontext, in dem diese Modelle und Verfahren entstanden sind, ist faszinierend. Dieser Abschnitt liefert nur einen kurzen Überblick über die Entwicklungen, enthält jedoch Verweise auf umfassendere Informationen für den interessierten Leser.

2.1 Thurstones Modell

Im Jahre 1927 veröffentlichte Thurstone einen Artikel, der weithin als grundlegend für das Gebiet der Wahrscheinlichkeitsmodelle für Vergleichsergebnissen angesehen wird [Th27]. Thurstone interessierte sich für das Problem der Messung in der Psychologie und entwickelte ein Verfahren, das die Antworten von menschlichen Probanden auf Vergleiche zwischen zwei von N Stimuli erklärt. Um die Tatsache zu berücksichtigen, dass die Reaktion

auf einen Stimulus variieren kann, schlug Thurstone vor, den wahrgenommenen Wert eines Stimulus i während eines Experiments mittels einer *zufälligen* Variablen $x_i \in \mathbf{R}$ zu modellieren. Das Ergebnis des Vergleichs zwischen den Stimuli i und j ist durch eine Realisierung der entsprechenden zwei Zufallsvariablen gegeben, d. h. durch das Ereignis $x_i > x_j$. Thurstone postulierte ferner, dass diese Zufallsvariablen einer multivariaten Gauss-Verteilung $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ folgen. Hierbei bezeichnet $i \succ j$ das Ereignis „i wird gegenüber über j bevorzugt“.

$$\mathbf{P}\left[i \succ j\right] = \mathbf{P}\left[x_i > x_j\right] = \Phi\left(\frac{\theta_i - \theta_j}{\sqrt{\Sigma_{ii} + \Sigma_{jj} - 2\Sigma_{ij}}}\right).$$

Hierbei ist $\Phi(\cdot)$ die kumulative Dichtefunktion der Standardnormalverteilung. Die letzte Gleichung ergibt sich daraus, dass $x_i - x_j \sim \mathrm{N}(\theta_i - \theta_j, \Sigma_{ii} + \Sigma_{jj} - 2\Sigma ij)$. Thurstone betrachtete mehrere Varianten des Modells, die Schritt für Schritt restriktivere Annahmen über die Kovarianzmatrix $\boldsymbol{\Sigma}$ machen. Die heutzutage am häufigsten benutzte Variante erhält man durch Einsetzen von $\boldsymbol{\Sigma} = \frac{1}{2}\boldsymbol{I}$. In diesem Fall ist

$$\mathbf{P}\left[i \succ j\right] = \Phi(\theta_i - \theta_j). \tag{1}$$

Der Vektor der N Parameter $\boldsymbol{\theta} = [\theta_1 \ \cdots \ \theta_N]^\top \in \mathbf{R}^N$ bestimmt die Wahrscheinlichkeiten aller $\binom{N}{2}$ möglichen paarweisen Vergleiche. Intuitiv lässt sich θ_i als die *Wertung* des Stimulus i verstehen, und die Wahrscheinlichkeit eines Vergleichsergebnisses für i und j, das mit der tatsächlichen Reihenfolge konsistent ist, steigt mit der Distanz $\theta_i - \theta_j$. Man beachte, dass, weil (1) nur paarweise Distanzen enthält, die Parameter $\boldsymbol{\theta}$ nur bis auf eine Konstante bestimmt werden können. Um diese Unbestimmtheit aufzulösen, werden die Parameter oft derart gewählt, dass $\sum_i \theta_i = 0$.

Die vielleicht erste Anwendung, die Thurstone im Sinn hatte, betrifft das Gebiet der Psychophysik. Man stelle sich vor, dass man zwei Bälle erhält und gefragt wird: „Welcher dieser beiden Bälle ist schwerer?“ Hat man eine Sammlung von Observationen dieser Art (von denen wohl einige inkonsistent sein werden), könnte das Modell (1) benutzt werden, um durch Schätzen der Parameter $\boldsymbol{\theta}$ die Stimuli auf einer reellwertigen Skala einzuordnen (wodurch alle Daten kompakt zusammengefasst werden).

