BestMasters

Daniel Lückehe

Hybride Optimierung für Dimensionsreduktion

Unüberwachte Regression mit Gradientenabstieg und evolutionären Algorithmen

Springer Vieweg

BestMasters

Mit „BestMasters“ zeichnet Springer die besten Masterarbeiten aus, die an renommierten Hochschulen in Deutschland, Österreich und der Schweiz entstanden sind. Die mit Höchstnote ausgezeichneten Arbeiten wurden durch Gutachter zur Veröffentlichung empfohlen und behandeln aktuelle Themen aus unterschiedlichen Fachgebieten der Naturwissenschaften, Psychologie, Technik und Wirtschaftswissenschaften.
Die Reihe wendet sich an Praktiker und Wissenschaftler gleichermaßen und soll insbesondere auch Nachwuchswissenschaftlern Orientierung geben.

Daniel Lückehe

Hybride Optimierung für Dimensionsreduktion

Unüberwachte Regression mit Gradientenabstieg und evolutionären Algorithmen

Springer Vieweg

Daniel Lückehe
Oldenburg, Deutschland

BestMasters
ISBN 978-3-658-10737-6 ISBN 978-3-658-10738-3 (eBook)
DOI 10.1007/978-3-658-10738-3

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Vieweg

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Fachmedien Wiesbaden ist Teil der Fachverlagsgruppe Springer Science+Business Media (www.springer.com)

Vorwort

In diesem Buch wird ein neues hybrides Verfahren zur Dimensionsreduktion methodisch erarbeitet und durch Tests mit vorhandenen Methoden verglichen. Das Buch ist sowohl für Experten im Gebiet der Dimensionsreduktion als auch für Einsteiger mit Grundkenntnissen über maschinelles Lernen geeignet. Hochdimensionale Daten liegen heutzutage in vielen Bereichen wie beispielsweise der Medizin und Astronomie vor. Dabei ist vom Begriff *Big Data* zu hören. Häufig ist es schwierig Informationen aus den sehr großen, komplexen Datensätzen zu gewinnen. Hierbei kann eine Dimensionsreduktion helfen. So können die Daten beispielsweise auf einen zweidimensionalen, latenten Raum abgebildet und für den Menschen visuell erfassbar gemacht werden. Eine Dimensionsreduktion kann auch als vorverarbeitender Schritt für eine Klassifikation oder Regression verwendet werden. In diesem Buch wird die unüberwachte Kernel-Regression als Verfahren zur Dimensionsreduktion genutzt und die damit erzeugten Ergebnisse mit dem Gradientenabstiegsverfahren optimiert. Die Parameter werden durch einen evolutionären Algorithmus gesteuert und zur Verbesserung der Flexibilität wird eine variable Kernel-Funktion entwickelt. In Tests werden verschiedene, alternierende Variationen des Verfahrens qualitativ und quantitativ bewertet und mit vorhandenen Methoden wie *Locally Linear Embedding* und *Isometric Mapping* verglichen.

Da dieses Buch auf einer Masterarbeit basiert, ist im weiteren Verlauf des Buches die Formulierung *diese Arbeit* als Synonym für *dieses Buch* zu betrachten.

Ausschnitte und Erweiterungen der Arbeit sind auf internationalen Konferenzen veröffentlich worden und können als weiterführende Literatur genutzt werden. Die wichtigsten neuen Erkenntnisse erweitert um Experimente auf zusätzlichen Datensätzen und einer ausführlichen

Analyse verschiedener Populationsgrößen für den Einbettungsprozess sind in dem Paper *Alternating Optimization of Unsupervised Regression with Evolutionary Embeddings* publiziert, welches auf der *Applications of Evolutionary Computation - 18th European Conference (EvoApplications), 2015* in Kopenhagen präsentiert und dabei mit dem *Best Paper Award* in der Rubrik *Evolutionary Computation in Image Analysis, Signal Processing and Pattern Recognition (EvoISAP)* ausgezeichnet wurde. Die variable Kernel-Funktion wurde auf der *Genetic and Evolutionary Computation Conference (GECCO), 2014* in Vancouver mit dem Titel *A Variable Kernel Function for Hybrid Unsupervised Kernel Regression* veröffentlicht. Eine Erweiterung, in der Muster mit relativ hohem Einfluss auf den Datenraumrekonstruktions-Fehler entfernt und neu im latenten Raum platziert werden, wurde auf der *24th International Conference on Artificial Neural Networks (ICANN), 2014* in Hamburg mit dem Titel *Leaving Local Optima in Unsupervised Kernel Regression* vorgestellt.

Herzlich möchte ich mich bei allen Menschen bedanken, die mit moralischer Unterstützung, inspirativen Ideen und fachlichen Tipps dieses Buch erst möglich gemacht haben. Besonders zu erwähnen sind die Fachhochschule Wilhelmshaven und die Universität Oldenburg, die mich optimal während des Studiums unterstützt haben. Dabei gilt insbesondere mein Dank Jun.-Prof. Dr. Oliver Kramer und Dipl.-Phys. Nils André Treiber für die engagierte Betreuung meiner Masterarbeit. Auch dem Team PTE1 2 (ATLAS ELEKTRONIK GmbH), ausdrücklich Jörg Bade, gilt mein Dank für die tolle Zusammenarbeit, die mir die Vereinbarkeit von Beruf und parallelem Studium ermöglicht hat. Besonders möchte ich mich bei meinen Eltern und meinen Freunden, insbesondere Mandy Ludwig und Julia Volmer für das konstruktive Korrekturlesen, bedanken.

