Karsten Borgwardt, der Spezialist für Data Mining

Mit den modernen Methoden der Genomik, Proteomik und Metabolomik werden in immer kürzerer Zeit immer größere „Datenberge" produziert. Um durch Mustererkennung daraus relevante Informationen zu gewinnen, sind spezielle mathematische und informatorische Verfahren gefragt. Der Tübinger Data-Mining-Spezialist Karsten Borgwardt entwickelt sie speziell für Anwendungen in den Lebenswissenschaften.

Mit 30 schon Professor: Der Data-Mining-Spezialsit Karsten Borgwardt überzeugt bereits in jungen Jahren mit außerordentlichen Leistungen. © Max-Planck-Institute Tübingen

Die Analogie mit dem Bergbau ist treffend: Es gilt, aus Datenmassen jeweils das "Datengold" zu extrahieren, das zu neuen Erkenntnissen führt. Dabei fallen in den Lebenswissenschaften heute immer größere Datenmengen in immer kürzerer Zeit an. Ein Beispiel: Während die erste Sequenzierung eines menschlichen Genoms bis zum Abschluss des Projektes noch über ein Jahrzehnt gedauert und Unsummen verschlungen hat – Schätzungen gehen von 100 Millionen US-Dollar aus – , spucken heutige Geräte des „Next Generation Sequencing" innerhalb weniger Wochen die Sequenzen ganzer Genome von Menschen, Tieren und Pflanzen zu moderaten Kosten im vierstelligen Dollar-Bereich aus.

Noch schneller und günstiger sind Genotypisierungen in der medizinischen Diagnostik. Dabei wird im Erbgut des Menschen nach speziellen Varianten im Genom, den Einzelnukleotid-Polymorphismen (SNP, engl. single nucleotide polymorphism), gesucht, die immer wieder bei erkrankten Personen zu finden sind und deshalb mit bestimmten Krankheiten korrelieren. So richtig komplex wird es, wenn nicht, wie im Fall der Sichelzellenanämie, ein einzelner SNP zu einer Krankheit führt, sondern wenn mehrere SNPs in einem bestimmten Muster auftreten müssen, um eine Krankheit auszulösen oder das Erkrankungsrisiko zu erhöhen. Im letzteren Fall stellt sich auch die Frage, inwieweit muss ein Muster vollständig sein, um mit einer bestimmten Wahrscheinlichkeit eine Krankheit auszulösen. Auch die Verträglichkeit von Medikamenten und das Ansprechen auf eine Therapieform versucht man mit SNPs bzw. SNP-Mustern in Verbindung zu setzen.

Kombistudium Biologie und Informatik

Solche Fragestellungen sind das Metier von Prof. Dr. Karsten Borgwardt. Der 32-jährige Informatiker forscht an den Tübinger Max-Planck-Instituten für Entwicklungsbiologie und für Intelligente Systeme. Außerdem ist er Professor für Data Mining in den Lebenswissenschaften an der Universität Tübingen. „Für mich heißt Data Mining, dass ich die speziellen Eigenschaften von Daten ausnutze, um effiziente Algorithmen für statistische Analysen in den Lebenswissenschaften zu entwickeln“, definiert Borgwardt seine Arbeit.

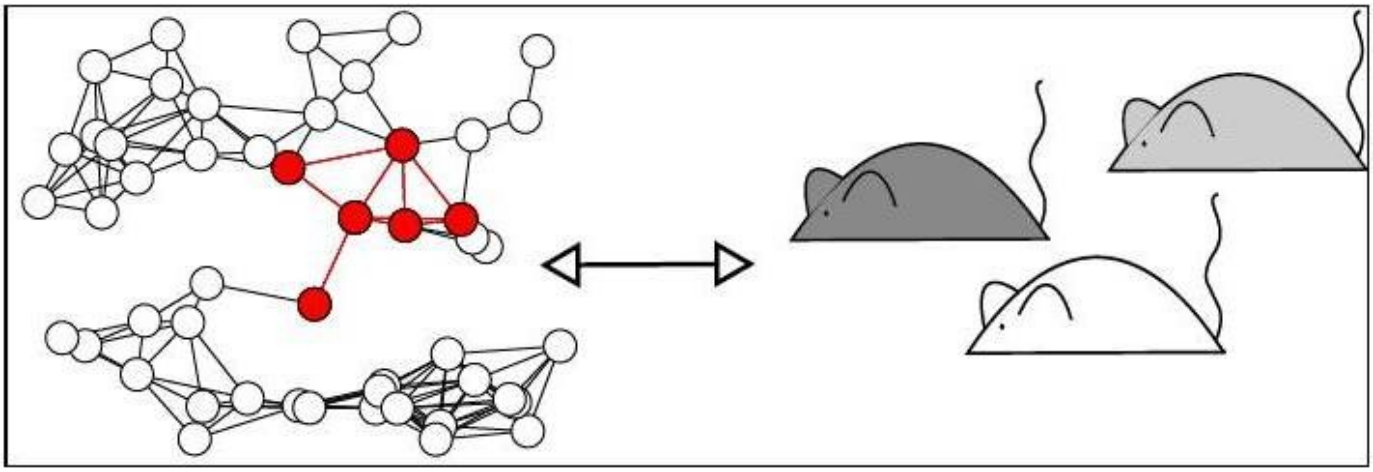
Er hat bereits während des Informatik-Studiums an der Ludwig-Maximilians-Universität (LMU) München im Nebenfach Biologie studiert und zwischendurch, während eines Auslandsjahres an der University of Oxford, einen Master in Biologie erworben. Damit hatte er optimale Voraussetzungen für ein Fachgebiet, das ihn schon lange in den Bann gezogen hatte. „Schon als Jugendlicher haben mich die Möglichkeiten fasziniert, die sich durch die Sequenzierung von Genomen ergeben und mir war bewusst, dass die gewaltigen Datenmengen mathematisch-informatorische Probleme mit sich bringen. Während der Diplomarbeit wurde mir klar, dass es viele algorithmische Probleme gibt mit Anwendungen in der Biologie“, so Borgwardt.