2.2 Bradley-Terry-Modell

Fast zeitgleich mit Thurstone schlug Zermelo (in deutscher Sprache) eine Methode zum Bewerten von Schachspielern basierend auf Spielergebnissen vor [Ze28]. Er betrachtete zwei Probleme: (*a*) Den Umgang mit *unausgeglichenen* Tournieren, bei denen Spieler eine ungerade Anzahl von Spielen gegen verschiedene Gegnergruppen spielen. (*b*) Das Schätzen der *relativen Stärke* der Spieler derart, dass die Schätzung Vorhersagekraft für künftige Spielergebnisse hat. Dazu führte er ein Wahrscheinlichkeitsmodell für die Spielergebnisse ein. In seinem Modell ist jeder Spieler $i \in [N]$ durch einen latenten Stärkeparameter

$\gamma_i \in \mathbf{R}_{>0}$ charakterisiert. Die Wahrscheinlichkeit, dass Spieler i gegen Spieler j gewinnt, ist eine Funktion der relativen Stärken der Spieler:

$$\mathbf{P}[i > j] = \frac{\gamma_i}{\gamma_i + \gamma_j}. \tag{2}$$

Es sei angemerkt, dass die Parameter $\boldsymbol{\gamma}$ nur bis auf einen multiplikativen Faktor bestimmt werden können. Aus diesem Grund wird oft angenommen, dass $\sum_i \gamma_i = 1$. Zermelo schlug vor, die Parameter $\boldsymbol{\gamma}$ durch Maximieren ihrer Likelihood angesichts der beobachteten Daten zu finden, eine für die damalige Zeit sehr fortschrittliche Idee. Er formulierte eine notwendige und hinreichende Bedingung[4] für die Existenz einer eindeutigen Maximum-Likelihood-Schätzung, entwickelte einen iterativen Algorithmus zu ihrer Ermittlung und bewies, dass der Algorithmus konvergiert. Insgesamt behandelt er das Modell sehr gründlich und vollständig; leider wurde es anscheinend für ungefähr 50 Jahre größtenteils ignoriert [Da88]. Eine spannende Einführung in die Arbeit von Zermelo findet sich bei Glickman [Gl13]. Abschließend sei angemerkt, dass das aktuell vom Weltschachbund eingesetzte Wertungssystem direkt auf Zermelos Modell basiert [El78].

Verhältnis zu Thurstones Modell Fast zwei Jahrzehnte später entdeckten Bradley; Terry [BT52], die Zermelos Arbeit offensichtlich nicht kannten, das Modell im Kontext der Ranganalyse von Experimenten auf der Grundlage paarweiser Vergleiche wieder, und verbanden somit das Modell wieder mit der Analyse von menschlichen Meinungen. Die Verbindung zu Thurstones Modell wurde in Bradley [Br53] deutlich, wo Bradley zeigt, dass sich durch Einsetzen von $\theta_i = \log \gamma_i$ für alle i die Wahrscheinlichkeit (2) schreiben lässt als

