Maschinelles Lernen mit multiplen Kernen

Marius Kloft

Abteilung Maschinelles Lernen, Technische Universität Berlin
kloft@tu-berlin.de

Abstract: Diese Arbeit gibt zunächst eine grundlegende Einführung in Theorie und Praxis des Maschinellen Lernens mit multiplen Kernen und skizziert den Stand der Forschung. Weiter entwickelt die Arbeit eine neue Methodologie des Lernens mit mehreren Kernen und beweist deren Effizienz und Effektivität. Sie entwickelt Algorithmen zur Optimierung des assoziierten mathematischen Programmes, die im Vergleich zu vorherigen Ansätzen um bis zu zwei Größenordnungen schneller sind. Unsere theoretische Analyse des Generalisierungsfehlers zeigt dabei Konvergenzraten mit Ordnungen von maximal $O(M/n)$, frühere Analysen präzisierend, die bisher nur $O(\sqrt{M/n})$ erreichten. In Anwendungen auf zentrale Fragestellungen der Bioinformatik und des Maschinellen Sehens werden Vorhersagegenauigkeiten erreicht, die den bisherigen Stand der Forschung signifikant übertreffen, wodurch eine Grundlage zur Erschließung neuer Anwendungsfelder und Forschungsansätze geschaffen wird.

1 Einführung

Ziel des Maschinellen Lernens ist das Erlernen des unbekannten Zusammenhanges zweier Variablen X und Y aus Daten $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$, um bei Beobachtung eines neuen Musters x eine möglichst präzise Vorhersage für dessen unbekanntes Konzept y abgeben zu können. Einen besonders eleganten Lösungsansatz hierfür bieten nicht-lineare, „kernbasierte“ Lernverfahren [MMR$^+$01]: Mit der Substitution aller Skalarprodukte $\langle \phi(x), \phi(\widetilde{x}) \rangle$ durch eine Nicht-Linearität $k(x, \widetilde{x})$ – dem sogenannten *Kern* – werden die Muster *implizit* in einen hoch-dimensionalen Merkmalsraum eingebettet, in welchem sie linear getrennt werden können. So erzeugen wir auf systematische Art und Weise aus einfacheren Lernmaschinen sehr viel komplexere und leistungsstärkere – im Merkmalsraum lineare – was den Lernschritt von der Datenrepräsentation modular entkoppelt. Aufgrund ihrer Ausdruckskraft und Leistungsstärke – bei gleichzeitig sehr geringer Lauf- und Ausführungszeit – stellen kernbasierte Lernverfahren in Anwendungsbereichen mit komplexen Problemstellungen und großen Datenmengen, wie beispielsweise der Bioinformatik und dem Maschinellen Sehen, den gegenwärtigen Standard dar.

Klassische kernbasierte Lernverfahren verwenden einen *einzelnen* Kern, der in der Regel aus einer im Vorfeld zu spezifizierenden Menge von Kandidatenkernen $\mathcal{K} = \{k_1, \ldots, k_M\}$ durch Kreuzvalidierung ausgewählt wird. Problematischerweise können die Kerne jedoch *komplementäre* Eigenschaften des Lernproblems charakterisieren: Zum Beispiel treten im Maschinellen Sehen eine Vielzahl von gegensätzlichen

Informationselementen auf, basierend auf der Farbverteilung eines Bildes oder den auftretenden Formen und Kanten et cetera. Nur einen einzelnen Kern, z. B. den „Farbkern“, auszuwählen, bedeutet daher zugleich auch wertvolle, komplementäre Information zu verwerfen! Beispielsweise mag Farbinformation hilfreich sein zur Erkennung von Stoppschildern, aber weniger hilfreich zur Annotation von Bildern, die Autos oder Luftballons enthalten.

Alle in dieser Dissertation entwickelten Methoden basieren daher auf einer optimierten Gewichtung mehrerer Kerne, *um die darin enthaltende Information zu fusionieren*:

$$k = \theta_1 k_1 + \cdots + \theta_M k_M \,.$$

Bis auf sehr kleine M (typischerweise $M \leq 3$) ist der Suchraum der θ_i allerdings zu groß für gewöhnliche Suchverfahren. Eine grundlegende Einsicht an dieser Stelle ist, dass viele Methoden des Maschinellen Lernens durch mathematische Programme definiert sind, wie beispielsweise die Support-Vektor-Maschine:

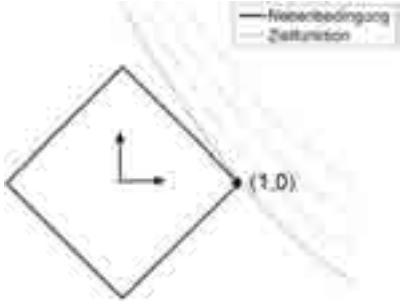
$$\begin{aligned} &\mathrm{SVM}(k, \mathcal{D}) \\ &:= \min_{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{n} \max\Big(0, 1 - y_i\big(\langle \boldsymbol{w}, \phi(x_i)\rangle - b\big)\Big) \qquad (1) \\ &= \max_{\boldsymbol{\alpha}\in\mathbb{R}^n : \mathbf{0}\leq\boldsymbol{\alpha}\leq C, \boldsymbol{y}^\top\boldsymbol{\alpha}=0} \quad \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i, x_j)\,. \end{aligned}$$

Können wir die Parameter $\theta_1, \ldots \theta_M$ ebenfalls als Variablen in ein solches Optimierungsproblem mit aufnehmen?

In der klassischen Arbeit [LCG+04] wird der Suchraum auf positive Gewichte $\theta_m \geq 0$ eingeschränkt, die sich zu eins aufsummieren,

$$\sum_{m=1}^{M} \theta_m = 1\,, \qquad (2)$$

weil dies sonst zu trivialen Lösungen $\theta_1 = \cdots = \theta_M = \infty$ führen würde. Die Nebenbedingung (2) führt zu *dünn-besetzten* (oder „spärlichen“) Gewichten $\boldsymbol{\theta}$, wie man anhand der nebenstehenden Abbildung erkennen kann: In dem Optimum des mathematischen Programms (1) berührt eine der Höhenlinien der Zielfunktion die Nebenbedingung (2) in einer ihrer (dünn besetzten) Ecken.



Dünn besetzte Kerngewichte sind zwar leicht interpretierbar und können auch numerisch von Vorteil sein, jedoch erlauben wir uns, an dieser Stelle zu betonen, dass die Fokussierung auf dünn besetzte Kernmischungen zum Teil eher einer generellen Präferenz der gegenwärtigen Forschung für sogenannte *sparse* Methoden geschuldet ist. Dabei kann die Beschränkung auf dünn-besetzte Gewichte bei der *Fusion* von multiplen Kernen äußerst unlogisch und geradezu kontraintuitiv sein, insbesondere wenn die Kerne *komplementäre*

Eigenschaften des Lernproblems codieren. Dies ist beispielsweise in den Bereichen der Bioinformatik und des Maschinellen Sehens der Fall, wo – wie oben erwähnt – Farb-, Form- und Kanteninformationen synergetisch wirken und die bisher übliche Selektion die potenzielle Leistungsfähigkeit massiv einschränkt.

In der vorliegenden Dissertation positionieren wir uns gegen diesen vorherrschenden Trend und zeigen, dass ausgewogene, nicht-spärliche Kernkombinationen weit höhere Vorhersagegenauigkeiten ermöglichen können als sparse Methoden. Weiterhin beweisen wir theoretische Schranken, die weit präzisere Konvergenzraten aufweisen als bisher existierende. So lässt sich nun erklären, *warum* nicht-spärliche Kernmischungen oftmals effektiver sind. In numerischer Hinsicht leiten wir Algorithmen her, die schneller sind als die bisherigen und es erlauben, auch gewaltige Datenmengen, wie sie etwa in der Bioinformatik auftreten können, zu verarbeiten.

Die Hauptbeiträge der Arbeit [Klo11] können danach wie folgt zusammengefasst werden:

- Wir entwickeln eine neue *Methodologie* des Lernens mit multiplen Kernen, die zu nicht-spärlichen Kernkombinationen führt – effizienter und effektiver als vorherige Ansätze.
- Zur Lösung des mit der Methodologie assoziierten mathematischen Programms leiten wir Algorithmen her, die gleichzeitig Zehntausende von Trainingsbeispielen und Tausende von Kernen verarbeiten können, und beweisen deren Konvergenz, welche um bis zu zwei Größenordnungen schneller erfolgt als jene der überkommenen Algorithmen.
- *Theoretische* Analysen des Generalisierungsfehlers zeigen Konvergenzraten einer Ordnung von maximal $O(M/n)$ – was alle früheren Analysen präzisiert, die nur $O(\sqrt{M/n})$ erreichten. Auf Grundlage der theoretischen Schranken können wir erklären, *warum* nicht-spärliche Kerngewichte oftmals effektiver sind.
- In *Anwendungen* auf zentrale Fragestellungen der Bioinformatik und des Maschinellen Sehens werden Vorhersagegenauigkeiten erreicht, die den bisherigen Stand der Forschung weit übertreffen.

Im Folgenden gehen wir auf die Hauptergebnisse der Dissertation [Klo11] ausführlicher ein.