Bremen, April 2015
Daniel Lückehe

Inhaltsverzeichnis

1. Motivation

Ziel dieser Arbeit ist es, ein neues hybrides Verfahren zur Dimensionsreduktion methodisch zu erarbeiten, zu implementieren und durch Tests mit vorhandenen Methoden zu vergleichen. Hochdimensionale Daten liegen heutzutage in vielen Bereichen vor. Darunter fallen beispielsweise Datensätze über Nutzerverhalten, die zur zielgerichteten Werbung genutzt werden können, visuell erfasste Daten, in denen Zeichen erkannt werden sollen, Daten über Kontobewegungen zur Erkennung von Betrugsfällen, Anwendungen im medizinischen Bereich, auditiv erfasste Daten, die zur Spracherkennung genutzt werden, Daten aus dem Gebiet der Astronomie und viele weitere. Dabei ist vom Begriff *Big Data* zu hören: Sehr große, komplexe Datensätze, die viele Informationen beinhalten, die jedoch aufwendig aus dem Datensatz gewonnen werden müssen [2]. Um diese hochkomplexen Daten besser verarbeiten und erfassen zu können, kann eine Dimensionsreduktion helfen. In diesem Gebiet besteht noch ein hoher Forschungsbedarf.

Bevor näher auf die Grundlagen eingegangen wird, soll ein Überblick über das Verfahren gegeben werden. Inspirationsquelle für das neue Verfahren war die Dimensionsreduktion *Unüberwachte Nächste Nachbarn* (UNN) von Kramer [9]. Als Regressions-Modell soll dabei anstatt der nächsten Nachbarn der Nadaraya-Watson-Schätzer zum Einsatz kommen, wie von Klanke [7] bekannt. Das Verfahren soll den latenten Raum wie UNN iterativ erzeugen. Die so erzeugten Lösungen sollen durch den Datenraumrekonstruktions-Fehler, welcher mit Hilfe einer Fehler-Funktion (Loss-Funktion) bestimmt wird, bewertet und optimiert werden. Zum Einsatz kommt dabei das Verfahren des Gradientenabstiegs, siehe [13]. Es geht in dieser Arbeit darum, zu zeigen, welche Ergebnisqualität mit dem entwickelten Verfahren realisierbar ist.

D. Lückehe, *Hybride Optimierung für Dimensionsreduktion,* BestMasters,
DOI 10.1007/978-3-658-10738-3_1, © Springer Fachmedien Wiesbaden 2015

1.1. Dimensionsreduktion

Bei einer Dimensionsreduktion geht es darum, Muster aus einem hochdimensionalen Datenraum auf einen niedrigdimensionalen latenten Raum abzubilden. Dabei ist für die Abbildung eine Funktion $F : \mathbf{y} \rightarrow \mathbf{x}$ für Muster $\mathbf{y} \in \mathbf{Y} \subset \mathbb{R}^d$ und latente Punkte $\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^q$ mit $q < d$ gesucht, die Nachbarschaften und Abstände bestmöglich erhält [10].

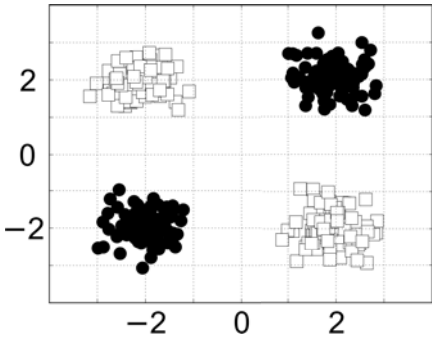

Abbildung 1.1.: Datenraum

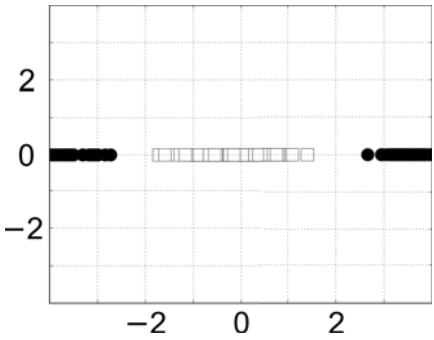

Abbildung 1.2.: Latenter Raum

Die Abbildungen 1.1 und 1.2 zeigen beispielhaft einen Datenraum und einen durch eine Abbildungsfunktion F entstandenen latenten Raum. Die Muster im Datenraum liegen in der Form $(y_1, y_2) \in \mathbb{R}^2$ vor. Im latenten Raum haben die Muster die Form $(x_1) \in \mathbb{R}^1$. Für F gilt nun die einfache Abbildung: $F_{\text{Beispiel}} : (y_1, y_2) \rightarrow (x_1) = (y_1 + y_2)$. Würden beispielsweise die schwarzen Punkte das Label *nein* und die weißen Quadrate das Label *ja* besitzen, so wäre mit Hilfe des latenten Raums ein sehr einfacher Klassifikator denkbar:

```
if abs(x) < 2:
  print "Die Antwort lautet: Ja."
else:
  print "Die Antwort lautet: Nein."
```

Es ist möglich mit einem Klassifikator, der die Muster aus dem Datenraum nutzt, ebenso gute Ergebnisse zu erzielen, wobei die