Kritik: *An Introduction to Kernel-Based Learning Algorithms* von Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda und Bernhard Schölkopf, IEEE TRANSACTIONS ON NEURAL NETWORKS 2001

Felix Gessert, 01.07.2012, Seminar maschinelles Lernen (Uni Hamburg)

Autoren

Der vorliegende Artikel ist ein Survey zu dem Thema kernel-basiertes Machine Learning, verfasst von den fünf Autoren Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda und Bernhard Schölkopf. Ihr Hintergrund und Forschungsschwerpunkt, sowie eine Einschätzung ihrer thematischen Kompetenz wird im folgenden Abschnitt diskutiert.

Robert Klaus Müller

Person	Professor for Machine Learning an der TU Berlin, Diplom in mathematischer Physik, Doktor in theoretischer Informatik
Forschungs-Schwerpunkt	Statistical Learning Theory für neuronale Netze, Support Vector Maschinen und Ensemble Methoden
Publikationen	Insgesamt ca. 290, darunter viele Journal-Artikel sowie Konferenzpapiere und Buchkapitel
Zitationen	22679 (nach Google Scholar)
Awards	SEL Alcatel Research Award 2006, Olympus Award for Pattern Recognition 1999 , GMD best project award 1998 und 1996

Müller hat wichtige Forschungsbeiträge auf dem Gebiet der Signalverarbeitung von Zeitdaten und statistische Machine Learning Methoden (z.B. Kernel PCA) geleistet. Sein derzeitiger Forschungsschwerpunkt liegt in der Analyse biologischer und medizinischer Daten (z.B. sequenzierter Genome). Müller hat sich in vielen Konferenzen und Workshops engagiert, die er z.T. selbst ausgerichtet hat. Er kann als ein Forschungsschwergewicht auf dem Gebiet dieses Artikels eingeschätzt werden.

Sebastian Mika

Person	Geschätsführer der Idalab, Informatik und Mathematik Studium an der TU Berlin, Doktor über statistische Methoden
Forschungs-Schwerpunkt	Statistische Methoden, Kernel-Methoden, Support Vector Maschinen
Publikationen	ca. 44 (nach Microsoft Academic Search), Journal- und Konferenzpapiere
Zitationen	3851 (nach Microsoft Academic Search)
Awards	Keine ausführliches CV oder umfassende Publikationsliste im Internet

Aus der Doktorarbeit von Sebastian Mika entwickelte sich die Idalab GmbH, ein Dienstleister für statistische Anwendungen speziell für Datenanalyse und Individualsoftware zur statistischen Auswertung. Als Geschäftsführer der Bereiche Forschung und Finanzen ist sein Tätigkeitsschwerpunkt nicht primär akademisch sondern anwendungsbezogen. Seine vergangene Mitautorenschaft in vielzitieren Papieren über Kernel-Methoden und sein Erfolg bei der kommerziellen Umsetzung statistischer Methoden kann als deutliches Indiz für seine Kompetenz auf dem Gebiet gewertet werden.

Gunnar Rätsch

Person	Leiter der *Machine Learning in Genome Biology* Gruppe am Friedrich Miescher Laboratory in Tübingen, Doktor über Boosting und Optimierung
Forschungs-Schwerpunkt	Kernel-Methoden, Boosting, DNA/RNA Analyse
Publikationen	ca. 189 (nach Google Scholar), Journal- und Konferenzpapiere, noch immer sehr aktiv
Zitationen	11725 (nach Google Scholar)
Awards	Bedeutende, z.B. Olympus prize 2007 for pattern recognition, Michelson award for an outstanding doctoral dissertation

Wie die anderen Autoren, ist auch Gunnar Rätsch eine wichtige Figur im Feld des kernel-basierten Machine Learning. Seine vergangenen Arbeiten (d.h. zum Zeitpunkt der Entstehung dieses Artikels) beschäftigen sich vorwiegend mit Boosting und Support Vektor Maschinen. Derzeit richtet Rätsch seine Forschung jedoch verstärkt bezüglich Methoden der Genom-Analyse aus. Nach eigener Aussage interessiert ihn in erster Linie die praktische Perspektive des Machine Learning, so dass es vermutlich seinem Hinwirken zu verdanken ist, dass der vorliegende Artikel die vorgestellten Verfahren an praktischen Beispielen wie OCR und DNA Analyse verdeutlicht.

