technische universität
dortmund

Fakultät für
Informatik

Lehrstuhl für
Künstliche Intelligenz
Prof. Dr. Katharina Morik

Big Data und Industrie 4.0

Big Data

Große Datenmengen sind seit einiger Zeit in aller Munde. Google Trend verzeichnet seit 2011 einen steilen Aufwärtstrend für den Begriff „Big Data“.

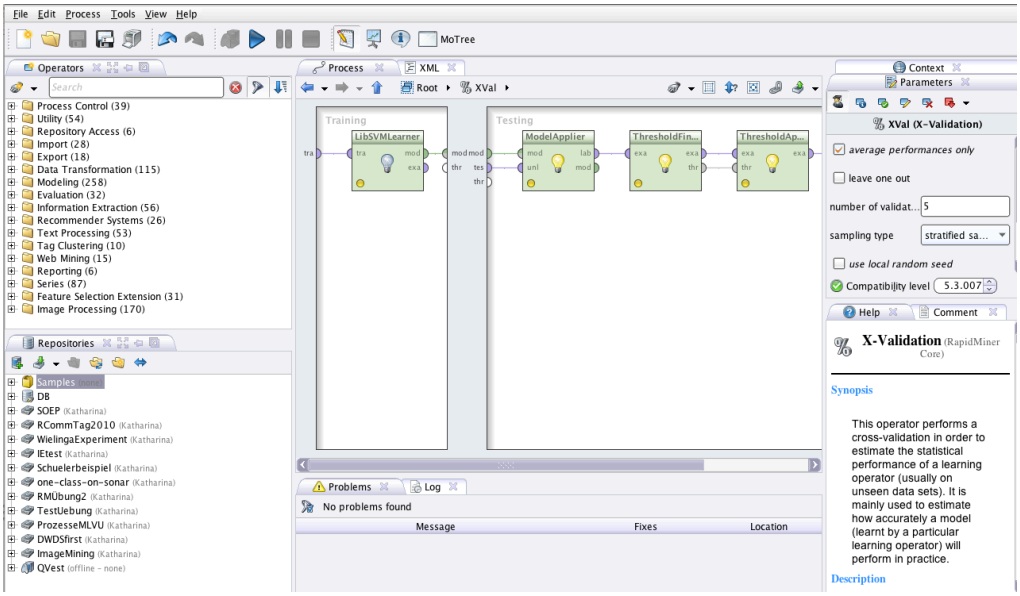
Suchbegriff „Big Data“ weltweit, Anfrage Google Trends vom 21.6.2014

Für Deutschland sagt die Bitkom Experton Group einen steigenden Umsatz im Geschäft mit Big Data (Hard-, Software, Dienstleistungen) auf € 1,7 Milliarden Euro im Jahre 2016 voraus.
Der Wert der Daten kommt erst durch die Analyse zum Vorschein. Interessanterweise korreliert in der Google Suche „Big Data“ genau mit „Big Data Analytics“. Ohne Analyse können große Datensammlungen zu Daten-friedhöfen verkommen. Die Datenfülle und möglicherweise hohe Dimension der Daten erleichtert die Analyse, wenn große Datenräume endlich dichter besetzt sind. Dies ist insbesondere bei Sprachdaten der Fall. Das Volumen ist eine Herausforderung, wenn in der Datenfülle nach nützlicher Information gesucht werden muss wie nach der Nadel im Heuhaufen.