$$\mathbf{P}[i > j] = \frac{1}{1 + \exp[-(\theta_i - \theta_j)]}. \tag{3}$$

Das Bradley-Terry-Modell (wie es in der Regel bezeichnet wird) ist somit ein weiterer Fall eines generalisierten linearen Modells [Ag15] für paarweise Vergleiche: Die Wahrscheinlichkeit eines Ergebnisses hängt von der Distanz $\theta_i - \theta_j$ zwischen zwei Parametern ab, die den Wertungen der Alternativen entsprechen. Yellot [Ye77] arbeitete die Verbindung weiter heraus, indem er zeigte, dass $\mathbf{P}[i > j]$ in (3) als $\mathbf{P}\left[x_i > x_j\right]$ für unabhängige zufällige Variablen $\{x_k : k \in [N]\}$ umgeschrieben werden kann, so dass $x_k \sim \text{Gumbel}(\theta_k, 1)$, das heißt, $\mathbf{P}[x_k \leq y] = \exp\{-\exp[-(y - \theta_k)]\}$. Ergebnisse kann man sich daher auch als den Vergleich der Realisierungen zweier zufälliger Variablen vorstellen, die um die Wertungen der Alternativen zentriert sind, woraus eine Interpretation im Rahmen von *Random Utility* folgte. Schließlich zeigte Stern [St92], dass sich das Thurstone-Modell und das Bradley-Terry-Modell zu einem einheitlichen Modell verallgemeinern lassen, wobei die Gamma-Verteilung benutzt wird. In der Praxis liefern beide Modelle in den meisten Fällen quantitativ ähnliche Ergebnisse [TG11].

[4] Die Maximum-Likelihood-Schätzung existiert, wenn und nur wenn es keinen Weg gibt, alle Spieler derart in zwei disjunkte, nicht leere Teilmengen $A, B \subset [N]$ zu unterteilen, dass es keinen Spieler in A gibt, der ein Spiel gegen einen Spieler in B gewonnen hat.

2.3 Das Auswahl-Axiom von Luce

Die zwei vorstehend besprochenen Modelle sind auf Vergleiche zwischen Elemente-*Paaren* beschränkt. Wie lassen sich diese Modelle auf *multivariate* Vergleiche verallgemeinern? Gegeben sei eine Menge von Alternativen $\mathcal{A} \subseteq [N]$ und ein Element $i \in \mathcal{A}$, und dabei bezeichne $i \succeq \mathcal{A}$ das Ereignis „i wird unter den Alternativen $\mathcal{A}$ ausgewählt“. Eine natürliche Art und Weise, das Bradley-Terry-Modell (2) auf die Auswahl unter beliebig vielen Alternativen zu erweitern, lautet dann wie folgt:

$$\mathbf{P}[i \succeq \mathcal{A}] = \frac{\gamma_i}{\sum_{j \in \mathcal{A}} \gamma_j}. \tag{4}$$

Einfach ausgedrückt ist die Wahrscheinlichkeit für eine Auswahl immer proportional zu der Stärke γ_i des Elements i, egal welche Menge an Alternativen vorliegt. Dieses Auswahlmodell geht auf Luce [Lu59] zurück, der zeigte, dass es eng mit der nachstehenden Eigenschaft zusammenhängt.

Definition (Unabhängigkeit irrelevanter Alternativen)**.** Ein probabilistisches Auswahlmodell erfüllt die Eigenschaft der *Unabhängigkeit von irrelevanten Alternativen* (UIA), wenn für jedes $\mathcal{A} \subseteq [N]$ und jedes $i, j \in \mathcal{A}$ gilt:

$$\frac{\mathbf{P}[j \succeq \mathcal{A}]}{\mathbf{P}[i \succeq \mathcal{A}]} = \frac{\mathbf{P}[j \succ i]}{\mathbf{P}[i \succ j]}.$$

Die UIA-Eigenschaft ist im Wesentlichen äquivalent[5] zum *Auswahl-Axiom* von Luce (1959, S. 6), und im Rahmen der vorliegenden Arbeit nehmen wir in austauschbarer Weise auf diese beiden Konzepte Bezug. Der wesentliche Beitrag von Luce bestand darin, zu zeigen, dass die UIA-Eigenschaft eine axiomatische Charakterisierung der Auswahl-Wahrscheinlichkeiten erlaubt.