Koji Tsuda

Person	Senior Researcher am AIST Computational Biology Research Center, Doktor (Engineering) an der Kyoto University 1998
Forschungs-Schwerpunkt	Graphentheorie, Kernel-Methoden, bioinformatische Anwendung von Machine Learning
Publikationen	94 Publikationen in Journalen und auf Konferenzen zu Machine Learning, Bioinformatik und Data Mining
Zitationen	4603 (nach Google Scholar)
Awards	Best Paper Award beim *International Workshop on Mining and Learning with Graphs* und der IPSJ Nagao Award 2009

Koji Tsuda forscht erfolgreich an einer industriellen Forschungseinrichtung, war aber bereits bei verschiedenen anderen internationalen Einrichtungen tätig, u.a. dem Max Planck Institut für biologische Kybernetik in Tübingen. Tsuda war Area Chair und Program Committee Mitglied bei Konferenzen wie NIPS und ICML. Unter seinen Arbeiten ist der vorliegende

Artikel der bei weitem meistzitierte, seine übrigen Veröffentlichungen scheinen jedoch zumindest von der jeweiligen Community interessiert aufgenommen zu werden.

Bernhard Schölkopf

Person	Direktor des MPI im Bereich Intelligente Systems, MS Mathematik, Diplom Physik, Forschung an zahlreichen Universitäten und Einrichtungen
Forschungs-Schwerpunkt	Inferenz aus empirischen Daten, Kernel -Methoden für hochdimensionale Daten
Publikationen	Sehr viele, 8 Bücher, 97 Journal Artikel, 181 Konferenz Paper
Zitationen	44592 insgesamt, 29790 davon seit 2007 (nach Google Scholar)
Awards	Sehr viele, zuletzt der Max-Planck Forschungspreis 2011, der mit 750.000 Euro dotiert ist

Professor Bernhard Schölkopf ist ohne Frage eine der führenden Figuren im Bereich Machine Learning. Als Direktor des MPI Bereichs „Intelligent Systems“, ehemaliger Program Chair namhafter Konferenzen (z.B. NIPS, COLT) und Editor (z.B. JMLR), sowie Forschungserfahrung bei mehreren renommierten Wissenschaftseinrichtungen (z.B. AT&T Bell Labs, Microsoft Research) hat Schölkopf als Begründer der kernel-basierten Methoden Weltrang. Mit einem h-Index 79 zählt er zu den bedeutendsten aktiven Forschern der Informatik [1]. Viele der in diesem Papier erläuterten Methoden gehen direkt auf ihn zurück.

Der Inhalt des Artikels

Der Artikel gibt einen umfassenden Überblick über den Forschungsstand der kernel-basierten Methoden im Machine Learning. Der Artikel beginnt damit, das Problem der Klassifikation mithilfe der Vapnik-Chevronenkis (VC) Theorie einzuführen. Das Grundproblem eines Klassifikators besteht darin, einen guten Kompromiss zwischen *Underfitting* (fehlende Komplexität) und *Overfitting* (zu hohe Komplexität) zu finden. Die VC-Dimension liefert ein geeignetes Maß, um die Komplexität einer Funktionenklasse anhand der Anzahl erreichbarer Klassifikationen von 2^n möglichen Klassifikationen bei n gegebenen Trainingsdaten anzugeben. Das Ziel ist es, die Trainingsdaten möglichst gut zu klassifizieren, d.h. einen kleinen empirischen Fehler zu erzielen und dabei gleichzeitig eine Funktion mit niedriger VC-Dimension einzusetzen, um Overfitting und damit fehlende Generalisierung zu vermeiden. Die VC-Dimension exakt zu bestimmen, ist im Allgemeinen schwer. Für Hyperebenen $(\boldsymbol{w} \cdot \boldsymbol{x}) + b$ in kanonischer Form jedoch ist die VC-Dimension eine Funktion des Margins $\boldsymbol{w}$.