Datenanalyse

Das maschinelle Lernen als automatischer Erwerb von Regeln aus Daten hat sich etwa **1984** als Teilgebiet der Künstlichen Intelligenz etabliert. Die Lernverfahren wurden dann, etwa **1994**, mit relationalen Datenbanken direkt verbunden[i]. Erste Werkzeuge boten direkten Zugang zu SQL-Datenbanken und Sammlungen von Lernverfahren. Der Prozess der Datenanalyse wurde schlicht als „Pipeline“ betrachtet. Die Nachteile:

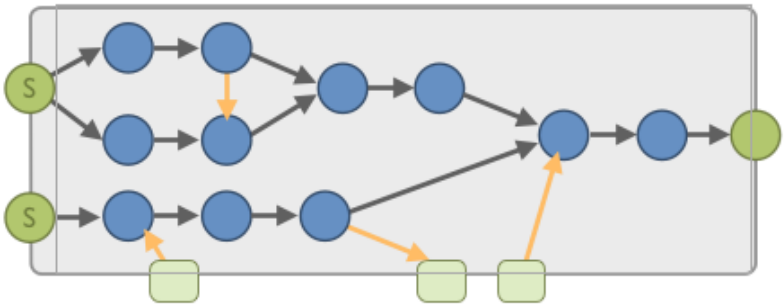
Subprozesse können nicht eingeschachtelt werden, die Kreuzvalidierung kann nicht als eingebetteter Prozess eingebunden werden, es gibt keine automatische Parameteroptimierung.
Diese Nachteile wurden mit dem Werkzeug *Yet Another Learning Environment* **2001** überwunden, das am LS 8 von Ingo Mierswa und anderen Doktoranden entwickelt wurde.



Inzwischen heißt das Werkzeug *RapidMiner* und wird von der gleichnamigen Firma vertrieben und weiter entwickelt. Es erfordert keine Programmierung: man zieht einfach Operatoren in den Prozess. Es gibt allein 115 Operatoren zur Transformation der Rohdaten und 258 Lernverfahren[ii]. Auch **2014** wurde RapidMiner wieder zum beliebtesten Werkzeug gekürt: 3285 Entwickler stimmten bei dem Fachportal *KDnuggets* ab: 44,2% wenden RapidMiner, 38,5% wenden R an. Dabei wenden 35,1% ausschließlich RapidMiner an, während nur 2,1% ausschließlich R anwenden[iii]. *Radoop* ist ein speziell auf Big Data zugeschnittenes Modul von RapidMiner.

Sonderforschungsbereich SFB 876

Weiterhin werden in der Forschung neue Lernverfahren entwickelt. Gerade bezogen auf Big Data und Cyber-Physical Systems gibt es eine Reihe von Herausforderungen, denen sich der SFB 876 stellt: es geht darum, durch Analyse unter Ressourcenbeschränkungen Information zu gewinnen, Stichwort: große Daten – kleine Geräte. 50 Promovenden forschen unter der Leitung von 19 Professorinnen und Professoren in 12 Projekten von 6 Disziplinen[iv]. Der SFB 876 ist in der Fakultät für Informatik angesiedelt, Sprecherin ist Prof Dr. Katharina Morik. Ein methodischer Schwerpunkt sind Datenstromalgorithmen: eingehende Datenströme sollen in Realzeit analysiert werden. Der Doktorand Christian Bockermann hat dazu ein Werkzeug entwickelt, *streams*, mit dem solche online Prozesse leicht konfiguriert, parallelisiert und verteilt ausgeführt werden können[v]. Aus den theoretischen Grundlagen, die im SFB 876 in zwei Projekten von Frau Morik, A1 und B3[vi], erarbeitet wurden, ergeben sich Anwendungsmöglichkeiten.



Graph-Darstellung von *streams* für Datenströme: jeder Knoten rechnet über Elementen des Datenstroms

Industrie 4.0

Zum ersten Mal wurde der Begriff „Industrie 4.0“ auf der Hannover-Messe 2011 verwendet. Der gleichnamige Arbeitskreis unter der Leitung von Dais, Bosch, Kagermann, acatech (Deutsche Akademie der Technikwissenschaften) legte 2013 seinen Bericht vor. Die industrielle Revolution wird danach in vier Etappen aufgeteilt:

1. Mechanisierung: Wasser und Dampfkraft
2. Massenproduktion: Elektrizität
3. Automatisierung: Digitale Information
4. Smart Factory: Integration adaptiver Cyber-Physical Systems in die Produktion

An der leichteren Kommunikation der Bediener mit den Maschinen und verschiedener Maschinen untereinander wird derzeit gearbeitet. Einige gehen aber schon weiter: Wichtig ist, dass sich Abläufe automatisch adaptieren können [vii]. Dies bieten lernfähige Systeme, die anhand der Daten des Produktionsprozesses eine Feinjustierung der Produktion vornehmen können. Genau dies ist die Innovation, die in Zusammenarbeit mit SMS SIEMAG und der Dillinger Hütte entwickelt wurde.