Satz 1 ([Lu59])**.** *Ein Auswahlmodell erfüllt die UIA-Eigenschaft, wenn und nur wenn ein Vektor $\boldsymbol{\gamma} \in \mathbf{R}_{>0}$ existiert, so dass die Auswahl-Wahrscheinlichkeiten durch* (4) *gegeben sind.*

Die Unabhängigkeit von irrelevanten Alternativen ist eine mächtige Eigenschaft, da sie zu einem Auswahlmodell führt, das *kombinatorisch* viele Auswahlwahrscheinlichkeiten mit Hilfe von lediglich N Parametern repräsentiert. Dies ermöglicht die Ermittlung von Auswahlwahrscheinlichkeiten anhand einer möglicherweise kleinen Anzahl von Observationen. Es schränkt jedoch unvermeidlich die Aussagekraft des Modells ein. In Fällen, in denen einige Alternativen sehr ähnlich sind, kann UIA eine unrealistische Annahme sein, wie Debreu [De60] anhand eines einfachen Beispiels zeigt. Im Kontext moderner Anwendungen mit einer großen Anzahl von Elementen, worauf der Schwerpunkt der vorliegenden Arbeit liegt, halten wir diesen Kompromiss für akzeptabel (und vielleicht sogar für notwendig).

[5] Luce [Lu59] führt die UIA-Eigenschaft als Konsequenz des Auswahl-Axioms ein, was etwas allgemeiner ist: Seine Formulierung lässt auch $\mathbf{P}[i \succeq \mathcal{A}] = 0$ zu, ein Detail, das in der vorliegenden Arbeit unberücksichtigt bleibt.

3 Überblick und Beiträge

Die vorliegende Arbeit behandelt das Problem, auf *effiziente* Weise eine Rangfolge für eine Menge von Elementen zu ermitteln (was in der Regel durch das Schätzen von Auswahlmodell-Parametern erfolgt). Effizienz ist dabei der rote Faden.

- Je weiter die Größe der Datensätze ansteigt, desto wichtiger wird es, Inferenz-Methoden zu entwickeln, die *rechnerisch* effizient sind, ohne dabei Abstriche bei der *statistischen* Effizienz, d. h. der Genauigkeit, zu machen.
- Bei hohen Anzahlen unterschiedlicher Elemente wird es wichtig, Observationen mit Bedacht derart zu nehmen, dass die Observationen so viel Informationen wie möglich beitragen. Dies bezeichnen wir als *Daten*-Effizienz.

In Kapitel 2 konzentrieren wir uns auf Algorithmen für Parameter-Inferenz und entwickeln zwei Verfahren für Modelle auf der Grundlage des Auswahl-Axioms von Luce. Dazu wird das Inferenz-Problem als das Problem gefasst, die stationäre Verteilung einer Markow-Kette zu finden – ein Ansatz, der von Negahban et al. [NOS12] bereits im Kontext paarweiser Vergleiche vorgeschlagen wurde. Die Ermittlung der stationären Verteilung einer Markow-Kette ist ein gut erforschtes Problem, und es stehen schnelle Löser zur Verfügung. Zunächst wird gezeigt, wie die Markow-Kette aus der Likelihood-Funktion abgeleitet werden kann – eine wesentliche Erkenntnis, welche die Verallgemeinerung der Ideen von Negahban et al. auf andere auf dem Auswahl-Axiom von Luce basierende Modelle ermöglicht. Der erste Algorithmus, LSR, ermittelt eine *spektrale* Schätzung der Modellparameter durch Lösen einer homogenen Markow-Kette: Er ist rechnerisch sehr effizient, und die Schätzung erweist sich als akkurater als Schätzungen alternativer Methoden mit vergleichbarer Laufzeit. Der zweite Algorithmus, I-LSR, ermittelt die Maximum-Likelihood-Schätzung (MLE) durch Lösen einer nichthomogenen Markow-Kette. Die MLE ist statistisch effizienter als die spektrale Schätzung, jedoch auch rechenintensiver. Doch selbst dann erweist sich I-LSR als deutlich schneller als andere oft verwendete Algorithmen zum Ermitteln der MLE.

