Schriftenreihe CIplus, Band 1/2014

Herausgeber: T. Bartz-Beielstein, W. Konen, H. Stenzel, B. Naujoks

SOMA – Systematische Optimierung von Modellen in IT- und Automatisierungstechnik

Wolfgang Konen und Patrick Koch

Fachhochschule Köln
Cologne University of Applied Sciences

SOMA

Systematische Optimierung von Modellen in IT- und Automatisierungstechnik

Schlussbericht

Förderlinie IngenieurNachwuchs 2009 (Informatik)
im Rahmen des Programms Forschung an Fachhochschulen

Prof. Dr. Wolfgang Konen
Dr. Patrick Koch

Institut für Informatik
Fakultät für Informatik und Ingenieurwissenschaften
Fachhochschule Köln
14. März 2014

Inhaltsverzeichnis

Kurzfassung Das im Rahmen der Förderlinie IngenieurNachwuchs geförderte Forschungsvorhaben “Systematische Optimierung von Modellen für Informations- und Automatisierungstechnik” (kurz: SOMA) startete im August 2009. Eine wesentliche Zielsetzung war die Entwicklung und Optimierung von Modellen zur Prognose von Zielgrößen. Ein wichtiges Merkmal ist dabei die effiziente Optimierung dieser Modelle, welche es ermöglichen soll, mit einer streng limitierten Anzahl an Auswertungen gute Parametereinstellungen zu bestimmen. Mithilfe dieser genaueren Parametrierungen der unterliegenden Modelle können unter Einbeziehung neuer merkmalserzeugender Verfahren insbesondere für kleine und mittelständische Unternehmen verbesserte Lösungen erzielt werden. Als direkter Gewinn derartiger Verbesserungen konnte für KMUs ein geeignetes Framework für Modellierungs- und Prognoseaufgaben bereitgestellt werden, sodass mit geringem technischem und personellen Aufwand performante und nahezu optimale Lösungen erzielt werden können. Dieser Schlussbericht beschreibt die im Projekt durchgeführten Maßnahmen und Ergebnisse.

1 Kurzdarstellung

1.1 Aufgabenstellung

In dem Forschungsprojekt *Systematische Optimierung von Modellen für IT- und Automatisierungstechnik* (SOMA) sollten spezielle Modelle zur Prognose von Zielgrößen entworfen und für ihre jeweiligen Einsatzzwecke optimiert werden. Als Anwendungsgebiete waren einerseits ingenieurwissenschaftliche Anwendungen als auch Anwendungen im Business Intelligence vorgesehen. Hierbei ist hervorzuheben, dass durch die gegebene Generalisierbarkeit der im Projekt SOMA eingesetzten Modelle auch Verwendungen in anderen Gebieten vorstellbar sind. Ein besonderer Fokus lag bei der Modellbildung insbesondere auf der Optimierung der freien Modellparameter. Hierbei wurden speziell angepasste Optimierverfahren verwendet, um innerhalb kürzester Zeit nahezu optimale Parametereinstellungen für die Modelle bestimmen zu können. Dabei ist festzuhalten, dass die Parametereinstellung derartiger Prognosemodelle für den Anwender oftmals eine sehr komplexe Hürde darstellt, da sie ein spezielles Fachwissen über die ablaufenden Prozesse erfordert. Aus diesen Gründen ist die genaue Einstellung der Parameter oftmals ein zeitaufwändiges Unterfangen, sodass in der Praxis oftmals auf eine systematische Optimierung verzichtet wird.

Im Projekt SOMA wurde gezeigt, dass durch den Einsatz der sogenannten *modellbasierten Optimierung* sehr gute Parameter für derartige Modelle bestimmt werden werden können. Bei der modellbasierten Optimierung wird im Gegensatz zur klassischen Optimierung ein sog. Meta- oder Surrogat-Modell der Zielfunktion gelernt. Mit Hilfe dieses Modells können Auswertungen auf der realen Zielfunktion weitesgehend reduziert werden und ein großer Anteil der Optimierung findet auf dem Meta-Modell statt. Allein mit Hilfe der verbesserten Parameterwerte gelang es anschließend, die Prognosegenauigkeit der unterliegenden Anwendungen signifikant zu verbessern und zu stabilisieren. Verschiedene Modelloptimierungen wurden in dem Projekt SOMA vorgenommen, welche jeweils detailliert analysiert und der Fachwelt zugänglich gemacht worden. Heute ist es möglich mit den im Projekt SOMA entwickelten Software-Tools ohne größeren Aufwand und ohne besonderes Problemwissen eine deutlich verbesserte Genauigkeit der komplexen Modellierungsaufgaben in industriellen Prozessen zu bekommen.