Innovation Realzeitliche Prognose im Stahlwerk

Aus den Datenströmen des BOF-Konverters (*Basic Oxygen Furnace*) sowie statischen Daten aus dem Einsatzmodelle werden Modelle gelernt, die realzeitlich zu jeder Sekunde die Werte von vier Zielgrößen vorhersagen: Temperatur, Kohlenstoffgehalt, Phosphorgehalt und Eisengehalt der Schlacke. Beispiel:

> Welchen Phosphorgehalt hätte die Schmelze, wenn jetzt kein Sauerstoff mehr geblasen würde?

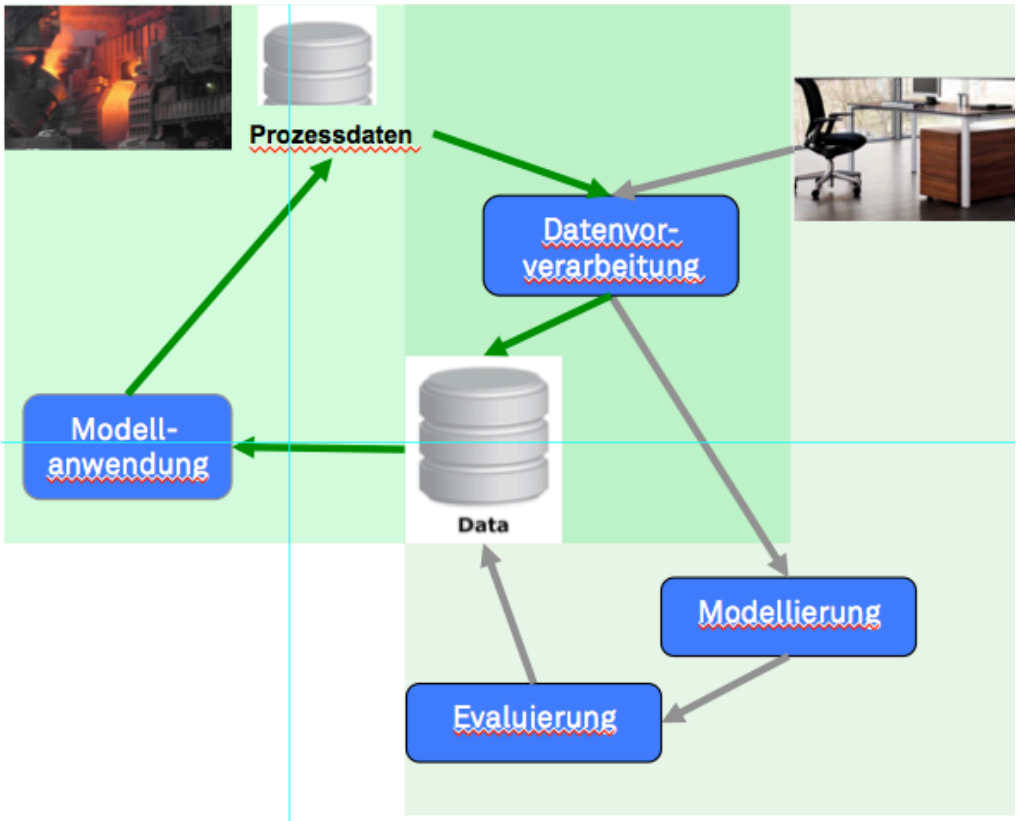
Diese Zielgrößen können nicht direkt beobachtet und müssen deshalb aus den gegebenen Merkmalen abgeleitet werden. Für diese Vorhersagen werden

- Merkmale aus den einströmenden Daten extrahiert, z.B. der Leistungsverlust der Lanze;
- Modelle mit Hilfe der Stützvektormethode (SVM) offline gelernt;
- die Modelle online in Echtzeit im Datenstrom angewandt.