2 Lernen nicht-spärlicher Kernkombinationen

In der vorliegenden Dissertation verwerfen wir die Beschränkung auf dünn-besetzte Kerngewichte und definieren das Lernen mit multiplen Kernen einschränkungslos durch ein rigoroses, mathematisches Optimierungskriterium unter Verwendung völlig beliebiger Normen $\|\cdot\|_O$ und Lossfunktionen l [KRB10]:

$$\begin{aligned} \inf_{\boldsymbol{w},b,\boldsymbol{t}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_{2,O}^2 + C\sum_{i=1}^{n} l\big(t_i,\ y_i\big) \\ \text{s.t.} \quad & \forall i:\ \langle \boldsymbol{w}, \phi(x_i)\rangle + b = t_i\ . \end{aligned} \tag{P}$$

Diese das Feld vereinheitlichende Formulierung enthält alle existierenden Ansätze zum Lernen mit multiplen Kernen als Spezialfälle, die nun *gemeinsam* analysiert werden können. Beispielsweise leiten wir eine völlig allgemeine duale Repräsentation mit Hilfe einer eigens zu dem Zweck in [Klo11] von uns entwickelten Dualitätsmethode her:

$$\sup_{\boldsymbol{\alpha}:\, \mathbf{1}^\top \boldsymbol{\alpha}=0} -C \sum_{i=1}^{n} l^*\Big(-\frac{\alpha_i}{C}, y_i\Big) - \frac{1}{2}\Big\| \sum_{i=1}^{n} \alpha_i \phi(x_i)\Big\|^2_{2,O^*} . \tag{D}$$

Der Einfachheit halber konzentrieren wir uns in der hier angefertigten Zusammenfassung auf Minkowski ℓ_p-Normen $\|\boldsymbol{\theta}\|_p := \big(\sum_{m=1}^{M} |\theta_m|^p\big)^{1/p}$ und die SVM-Verlustfunktion $l(t) := \max(0, 1-t)$, d. h. wir betrachten das folgende mathematische Programm:

$$\min_{\boldsymbol{\theta}\geq \mathbf{0},\, \|\boldsymbol{\theta}\|_p=1} \ \mathrm{SVM}\Big(\sum_{m=1}^{M} \theta_m k_m, \mathcal{D}\Big) \tag{3}$$

3 Algorithmen

Im Rahmen der Dissertation werden drei effiziente Algorithmen zur Optimierung von (3) vorgestellt [KBLS08, KBS+09, KBSZ11]:

- ein Newton-Verfahren
- ein Block-Koordinaten-Abstiegs-Algorithmus
- ein Cutting-Plane-Algorithmus, basierend auf sequentieller, quadratisch-bedingter, quadratischer Programmierung mit Höhenlinien-Projektionen.

Jeder dieser Algorithmen existiert in zwei Varianten: als einfacher, zweischrittiger Algorithmus sowie als effizienter, fest in die SVM integrierter Algorithmus. In dieser Zusammenfassung konzentrieren wir uns darauf, den Block-Koordinaten-Abstiegs-Algorithmus darzustellen. Er basiert auf einer einfachen, analytischen Optimalitäts-Formel, die innerhalb von Mikrosekunden ausgewertet werden kann:

$$\forall\, m = 1, \ldots, M: \qquad \theta_m = \frac{\|\boldsymbol{w}_m\|_2^{2-p}}{\Big(\sum_{m'=1}^{M} \|\boldsymbol{w}_{m'}\|_2^p\Big)^{(2-p)/p}} . \tag{4}$$

In der einfacheren, modularen Version werden alternierend die Gleichungen (1) und (4) gelöst, so dass dieser Wrapper-Algorithmus sogar einfacher als SimpleMKL [RBCG08] ist, welches in jeder Iteration eine heuristische Line Search ausführt. In der zweiten, in Algorithmen-Tafel 1 beschriebenen Version ist das Mehr-Kern-Modul fest in die SVM eingebettet, um maximale Effizienz zu erreichen.

Alle Algorithmen sind in C++ programmiert und in die Shogun Machine Learning Toolbox [SRH+10] integriert worden, welche Schnittstellen zu MATLAB, Octave, Python und R beinhaltet. Die Konvergenz beider Algorithmen wird durch das folgende Theorem, dessen Beweis in Abschnitt 3.2.1 in [Klo11] geführt wird, sichergestellt:

Algorithm 1 In die SVM integrierter analytischer Block-Koordinaten-Algorithmus.