Außergewöhnliche Leistungen ermöglichen Karriere im Laufschritt

Seine Kombi-Karriere ging rasant weiter: Borgwardt promovierte nach nur zweieinhalb Jahren mit Auszeichnung auf dem Gebiet des Data Mining an der LMU München. Danach ging er als wissenschaftlicher Mitarbeiter an die University of Cambridge und befasste sich mit maschinellem Lernen in der Biologie. Borgwardt arbeitete daran, computergestützte Systeme zu entwickeln, die Muster und Gesetzmäßigkeiten in Datenströmen erkennen und "lernen" konnten, daraus Gesetzmäßigkeiten abzuleiten. Zu dieser Zeit hatte Borgwardt über die internationalen wissenschaftlichen Netzwerke bereits Kontakt zu den Max-Planck-Instituten in Tübingen geknüpft. Zwei ihrer Direktoren, Prof. Dr. Bernhard Schölkopf und Prof. Dr. Detlef Weigel, erkannten das Potenzial des jungen Forschers und überzeugten ihn davon, nach Tübingen zu wechseln. Wobei allzu viel Überzeugungskraft wohl nicht vonnöten war, denn Borgwardt kannte den Ruf der Tübinger Bioinformatik: „Tübingen hat im Bereich des Maschinellen Lernens und der Bioinformatik sehr viel zu bieten, es gibt weltweit kaum einen anderen Standort, an dem maschinelles Lernen so gut mit der Molekularbiologie verknüpft ist.“

Kaum in Tübingen angekommen, wurde Borgwardt nach nicht mal einem Jahr schon Forschungsgruppenleiter und rund zwei Jahre später erhielt er 2011 die Professur für Data Mining. Damit steht er für den dritten Weg, auf dem es in Deutschland möglich ist, Professor in den Natur- und Ingenieurwissenschaften zu werden, denn Borgwardt hat weder habilitiert noch war er Juniorprofessor. Vielmehr konnte er durch seine außergewöhnliche wissenschaftliche Leistung überzeugen und dadurch, dass er als Forschungsgruppenleiter schon so erfolgreich war, dass er Doktoranden führender Universitäten aus Europa, Asien und den USA anzog. Seit Januar 2013 koordiniert er zudem das europaweite Marie-Curie-Netzwerk für "Maschinelles Lernen in der personalisierten Medizin". „Ziel des Netzwerkes ist es, neue statistische und algorithmische Verfahren zu entwickeln, um eine maßgeschneiderte Therapie des Einzelnen zu ermöglichen. Bei dieser Forschung stehen wir bei unterschiedlichen Krankheiten an sehr unterschiedlichen Punkten“, sagt Borgwardt.

Alfried Krupp-Preis: eine Million Euro für weitere Forschung

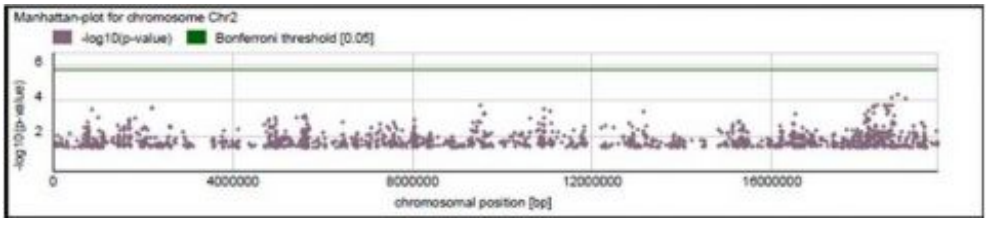


Die Gruppe um Borgwardt will in Zukunft ganze Netzwerke genetischer Veränderungen darauf untersuchen, ob sie mit Krankheiten oder anderen Merkmalen korrelieren. Hier als Beispiel ein hypothetisches Subnetzwerk von SNPs (links rot markiert), das die Fellfarbe von Mäusen beeinflusst. © Dr. C.-A. Azencott