Da Hyperebenen im Eingaberaum der Daten nur linear separierbare Daten trennen können, müssen zur Klassifikation von Daten durch trennende Hyperebenen die Eingabedaten in einen höherdimensionalen Raum transformiert werden: $\boldsymbol{x}_{new} = \Phi(\boldsymbol{x}_{input})$ (siehe Abb. 1). Um dem Problem der hohen Berechnungskomplexität in hochdimensionalen Räumen zu entgehen, kommt die zentrale Grundlage der Kernel Methoden zum Einsatz, der sogenannte **Kerneltrick**. Anstatt Berechnungen im hochdimensionalen Raum durchzuführen wird stattdessen das Skalarprodukt zweier transformierter Daten durch eine leicht zu berechnende Kernel-Funktion K ermittelt: $K(x_1, x_2) = \Phi(x_1) \cdot \Phi(x_2)$. Der höherdimensionale Raum kann dadurch implizit berechnet werden, ohne die Daten tatsächlich in den höherdimensionalen

Raum zu transformieren oder die Abbildung Φ auch nur zu kennen. Dadurch kann jeder Algorithmus, der nur auf dem Skalarprodukt der Eingabedaten arbeitet, durch den Kerneltrick im höherdimensionalen Raum ausgeführt werden.

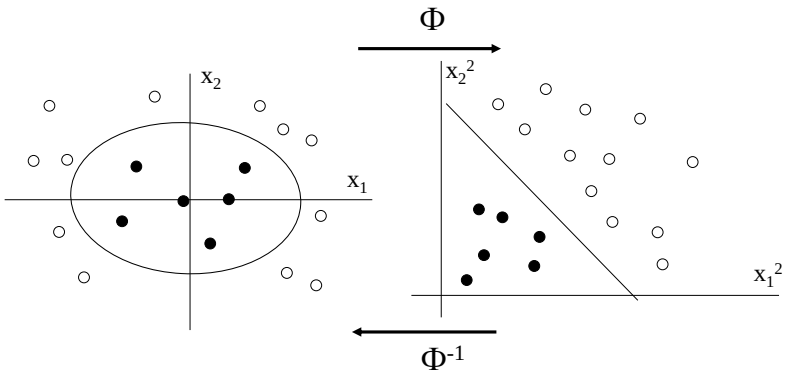

Abbildung 1 Durch eine Transformation Φ werden die Daten separierbar. Aus [2].

Ausgehend von dieser zentralen Erkenntnis beschreiben die Autoren Support Vector Machines (SVMs), einen binären Klassifikator, der Eingabedaten in einem höherdimensionalen Raum durch eine Hyperebene so separiert, dass der minimale Abstand der Ebene zu den Trainingsdatenpunkten (der Margin) maximal wird. Die Kernel Fisher Diskriminante, eine Projektion mit der Eigenschaft die Klassenzugehörigkeit der projizierten Daten möglichst gut zu erhalten, macht sich ebenfalls den Kerneltrick zunutze. Anschließend zeigen die Autoren, dass ein formaler Zusammenhang zwischen Boosting und SVMs besteht, beim Boosting die Berechnungen jedoch explizit im hochdimensionalen Raum ausgeführt werden. Darauffolgend wir die Anwendung des Kerneltricks auf Unsupervised Learning anhand der Prinicipal Component Analysis (PCA) und der Single Class Classification gezeigt. Anders als die klassische PCA, ist die Kernel-PCA in der Lage, nichtlineare niederdimensionale Richtungen hoher Varianz in den Daten zu identifizieren. Nach einer Diskussion der Problematik eine adäquate Kernelfunktion zu finden (z.B. durch Kreuzvalidierung), demonstrieren die Autoren an Beispieldaten die hohe Klassifikationsqualität der SVM für die Erkennung codierender Genomsequenzen und Buchstabenerkennung sowie die Anwendung der Kernel-PCA zum Entrauschen korrumpierter Bilddaten.