input: $p \in [1,\infty]\backslash\{2\}, \quad Q \in \mathbb{N}, \quad \epsilon > 0$
initialize: $\forall\, i,m: \; g_{m,i} = \hat{g}_i = \alpha_i = 0; \quad L = S = -\infty; \quad \theta_m := (1/M)^{(2-p)/p}$
iterate
  Select l variables $\alpha_{i_1}, \ldots, \alpha_{i_l}$ based on the gradient $\hat{\mathbf{g}}$ of SVM
  Store $\boldsymbol{\alpha}^{\text{OLD}} = \boldsymbol{\alpha}$ and then compute $\boldsymbol{\alpha} := \arg(\text{SVM}(K_{\boldsymbol{\theta}}))$ w.r.t. the selected variables
  Update gradient $\forall\, i,m: \; g_{m,i} := g_{m,i} + \sum_{q=1}^{Q} (\alpha_{i_q} - \alpha_{i_q}^{\text{OLD}}) k_m(x_{i_q}, x_i)$
  Compute the quadratic terms $\forall m: \; S_m := \frac{1}{2}\sum_i g_{m,i}\alpha_i, \;\; \|\boldsymbol{w}_m\|_2^2 := 2\theta_m^2 S_m$
  $L_{\text{OLD}} = L, \quad L = \sum_i y_i \alpha_i, \quad S_{\text{OLD}} = S, \quad S = \sum_m \theta_m S_m$
  if $|1 - (L-S)/(L_{\text{OLD}} - S_{\text{OLD}})| \geq \varepsilon$
    Update $\boldsymbol{\theta}$ according to Eq. (4)
    if $p \in [1,2]$
      For all m compute $\|\boldsymbol{w}_m\|^2 := \theta_m^2 \boldsymbol{\alpha} K_m \boldsymbol{\alpha}$
    end if
  else
    break
  end if
  $\hat{g}_i = \sum_m \theta_m g_{m,i}$ for all $i = 1, \ldots, n$
output: $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ as sparse vectors

Theorem 1. *Seien $K_1, \ldots, K_M$ strikt positiv-definite Kernmatrizen. Dann ist jeder Häufungswert der Sequenz der von Algorithmus 1 zurückgegebenen Lösungen ein globaler, optimaler Punkt des mathematischen Programmes* (3).

Anhand der nebenstehend dargestellten Ergebnisse unserer Laufzeituntersuchungen erkennen wir, dass unsere Algorithmen – erstmals! – die effektive Verwendung von Zehntausenden von Datenpunkten und Tausenden von Kernen erlauben. Wie in Abbildung 1 dargestellt, erweisen sie sich um bis zu *zwei Größenordnungen* schneller als die State-of-the-Art Algorithmen SimpleMKL [RBCG08] und HessianMKL [CR08]. Während letztere bei ca. 10 000 Trainingsbeispielen und 1 000 Kernen „out of memory“ meldeten, kann unser Algorithmus durch on-the-fly-Berechnung von Kernen auch für größere Trainingsmengen eingesetzt werden.

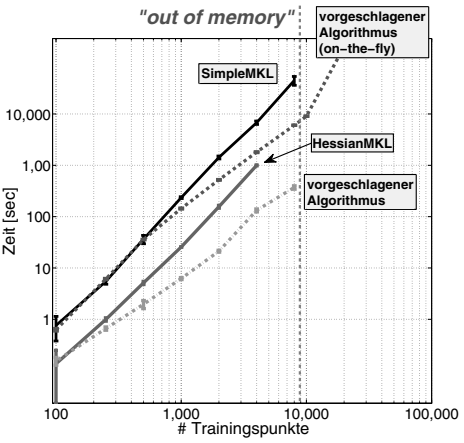


Abbildung 1: Laufzeit von Algorithmus 1 im Vergleich zu State-of-the-Art Verfahren.

4 Theoretische Analyse

Die vorgeschlagene Methodologie ist durch fundamentale Garantien der statistischen Lerntheorie untermauert [KB11, KB12] – wir zeigen die folgende obere Schranke auf die lokale Rademacher-Komplexität [BBM05] des Lernens mit multiplen Kernen:

Theorem 2 (Obere Rademacher-Schranke)**.** *Die lokale Rademacher-Komplexität des Lernens mit multiplen Kernen kann durch die folgende Schranke abgeschätzt*

werden:

$$R_r(H_p) \leq \min_{t\in[p,2]} \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rM^{1-\frac{2}{t^*}}, ceC^2t^{*2}\lambda_j^{(m)}\right)\right)_{m=1}^{M}\right\|_{\frac{t^*}{2}}} + \frac{\sqrt{Be}CM^{\frac{1}{t^*}}t^*}{n},$$

wobei $\lambda_j^{(m)}$ *den j-ten Eigenwert des m-ten Kerns (in absteigender Reihenfolge sortiert),* $t^* := \frac{t}{t-1}$ *den zu t konjugierten Exponenten und* $B^2 := \sup_x k(x,x)$ *bezeichnet.*

Beweisskizze. An dieser Stelle fassen wir die Schlüsselschritte des Beweises von Theorem 2 zusammen. Der vollständige Beweis ist auf den Seiten 51–59 in [Klo11] geführt.