1.2 Vorraussetzungen zur Durchführung des Projektes

Das Projekt SOMA hatte eine Gesamtlaufzeit von insgesamt drei Jahren und 11 Monaten. Die Durchführung des Projektes fand am Institut für Informatik an der Fachhochschule Köln, Campus Gummersbach statt. Hier ist der Projektleiter Wolfgang Konen Professor für angewandte Mathematik. Der Projektleiter war während des gesamten Projektzeitraums in dem an Fachhochschulen üblichen Umfang in Hochschulverwaltung und Lehre eingebunden. Zu Projektbeginn im August 2009 nahm Herr Dipl.-Inform. Patrick Koch seine Tätigkeit im Projekt SOMA auf. Herr Koch war während des gesamten Projektlaufzeit in Vollzeit als wissenschaftlicher Mitarbeiter im Projekt SOMA angestellt. Weitere Nachwuchswissenschaftler (studentische und wissenschaftliche Hilfskräfte) wurden nach Bedarf angeworben und eingestellt. Weitere Vorraussetzungen waren für dieses Forschungsvorhaben nicht notwendig.

1.3 Planung und Ablauf des Projektes

Das Projekt SOMA wurde anhand eines detaillierten Projektplans in sechs verschiedene Module eingeteilt. In Abb. 1 ist dazu der Arbeitsplan des Projektes SOMA in Form eines Gantt-Diagramms dargestellt. Die sechs beschriebenen Module umfassen im Wesentlichen die folgenden Arbeitspakete:

- B Bereitstellung und Test CI-Verfahren: Dieses Modul diente der Zusammenführung verschiedener CI-Verfahren, sodass eine Nutzung in einem gemeinsamen Optimierungs-Framework ermöglicht wurde
- C Case Studies: Beinhaltet die Projekte, welche in Form von studentischen Projekten (Case Studies) im Masterstudiengang Automation & IT bearbeitet wurden
- E Entwurfsalgorithmen Feature Generierung: Umfasst die Bereiche zur Merkmalsvorverarbeitung und Feature-Konstruktion, welche wesentlicher zu den im Projekt SOMA entwickelten Prognoseverfahren dient
- S SPO-Metastrategien und alternative Optimierer: In diesem Modul werden alle wesentlichen Optimierverfahren beschrieben, welche zur Optimierung der Parameter von Prognoseverfahren eingesetzt wurden.
- P Promovend: Zu der Teamzusetzung der wissenschaftlichen Nachwuchsgruppe gehörte der Promovend Patrick Koch. Herr Patrick Koch war mit voller Stelle als wissenschaftlicher Mitarbeiter in dem Projekt SOMA angestellt und begleitete die Entwicklung in den verschiedenen Arbeitspaketen.
- K Koordination: Dieses Modul diente der besseren Abstimmung der einzelnen Arbeitspakete und der Vernetzung der wissenschaftlichen Arbeiten untereinander.