Einordnung nach Qualität, Klarheit, Originalität und Signifikanz

Qualität

Der Artikel erfüllt einen sehr hohen wissenschaftlichen Qualitätsstandard. Die mathematische Darstellung ist meist prägnant, ohne durch unnötige Komplexität überfordernd zu wirken und die verwendete Notation wird sauber eingeführt. Auf Beweise oder Herleitungen verzichtet der Artikel naheliegender Weise. Die Autoren referenzieren eine große Anzahl relevanter Literatur und bieten so mit ihrem Survey einen geeigneten Einstiegpunkt, um in jedes der angesprochenen Themengebiete tiefer einzusteigen. Etwas erstaunlich ist lediglich, dass der Artikel noch mehrere kleine Flüchtigkeitsfehler enthält. Ein negativer Punkt bei der Qualität des Papers ist die Verwendung von Rastergrafiken mit niedriger Auflösung, die einerseits schlecht lesbar sind und andererseits sehr unprofessionell wirken.

Klarheit

Der Artikel bemüht sich durchgehend darum die mathematischen Darstellungen auch in Worte zu kleiden und am Ende jedes Abschnitts eine kurze Zusammenfassung der relevanten Erkenntnisse zu geben. An einigen Stellen setzen die Autoren profundes Vorwissen über höhere Mathematik voraus, was sich allerdings angesichts des Themas schwer vermeiden lässt. Die Gliederung in Hintergrund, Supervised Learning, Unsupervised Learning und Evaluation ist sinnvoll und lässt den Artikel strukturiert wirken.

Originalität

Anstatt neue Verfahren zu präsentieren, erörtern die Autoren bereits bekannte Algorithmen und Techniken. Diese Arbeit ist in diesem Sinne nicht originär und auch nicht der einzige Survey-Artikel zu dem Thema, aber dies ist auch nicht seine Funktion. Nach Aussage der Autoren soll der Artikel einen breiteren Überblick über die Kernelmethoden geben als dies andere Artikel bis zu dem Zeitpunkt taten. Die Tatsache, dass der Artikel bisher 2094 mal zitiert wurde, spricht klar für die Tatsache, dass ein Bedarf an der thematischen Zusammenstellung dieses Artikels bestand.

Signifikanz

Die Signifikanz eines Survey-Artikels lässt sich wohl am ehesten daran messen, wie viele Wissenschaftler dazu inspiriert werden, an dem vorgestellten Thema oder einzelnen Aspekten zu arbeiten und daran ob es gelang, eine saubere und kohärente Taxonomie für nachfolgende Arbeiten aufzustellen. In diesem Sinne ist der Artikel signifikant, wird er doch durch zahlreiche andere Papiere referenziert. Die Signifikanz des Artikels hätte möglicherweise dadurch erhöht werden können, dass die einzelnen Verfahren auf eine Weise dargestellt werden, dass auch Anwender dieser Algorithmen (die nicht an zu mathematischen Details interessiert sind) auf diesen Artikel zurückgreifen können.

Zusammenfassung

Es werden nun abschließend jeweils drei Stärken und Schwächen des Papers genannt.

Stärken des Papers

- Prägnante Sprache und mathematische Notation.
- Eine große Breite an Verfahren wird abgedeckt und die Kernideen jeweils erklärt.
- Der praktische Teil illustriert eindrucksvoll die tatsächlichen Vorteile von kernel-basierten Methoden.

Schwächen des Papers

- Die Rastergrafiken sind alle anderen Arbeiten entnommen und sehr schlecht lesbar.
- Der mathematische Anspruch des Artikels ist zu hoch um Anwendern als Überblick zu dienen.
- Eine deutliche Abgrenzung und differenzierte Betrachtung von Vor- und Nachteilen im Vergleich zu anderen Algorithmen fehlt.

Referenzen

[1] Liste der Informatiker mit höchstem h-Index, Stand Juni 2012: http://www.cs.ucla.edu/~palsberg/h-number.html

[2] http://www.meduniwien.ac.at/user/georg.dorffner/lv/mlnc.html