1. Bestimmung der Komplexität der Originalklasse durch die zentrierte Klasse:
$$R_r(H_p) \leq R_r(\widetilde{H}_p) + n^{-\frac{1}{2}} \min\left(\sqrt{r}, C\sqrt{\left\|\left(\operatorname{tr}(J_m)\right)_{m=1}^{M}\right\|_{\frac{p^*}{2}}}\right)$$
2. Abschätzung der Komplexität der zentrierten Klasse:
$$R_r(\widetilde{H}_p) \leq \sqrt{\frac{r\sum_{m=1}^{M}h_m}{n}} + C\,E\left\|\left(\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle\widetilde{\boldsymbol{u}}_j^{(m)}\right)_{m=1}^{M}\right\|_{2,p^*}$$
3. Verwendung der Ungleichungen von Khintchine-Kahane und Rosenthal:
$$E\left\|\left(\sum_{j=h_m+1}^{\infty}\langle\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle\widetilde{\boldsymbol{u}}_j^{(m)}\right)_{m=1}^{M}\right\|_{2,p^*} \leq \sqrt{\frac{p^*}{n}}\,E\left(\sum_{m=1}^{M}\left(\sum_{j=h_m+1}^{\infty}\frac{1}{n}\sum_{i=1}^{n}\langle\widetilde{\phi}_m(x_i), \widetilde{\boldsymbol{u}}_j^{(m)}\rangle^2\right)^{\frac{p^*}{2}}\right)^{\frac{1}{p^*}}$$
4. Abschätzung der Komplexität der Originalklassse:
$$R_r(H_p) \leq \sqrt{\frac{4r\sum_{m=1}^{M}h_m}{n}} + \sqrt{\frac{4ep^{*2}C^2}{n}\left\|\left(\sum_{j=h_m+1}^{\infty}\lambda_j^{(m)}\right)_{m=1}^{M}\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}CM^{\frac{1}{p^*}}p^*}{n}$$
5. Charakterisierung bezüglich der Trunkierung der Spektren der Kerne:
$$R_r(H_p) \leq \sqrt{\frac{16}{n}\left\|\left(\sum_{j=1}^{\infty}\min\left(rM^{1-\frac{2}{p^*}}, ep^{*2}C^2\lambda_j^{(m)}\right)\right)_{m=1}^{M}\right\|_{\frac{p^*}{2}}} + \frac{\sqrt{Be}CM^{\frac{1}{p^*}}p^*}{n} \qquad \square$$

Sind die oberen Schranken präzise oder möglicherweise verbesserbar? Diesbezüglich zeigen wir eine *untere* Schranke, deren Konvergenzrate mit jener der oberen Schranke übereinstimmt. Wir können daher zu dem Schluss kommen, dass die erzielten Raten nicht verbesserbar und theoretisch-optimal sind:

Theorem 3 (Untere Rademacher-Schranke)**.** *Seien die Kerne zentriert und unabhängig, identisch verteilt und* $c > 0$ *eine Konstante mit* $\lambda^{(1)} \geq \frac{1}{nD^2}$*. Dann gilt für alle* $r \geq \frac{1}{n}$ *und* $p \geq 1$*:*

$$R_r(H_p) \;\geq\; \sqrt{\frac{c}{n}\sum_{j=1}^{\infty}\min\left(rM, D^2M^{2/p^*}\lambda_j^{(1)}\right)}. \tag{5}$$

Da die Generalisierungsfähigkeit einer Lernmaschine durch die lokale Rademacher-Komplexität genauestens charakterisiert ist [BBM05], folgt aus Theorem 2 die folgende Generalisierungsschranke für ℓ_p-Norm MKL:

Theorem 4 (Theoretische Garantie)**.** *Angenommen* $\|k\|_\infty \leq B$ *und* $\exists d > 0$, $\alpha := \alpha > 1$, *so dass* $\forall m : \lambda_j^{(m)} \leq d_{\max} j^{-\alpha}$. *Dann gilt: Der Verlust des Lernens mit multiplen Kernen ist für jedes* $p \in [1, \ldots, 2]$ *und* $z > 0$ *mit Wahrscheinlichkeit größer gleich* $1 - e^{-z}$ *beschränkt durch*

$$P(l_{\hat{f}} - l_{f^*}) \leq \min_{t \in [p,2]} \; 186\sqrt{\frac{3-\alpha_m}{1-\alpha_m}} \left(d_{\max} D^2 L^2 t^{*2}\right)^{\frac{1}{1+\alpha}} F^{\frac{\alpha-1}{\alpha+1}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}} + \frac{47\sqrt{B} D L M^{\frac{1}{t^*}} t^*}{n} + \frac{(22 B D L M^{\frac{1}{t^*}} + 27F)z}{n}.$$