In Kapitel 3 wenden wir unsere Aufmerksamkeit der Aufgabe zu, auf „intelligente“ Weise Ergebnisse von paarweisen Vergleichen zu sammeln, und zwar basierend auf den observierten Ergebnissen vorheriger Vergleiche. Unter der Annahme, dass wir adaptiv auswählen können, welches Paar von Elementen zu jedem Zeitpunkt abgefragt wird, streben wir danach, die über das Modell (insbesondere über die Rangfolge der N Elemente) erhaltenen Informationen zu maximieren und gleichzeitig die Anzahl zu minimieren. In der Literatur über maschinelles Lernen ist dies als das Problem des *aktiven Lernens* bekannt [Se12]. Wir starten mit einer Analyse von Quicksort [Ho62], einem bekannten Sortieralgorithmus, der eine Rangfolge berechnet, wobei Vergleiche stets mit der tatsächlichen Reihenfolge konsistent sind. Unter einigen natürlichen Annahmen über die Verteilung von Bradley-Terry-Modellparametern (welche die Schwierigkeit von Rangfolgen charakterisieren) zeigen wir, dass Quicksort erstaunlich resilient gegenüber inkonsistenten Vergleichsergebnissen ist.

Dies führt uns zu einer praktischen und dateneffizienten Abfragestrategie, die wiederholt einen Sortieralgorithmus ausführt, bis ein vorgegebenes Vergleichsbudget verbraucht ist. Im Bezug auf andere Strategien des aktiven Lernens erreicht die vorgeschlagene Methode eine vergleichbare Dateneffizienz, ist aber wesentlich weniger rechenintensiv.

In Kapitel 4 betrachten wir ein Szenario, bei dem Auswahlentscheidungen in einem Netzwerk stattfinden, was durch die Arbeit von Kumar et al. [Ku15] inspiriert ist. Es geht darum zu verstehen, wie Benutzer in einem Netzwerk navigieren (z. B. auf welche Links sie im Web klicken), und zwar unter der Annahme, dass wir Zugriff auf den aggregierten Datenverkehr an jedem Knoten im Netzwerk haben, jedoch nicht auf die individuellen Entscheidungen (d. h. die eigentlichen Übergänge). Wenn die Übergänge das Auswahl-Axiom von Luce erfüllen, können wir zeigen, dass der aggregierte Datenverkehr eine ausreichende statistische Größe für die Übergangswahrscheinlichkeiten ist. Als Nächstes entwickeln wir einen Inferenz-Algorithmus, der (*a*) robust gegenüber verschiedenen mangelhaft gestellten Szenarien ist und (*b*) effizient implementiert werden kann. Zum Beispiel skaliert der Algorithmus erfolgreich bis zu einem Schnappschuss eines WWW-Hyperlink-Graphen mit Milliarden Knoten. Schließlich zeigen wir anhand von realen Klickstream-Daten, dass die vorgeschlagene Methode Übergangswahrscheinlichkeiten gut schätzen kann, und zwar trotz der starken Annahmen, die das Axiom von Luce impliziert.

Schließlich verlassen wir in Kapitel 5 das Gebiet der menschlichen Meinungen und betrachten eine Anwendung in der Welt des Sports. Konkret beschäftigen wir uns mit der Vorhersage der Ergebnisse von Fußballspielen zwischen Nationalmannschaften. Dies ist ein schwieriges Problem, weil Nationalmannschaften nur wenige Spiele pro Jahr absolvieren, so dass ihre Stärke sich nur schwer allein aus den Ergebnissen der von ihnen gespielten Spiele schätzen lässt. Doch berücksichtigen wir, dass die meisten Nationalspieler in Spielen ihrer Clubs gegeneinander antreten, und versuchen, die (vergleichsweise) große Anzahl an Spielen zwischen den Clubs zu nutzen, um die Vorhersagen zu verbessern. Dazu stellen wir alle Spiele in einem *Spieler-Raum* dar und sorgen mit einer Kernel-Methode dafür, dass die Modellinferenz rechnerisch handhabbar wird. Wir evaluieren die erhaltene Vorhersage anhand von Daten der letzten drei Europameisterschaften. Dabei stellen wir fest, dass die Vorhersagen auf Grundlage des kombinierten Modells exakter sind als die Vorhersagen, die lediglich auf den Ergebnissen der Spiele zwischen den Nationalmannschaften beruhen.