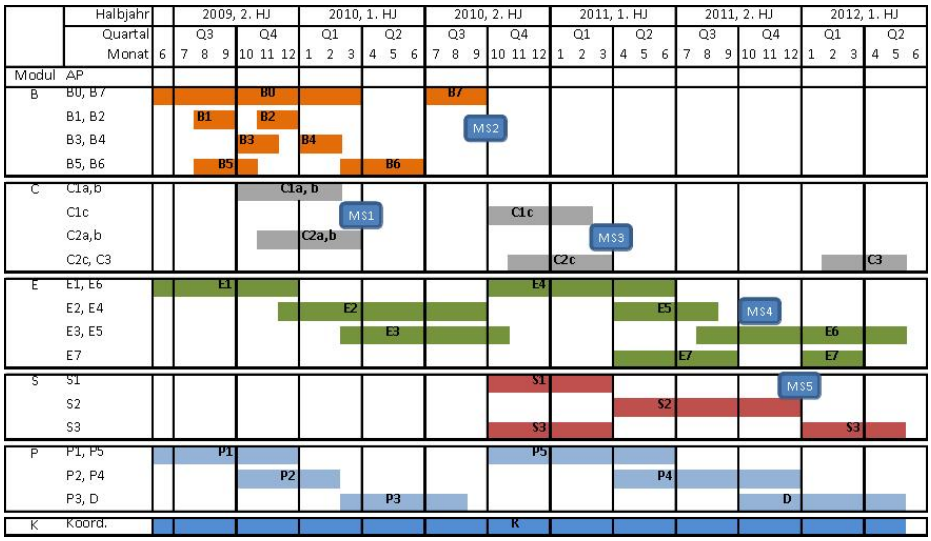

Abbildung 1: Projektplan SOMA

Bei der Definition der Arbeitspakete wurde insbesondere auf eine nicht zu langfristige Planung geachtet, sodass während des Projektablaufs ein interner Abgleich möglich war und auch durchgeführt wurde.

Bereitstellung und Test CI-Verfahren Ziel des Moduls *Bereitstellung und Test CI-Verfahren* war es, verschiedene Verfahren zur Prognose von Daten in einem geeigneten Framework zusammenzuführen. Zur Einstellung der freien Parameter erfolgte frühzeitig eine Einarbeitung in das Optimierungsframework *Sequential Parameter Optimization Toolbox* (SPOT), welches als mögliches Werkzeug für die verschiedenen Optimierungsaufgaben diente. Es gelang eine Auswahl geeigneter Prognose-Verfahren auf synthetischen und praktischen Beispielen gegenüber zu stellen. Als Verfahren wurden z.B. Methoden wie Support Vector Machines (SVM) und Ensembles von Entscheidungsbäumen (Random Forest) berücksichtigt. Zur Evaluierung der eingesetzten Prognosemodelle wurden verschiedene Arten der Aufteilung in Trainings- und Testmengen und verschiedene Fehlermaße berücksichtigt und miteinander verglichen.

Case Studies aus IT und Automatisierungstechnik In dem Masterstudiengang "Automation & IT" an der Fachhochschule Köln werden jeweils im Sommersemester verschiedene *Case Studies* angeboten. Die Case Studies sind Lerngruppen von ca. 2-4 Studierenden, welche ein vorgegebenes Thema als Projekt bearbeiten. Lernziele sind dabei Projektmanagement, Literaturrecherche, Präsentation von Ergebnissen, Programmierung, sowie praxisnahe Untersuchungen und statistische Auswertungen. Für die Studierenden stellten die für das Projekt SOMA angebotenen Case Studies eine interessante Umgebung dar, da praxisrelevante Themen bearbeitet werden konnten und eine direkte Verbindung zu aktuellen Forschungsthemen gegeben war.

Entwurfsalgorithmen Feature Generierung + Feature Evolution Verfahren zur Merkmalsextraktion und Merkmalsgenerierung spielen in immer komplexer werdenden Anwendungen eine wesentliche Rolle. Das Arbeitspaket *Entwurfsalgorithmen Feature Generierung + Feature Evolution* diente dazu, vorhandene Algorithmen in der Literatur zu implementieren und für verschiedene Anwendungsbeispiele einzusetzen. Bei den im Projekt SOMA durchgeführten Tests zeigte sich, dass durch geeignete Verfahren zur Merkmalsauswahl und Generierung deutliche Verbesserungen der Prognosequalität erzielt werden konnten.

SPO-Metastrategien und alternative Optimierer In dem Modul SPO-Metastrategien und alternative Optimierer sollten Verfahren zur Optimierung der Modellparameter erfasst werden. Zur Auswahl geeigneter Optimierer musste zunächst untersucht werden, welche verschiedenen Parameter überhaupt relevant für die Anwendung waren. Hierzu wurde etwa bestimmt, wie sensitiv das Prognosemodell auf Einstellungen einzelner Parameter reagiert. Weiterhin wurde experimentell analysiert, inwieweit Korrelationen zwischen Parametern auftreten können.