Wir beobachten, dass durch die obige Schranke Konvergenzraten von maximal $O(M/n)$ erzielt, was eine wesentliche Verbesserung der genauesten aus der Literatur [CMR10] bekannten Rate von $O\big(\sqrt{M/n}\big)$ darstellt, denn typischerweise wird M als konstant angesehen und $n \to \infty$. Bei der typischen Anzahl von $M = 10$ Kernen und $n = 100\,000$ Trainingspunkten enthält die Schranke von [CMR10] einen Faktor von $\sqrt{M/n} = 1/100$, während unser Resultat $M/n = 1/10\,000$ erzielt – ein Unterschied von zwei Größenordnungen.

5 Anwendungen in der Bioinformatik und dem Maschinellen Sehen

Die entwickelte Methodologie wird in [Klo11] auf aktuelle Fragestellungen der Bioinformatik und des Maschinellen Sehens angewendet. Diese Bereiche der Informatik zeichnen sich durch das Auftreten vielfältiger, komplementärer Sichtweisen auf die Daten aus, was die Verwendung multipler Kerne sinnvoll macht. Sämtliche frühere Analysen – mit Ausnahme der Arbeit von [ZO07] zu subzellularer Lokalisierung von Proteinen – scheiterten darin, die Effektivität des Lernens mit multiplen Kernen nachzuweisen, z. B. [SZR06] in der Bioinformatik und [GN09] im Maschinellen Sehen. In dieser Dissertation wird gezeigt, dass unter Verwendung der neuen nichtspärlichen Methodologie die Vorhersagegenauigkeit in beiden Bereichen signifikant erhöht wird.

Visuelle Objekterkennung Dieser Bereich des Maschinellen Sehens beschäftigt sich mit dem Erkennen von Objekten in Bildern – ein schwieriges Unterfangen, denn Objekte können rotiert, verschoben, beleuchtet oder auch von anderen Objekten partiell verdeckt sein. Weiterhin können gewisse Merkmale relevant zum Erkennen einiger, aber wirkungslos zum Erkennen anderer Objekte sein. Beispielsweise ist Farbinformation hilfreich zur Erkennung von Stoppschildern, aber von wenig Nutzen bei der Erkennung von Autos oder Luftballons. Obwohl sich daher ein Einsatz mehrerer Kerne anbietet, zeigten frühere Analysen keinen Vorteil von klassischem Mehr-Kern-Verfahren (siehe z. B. [GN09]).

In [Klo11] analysieren wir den offiziellen, aus 8780 Bildern und 20 Objektklassen bestehenden, Datensatz der PASCAL VOC Challenge 2008. Wir verwenden multiple

Kerne basierend auf Histogrammen gerichteter Gradienten, visueller Wörter, Pixelfarben (letzteres in zwei Farbkanälen) und verschiedenen Pyramidenebenen. Dies resultiert in insgesamt 12 Kernen. Als Gütekriterium verwenden wir das offizielle Fehlermaß des VOC 2008 Wettbewerbs („durchschnittliche Präzision“) sowie den offiziellen Testdatensatz. Die Ergebnisse sind in Abbildung 2 dargestellt: Vertikale Balken geben den Unterschied der durchschnittlichen Präzisionen des Mehr-Kern-Lernens zu einer uniformen SVM an. Wir erkennen, dass unsere neue Methodologie in 18 der 20 Klassen vorteilhaft ist, während das klassische, dünn-besetzte Verfahren keine konsistente Verbesserung ergibt.

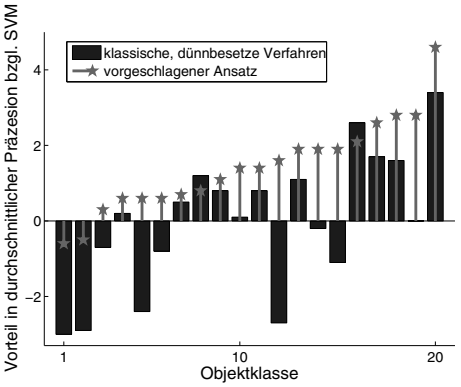


Abbildung 2: Empirische Ergebnisse bei der visuellen Objekterkennung.