Literaturverzeichnis

[Ag15] Agresti, A.: Foundations of Linear and Generalized Linear Models. Wiley, 2015.

[Br53] Bradley, R. A.: Some Statistical Methods in Taste Testing and Quality Evaluation. Biometrics 9/1, S. 22–38, 1953.

[BT52] Bradley, R. A.; Terry, M. E.: Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 39/3/4, S. 324–345, 1952.

[Da88] David, H. A.: The Method of Paired Comparisons. Charles Griffin & Company, 1988.

[De60] Debreu, G.: Review of Individual Choice Behavior: A Theoretical Analysis. The American Economic Review 50/1, S. 186–188, 1960.

[El78] Elo, A.: The Rating Of Chess Players, Past & Present. Arco Publishing, 1978.

[Fe54] Festinger, L.: A Theory of Social Comparison Processes. Human Relations 7/2, S. 117–140, 1954.

[Gl13] Glickman, M. E.: Introductory note to 1928 (= 1929). In: Ernst Zermelo - Collected Works II. Springer, S. 616–671, 2013.

[Ho62] Hoare, C. A. R.: Quicksort. The Computer Journal 5/1, S. 10–16, 1962.

[Ku15] Kumar, R.; Tomkins, A.; Vassilvitskii, S.; Vee, E.: Inverting a Steady-State. In: Proceedings of WSDM'15. Shanghai, China, Feb. 2015.

[Lu59] Luce, R. D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, 1959.

[Mc01] McFadden, D.: Economic Choices. American Economic Review 91/3, S. 351–378, 2001.

[Mc77] McFadden, D.; Talvitie, A.; Cosslett, S.; Hasan, I.; Johnson, M.; Reid, F.; Train, K.: Demand Model Estimation and Validation, Techn. Ber., Institute of Transportation Studies, University of California, Berkeley, 1977.

[NOS12] Negahban, S.; Oh, S.; Shah, D.: Iterative Ranking from Pair-wise Comparisons. In: Advances in Neural Information Processing Systems 25. Lake Tahoe, CA, Dez. 2012.

[Se12] Settles, B.: Active Learning. Morgan & Claypool Publishers, 2012.

[SL15] Salganik, M. J.; Levy, K. E. C.: Wiki Surveys: Open and Quantifiable Social Data Collection. PLoS ONE 10/5, S. 1–17, 2015.

[St92] Stern, H.: Are all linear paired comparison models empirically equivalent? Mathematical Social Sciences 23/1, S. 103–117, 1992.

[TG11] Tsukida, K.; Gupta, M. R.: How to Analyze Paired Comparison Data, Techn. Ber., Seattle, WA, USA: University of Washington, Mai 2011.

[Th27] Thurstone, L. L.: A Law of Comparative Judgment. Psychological Review 34/4, S. 273–286, 1927.

[Ye77] Yellot Jr., J. I.: The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the Double Exponential Distribution. Journal of Mathematical Psychology 15/2, S. 109–144, 1977.

[Ze28] Zermelo, E.: Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift 29/1, S. 436–460, 1928.

Lucas Maystre erhielt M.Sc. und Ph.D. Abschlüsse der Eidgenössische Technische Hochschule Lausanne (EPFL), Schweiz bzw. 2012 und 2018. Derzeit ist er wissenschaftlicher Mitarbeiter und Mitglied der Gruppe *Satisfaction, Interaction and Algorithms* bei Spotify, London, Vereinigtes Königreich. Seine Forschungsinteressen liegen im Bereich des maschinellen Lernens, insbesondere der probabilistischen Modellierung und des effizienten Algorithmusdesigns. Dr. Maystre wurde 2016 mit dem Google Fellowship in Machine Learning ausgezeichnet und erhielt 2018 die EPFL Auszeichnung für Abschlussarbeiten. Zusammen mit einem Kollegen der EPFL betreibt er bei `https://kickoff.ai` eine Fußball-Vorhersageplattform.