Aktuell werden Borgwardts Forscherqualitäten dadurch unterstrichen, dass er den mit einer Million Euro verbundenen Alfried Krupp-Förderpreis 2013 erhält. Diese Auszeichnung ist einer der höchstdotierten Forschungspreise für junge Professoren in Europa. Die feierliche Preisübergabe wird im November 2013 in der Villa Hügel in Essen, dem Sitz der Alfried Krupp von Bohlen und Halbach-Stiftung, stattfinden. Das Preisgeld will Borgwardt zum größten Teil dazu nutzen, seine Forschergruppe personell zu verstärken. „Ich habe keine großen experimentellen Ausgaben, deshalb kann ich die Gelder primär dafür nutzen, Doktorandenstellen zu schaffen. Ein kleiner Teil des Geldes wird dafür verwendet, unsere Rechner-Ressourcen zu erweitern“, so Borgwardt.

So wird der Alfried Krupp-Preis dazu beitragen, die SNP-Analysen der Borgwardt-Gruppe voranzutreiben. Das Team befasst sich unter anderem mit grundlegenden statistischen Fragen dazu, welche SNPs am häufigsten mit dem Auftreten einer bestimmten Krankheit korrelieren. „Wenn wir zum Beispiel die SNPs von 10.000 Patienten mit denen von 10.000 gesunden Kontrollpersonen vergleichen, dann müssen wir in der Regel Hunderttausende SNPs untersuchen. Mithilfe von Manhattan Plots stellen wir dar, wie stark die SNPs jeweils mit dem Auftreten der Krankheit korrelieren“, erklärt Borgwardt. Bei Manhattan Plots gibt der Ausschlag auf der Y-Achse die Korrelationsstärke der SNPs wieder, die auf der X-Achse entsprechend ihrer Position im Chromosom aufgereiht sind. Da die entstehenden Muster an die Skyline von Manhattan erinnern, wurde das Ganze „Manhattan Plot“ getauft.

Mit neuen Algorithmen große Netzwerke auf kleinen Rechnern analysieren



Manhattan Plots wie dieser sind ein Beispiel für die Auswertung genomweiter Assoziationsstudien (GWAS). Der Plot zeigt, wie häufig bei erkrankten Menschen an bestimmten Positionen im Chromosom genetische Veränderungen (SNPs) auftreten. © Prof. Dr. Karsten Borgwardt

Die Werkzeuge für die Suche nach solchen statistischen Zusammenhängen entwickelt und optimiert die Gruppe um Borgwardt selbst. So hilfreich die Ergebnisse sind, darf jedoch nicht zu viel daraus abgelesen werden, wie Borgwardt betont: „Korrelation ist nicht gleichzusetzen mit Kausalität. Die Manhattan Plots dienen uns dazu, Positionen im Genom zu finden, die man sich im nächsten Schritt näher anschauen sollte, um dort liegende Gene und Veränderungen im Genom zu finden." Erst im weiteren Verlauf der Untersuchungen können daraus Kausalitäten abgeleitet werden.

Um mehr Licht in das komplexe Dunkel der Zusammenhänge zu bringen, will Borgwardt neue Algorithmen entwickeln, um nicht nur die Korrelation von einem, sondern von einem Paar SNPs mit einem bestimmten Phänotyp zu finden. Dieser Phänotyp kann grundsätzlich alles mögliche sein, von allen denkbaren körperlichen Merkmalen bis hin zum Auftreten von Unverträglichkeiten oder einer Krankheit. Borgwardts Arbeit ist also nicht nur für die Medizin relevant und auch nicht nur für den Menschen, wie er sagt: „Die Pflanzenforschung kann davon genauso profitieren wie die Veterinärmedizin."

Wie schnell die Arbeiten zum Erfolg führen können, hängt nicht zuletzt von den eingesetzten Rechenkapazitäten ab. „Schon wenn wir uns Korrelationen von SNP-Paaren mit Phänotypen anschauen, haben wir es mit bis zu 10^{14} SNPs zu tun. Mit solchen Analysen kann man selbst große Rechencluster tagelang lahmlegen. Theoretisch sind wir heute jedoch schon in der Lage, hocheffizient ganze Netzwerke von SNPs, interagierenden Genen und Proteinen zu identifizieren, die mit Phänotypen korrelieren. Deshalb entwickeln wir zurzeit Algorithmen, um eine intelligente Suche zu ermöglichen und dadurch mit geringer Rechenkapazität auszukommen", sagt Borgwardt. Statt Supercomputer damit zu beschäftigen, will er die Berechnungen in Zukunft auf einem Laptop durchführen können.

Fachbeitrag

09.09.2013
leh
BioRegio STERN

Weitere Informationen

Max-Planck-Institut für Intelligente Systeme
Prof. Dr. Karsten Borgwardt
Spemannstr. 38
72076 Tübingen
Tel.: 07071 / 601-1784
E-Mail: karsten.borgwardt(at)tuebingen.mpg.de

- Max-Planck-Institut für Intelligente Systeme

Der Fachbeitrag ist Teil folgender Dossiers

Data-Mining: Neue Chancen für Medizin und Gesundheit