Lokalisation von Genen in DNA Die Entdeckung von Transkriptionsstarts (TSS) von RNA Polymerase II bindenden Genen in genomischer DNA stellt einen kritischen Schritt zur Dechiffrierung von Transkription regulierenden Elementen dar. Demzufolge wurde in diesem aktiven Bereich der Bioinformatik eine große Anzahl von Lösungsansätzen vorgestellt. In der unabhängigen Studie [AdPS09] wurden 19 solcher State-of-the-Art Programme verglichen und das Programm ARTS von [SZR06] als genauestes Programm identifiziert. Wir zeigen, dass durch die Verwendung der vorschlagenen Methodologie die Vorhersagegenauigkeit weiter gesteigert wird - über jene des bisher besten Programms [SZR06] hinaus.

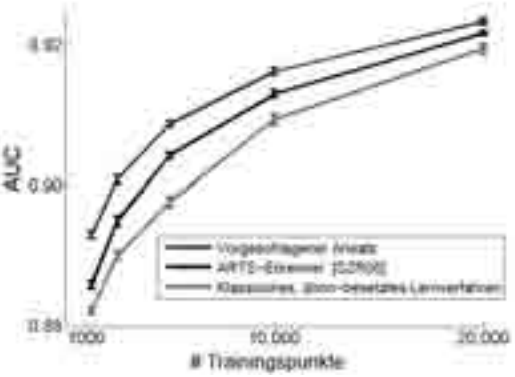


Abbildung 3: Empirische Ergebnisse in der Erkennung von Transkriptionsstarts.

Wie [SZR06] verwenden wir fünf verschiedene Kerne, die komplementäre Eigenschaften des Problems charakterisieren: das TSS Signal, die Promoter-Region, das 1. Exon, die bindende Energie und die Krümmung der DNA. In Übereinstimmung mit [SZR06] setzen wir die Fläche unter der ROC-Kurve (AUC) als Gütekriterium ein und experimentieren auf Grundlage der von [SZR06] bereitgestellten Datensätze. Die Ergebnisse der Analyse sind in Abbildung 3 dargestellt. Vertikale Balken stellen hierbei statistische Standardfehler dar. Wir beobachten, dass das klassische, dünn-besetzte Lernen dem Programm ARTS in Bezug auf alle Trainingsdatengrößen unterliegt – ARTS wiederum wird von der vorgeschlagenen, nicht-spärlichen Methodologie noch einmal deutlich übertroffen.

6 Zusammenfassung

Wir entwickelten eine neue Methodologie zum Lernen mit mehreren Nicht-Linearitäten (oder „Kernen“), einem hochaktuellen und bisher ungelösten Forschungsproblem des Maschinellen Lernens, die – im Gegensatz zu früheren Ansätzen – *keine* dünn besetzten Lösungen liefert. Unsere empirische Evaluierung zu herausfordernden Problemen aus den Bereichen Bioinformatik und Maschinelles Sehen zeigte, dass Vorhersagegenauigkeiten erreicht werden konnten, die den bisherigen Stand der Forschung weit übertreffen. Die vorgeschlagenen Algorithmen zur Optimierung erwiesen sich um bis zu zwei Größenordnungen schneller als existierende und erlaubten, zugleich Zehntausende von Trainingsbeispielen und Tausende von Kernen zu verarbeiten. Die entwickelten Techniken sind grundlegend untermauert durch die statistische Lerntheorie: Wir bewiesen Generalisierungsschranken der Ordnung $O(M/n)$, die weit höhere Konvergenzgeschwindigkeiten aufweisen als vorherige Schranken, welche bestenfalls $O(\sqrt{M/n})$ erzielten.[1]

Schließlich erlauben wir uns, zu bemerken, dass die aktuelle starke Präferenz von dünn besetzten Lernverfahren im Bereich Maschinelles Lernen – oder gar in den Wissenschaften im Allgemeinen – überdacht werden sollte und sich dem hier vorgeschlagenen Ansatz folgend bedeutende neue Perspektiven erschließen lassen. So kann bereits schwache Konnektivität in kausalen, grafischen Modellen dazu führen, dass *sämtliche* Variablen im optimalen Vorhersagemodell aktiv sind. Beispielsweise argumentiert Gelman [Gel11] in den Sozialwissenschaften: „Faktoren sind (fast) nie wirklich Null.“ Basierend auf nicht-spärlichen, multiplen kernbasierten Lernverfahren wurde durch die vorliegende Arbeit eine neue technologische Grundlage zur Fusion von Information geschaffen. Meine zukünftige Forschung wird sich auf deren Verwendung in bioinformatischen und technologischen Anwendungsbereichen konzentrieren, insbesondere in Bezug auf die Fragestellung abhängiger Datenströme.