Wenn für alle Zielgrößen ein Wert im gewünschten Bereich vorhergesagt wird, ist das Einblasen von Sauerstoff zu beenden. Im Gegensatz zu dem üblichen Vorgehen, ist hier die Prognose direkt in den Konverterprozess eingebunden und arbeitet im Datenstrom. Mit diesem innovativen Vorgehen können die Entscheidungen der Schmelzer unterstützt werden.
Erstmals ist nun die Lücke zwischen der Datennahme und der Modellanwendung geschlossen und das Lernergebnis, die gelernten Modelle, direkt in den Prozess integriert.

Innovation Stahlwerk im Büro

Die Datennahme, Datenaufbereitung und Datenhaltung ist mit der Datenstromumgebung so modelliert worden, dass auch die gelernten Modelle, ihre Vorhersagen und der tatsächliche Endzeitpunkt des Konverterprozesses gespeichert sind.



Das Stahlwerk im Büro: links das Stahlwerk, in dem die Prozessdaten vorverarbeitet und gelernte Modelle angewandt werden, rechts das Büro, in dem

- neue Modellierungen (d.h. Lernprozesse) entworfen,
- anhand aufgezeichneter Daten Modelle evaluiert und
- „was wäre wenn“ Fragen durchgespielt werden.

Erstmals sind der realzeitliche und der offline Zyklus durch eine innovative Datenhaltung integriert.

[i] Die Tagungsreihe Knowledge Discovery in Databases (KDD) begann als amerikanische Konferenz der Association of Computing Machinery. Die Tagungsreihe International Conference on Data Mining (IEEE ICDM) begann 2001 als internationale Tagung des Institute of Electrical and Electronics Engineers (IEEE).
[ii] Mehr Information und eine freie Testversion ist hier zu finden: http://rapidminer.com

[iii] http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html

[iv] Die Disziplinen sind: Informatik, Statistik, Elektrotechnik, Biomedizin, Physik, geplant ist die Einbeziehung der Logistik.
[v] http://www-ai.cs.uni-dortmund.de/SOFTWARE/streams/
Das Framework hat seine Skalierbarkeit bei einem Wettbewerb 2013 bewiesen, als es, mit einer Ereigniserkennung gekoppelt, ein Ereignis in durchschnittlich nur 2 Millisekunden verarbeiten konnte. http://jwall.org/TechniBall/award.html

[vi] http://sfb876.tu-dortmund.de/index.html

A1 ist das Teilprojekt, das mit Hilfe maschinellen Lernens den Ressourcenverbrauch von eingebetteten Systemen (Cyber-Physical Systems) reduziert. Die Projektleiter sind Katharina Morik und Olaf Spinczyk.
B3 ist das Teilprojekt, das am Beispiel eines Walzwerks der Deutschen Edelstahlwerke, durch frühzeitige Qualitätsprognose während des Walzprozesses erlaubt, Material auszuschleusen oder die Produktionsplanung umzustellen. Projektleiter sind Katharina Morik und Jochen Deuse.
[vii] VDI Zukunftskongress Düsseldorf 11.1.2013: Moderiert von Professor Birgit Vogel-Heuser der TU München diskutieren Entscheider aus Unternehmen wie ABB, Daimler, Festo, Phoenix Contact, Siemens über notwendige Ressourceneffizienz und Flexibilität sowie über die Weiterentwicklung der Automation in Bezug auf Sicherheit, Vernetzung und Echtzeitfähigkeit. Prof. Michael ten Hompel vom Fraunhofer Institut Materialfluss und Logistik in Dortmund und Jörg Murawksi von Würth Elektronik sprechen über erfolgreiche Flexibilisierung durch adaptive Logistik.