Danksagung Mein herzlicher Dank gilt meinem Doktorvater Prof. Dr. Klaus-Robert Müller und meinen Mentoren Prof. Peter L. Bartlett, PhD und Prof. Dr. Gilles Blanchard sowie den Mitarbeiterinnen und Mitarbeitern der Abteilungen Maschinelles Lernen der TU Berlin und der UC Berkeley.

Literatur

[AdPS09] T. Abeel, Y. Van de Peer und Y. Saeys. Towards a gold standard for promoter prediction evaluation. *Bioinformatics*, 25(12):313–320, 2009.

[BBM05] P. L. Bartlett, O. Bousquet und S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[CMR10] C. Cortes, M. Mohri und A. Rostamizadeh. Generalization Bounds for Learning Kernels. In *Proceedings, 27th ICML*, Seiten 247–254, 2010.

[CR08] O. Chapelle und A. Rakotomamonjy. Second Order Optimization of Kernel Parameters. In *Proc. of the NIPS Workshop on Kernel Learning*, 2008.

[1] Weitere Beiträge aus [Klo11] wurden aus Platzgründen in dieser Zusammenfassung ausgespart: z. B. die Anwendung der vorgeschlagenen Methodologie in anderen Bereichen der Informatik, wie beispielsweise der Netzwerk Sicherheit [KBD+08, KNB09].

[Gel11] A. Gelman. Causality and Statistical Learning. *American Journal of Sociology*, 117(3):955–966, 2011.

[GN09] P. V. Gehler und S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, Seiten 221–228, 2009.

[KB11] M. Kloft und G. Blanchard. The Local Rademacher Complexity of Lp-Norm Multiple Kernel Learning. In *NIPS 2011, in press*, 2011.

[KB12] M. Kloft und G. Blanchard. On the convergence rate of multiple kernel learning. *Journal of Machine Learning Research (JMLR), accepted*, 2012.

[KBD+08] M. Kloft, U. Brefeld, P. Düssel, C. Gehl und P. Laskov. Automatic feature selection for anomaly detection. In *AISec*, Seiten 71–76. ACM, 2008.

[KBLS08] M. Kloft, U. Brefeld, P. Laskov und S. Sonnenburg. Non-Sparse Multiple Kernel Learning. In *Proc. NIPS Workshop on Kernel Learning*, 2008.

[KBS+09] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller und A. Zien. Efficient and Accurate Lp-Norm Multiple Kernel Learning. In *Advances in Neural Information Processing Systems 22*, Seiten 997–1005. MIT Press, 2009.

[KBSZ11] M. Kloft, U. Brefeld, S. Sonnenburg und A. Zien. Lp-norm Multiple Kernel Learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.

[Klo11] M. Kloft. *ℓ_p-Norm Multiple Kernel Learning*. Dissertation, Technische Universität Berlin, Oct 2011.

[KNB09] M. Kloft, S. Nakajima und U. Brefeld. Feature Selection for Density Level-Sets. In *ECML/PKDD*, Seiten 692–704, 2009.

[KRB10] M. Kloft, U. Rückert und P. L. Bartlett. A Unifying View of Multiple Kernel Learning. In *ECML/PKDD*, Seiten 66–81, 2010.

[LCG+04] G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett und M. I. Jordan. Learning the kernel with semi-definite programming. *JMLR*, 5:27–72, 2004.

[MMR+01] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda und B. Schölkopf. An Introduction to Kernel-based Learning Algorithms. *IEEE N. Netw.*, 12(2):181–201, 2001.

[RBCG08] A. Rakotomamonjy, F. Bach, S. Canu und Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[SRH+10] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl und V. Franc. The SHOGUN Machine Learning Toolbox. *Journal of Machine Learning Research*, Seiten 1799–1802, 2010.

[SZR06] S. Sonnenburg, A. Zien und G. Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):472–480, 2006.

[ZO07] A. Zien und C. S. Ong. Multiclass multiple kernel learning. In *ICML*, Seiten 1191–1198. ACM, 2007.

Marius Kloft, geboren 1980, studierte von 2000 bis 2006 Mathematik, Physik und Informatik an der Philipps-Universität Marburg. Nach dem Diplom in Mathematik (2006) verfasste er zwischen 2007 und 2011 seine Dissertation an der Technischen Universität Berlin, dem Fraunhofer Institut FIRST und der University of California at Berkeley. Er hat Forschungsaufenthalte an dem Friedrich-Miescher-Laboratorium der Max-Planck Gesellschaft (Tübingen) und der Universität Tromso (Norwegen) verbracht und ist zur Zeit im Rahmen eines Projektes zur computergestützten Genomanalyse in Kooperation mit Laboratorien in Tübingen und New York an der Technischen Universität Berlin tätig.