Als weiterer Schritt war es wichtig festzuhalten, welche verschiedenen Parametertypen in der Modellbildung auftreten können. Dazu spielten insbesondere die folgenden Punkte eine wichtige Rolle:

- Handelt es sich um kontinuierliche oder diskrete Parameter?
- Sind die Parameter in bestimmten Bereichen restringiert, oder frei einstellbar?
- Handelt es sich um numerische oder faktorielle Parameter?
- Wie sensitiv reagiert das Prognosemodell auf eine Änderung der Parameter?
- Bestehen mögliche Wechselwirkungen mit anderen Parametern?

Darauf aufbauend konnte eine Auswahl an Optimierverfahren bestimmt werden, die anschließend an praxisnahen Anwendungen getestet worden sind. Zur Beurteilung der Qualität einzelner Optimierverfahren wurde der Prognosefehler auf unabhängigen Daten mit Bezug auf die notwendigen Funktionsauswendungen betrachtet.

Promovend Der in dem Projekt SOMA tätige wissenschaftliche Mitarbeiter Patrick Koch führte eigenständige Untersuchungen und Analysen zur Parameteroptimierung von Prognoseverfahren durch. Es fand eine enge inhaltliche Zusammenarbeit mit dem Projektleiter statt. Weiterhin gehörten koordinative Aufgaben der im Projekt SOMA durchgeführten Aufgaben zu seinem Tätigkeitsbereich. Herr Koch stellte in Abstimmung mit dem Projektleiter zu Beginn des Projektes SOMA den Kontakt zu der Partneruniversität *Universiteit Leiden* (Niederlande) her. Bereits Ende 2009 erklärte sich Prof. Dr. Thomas Bäck von der Universität Leiden (Niederlande) bereit, das Promotionsvorhaben von Herrn Koch zu unterstützen. Prof. Bäck ist Lehrstuhlinhaber an der Universität Leiden und bearbeitet fachlich am Leiden Institute of Advanced Computer Science (LIACS) ein thematisch vergleichbares Forschungsgebiet. Zur besseren Abstimmung des Promotionsvorhabens wurden frühzeitig die notwendigen Schritte (Einschreibung, Themenauswahl) eingeleitet. Herr Koch konnte seine Promotion kurz nach Ablauf des Projektes SOMA am 29.10.2013 erfolgreich abschließen.

1.4 Wissenschaftlicher und technischer Stand

Der wissenschaftliche und technische Stand umfasst die Auswahl geeigneter Prognoseverfahren und Algorithmen zur Merkmalsgewinnung. Außerdem war es von besonderer Bedeutung, die Modellparameter und auftretenden freien Parameter in der Prognoseanwendung einzustellen. Hierzu war es notwendig, einen aktuellen Überblick über zahlreiche Verfahren zur Optimierung von Modellparametern zu gewinnen.

Bei den Themengebieten handelt es sich um schnell wachsende und sich stetig ändernde Forschungsfelder, sodass es nicht einfach war, einen kompletten Überblick über die Gebiete zu geben. Es hat sich jedoch herausgestellt, dass sich einige Verfahren in den jeweiligen Forschungsgebieten und im praktischen Einsatz etabliert haben. Auf Seite der Prognoseverfahren sind hier insbesondere Support Vector Machines [5] sowie Ensembles von Entscheidungsbäumen, wie z.B. Random Forests [3] hervorzuheben. Auf Seite der Optimierverfahren ist zu differenzieren zwischen modellbasierten und klassischen Ansätzen. Auf klassischer Seite seien

hier die CMA-ES von Hansen und Ostermeier [11] und Differential Evolution von Storn und Price [31] genannt, welche beide gute Ergebnisse in der Praxis liefern können. Bei den modellbasierten Ansätzen gibt es Unterschiede bezüglich der Methoden aufgrund der verwendeten Surrogat-Modelle. Die sequentielle Parameter Optimierung von Bartz-Beielstein u.a. [1] stellt hier ein sehr universell einsetzbares Tool dar, welches eine flexible Anpassung der verwendeten Surrogat-Modelle erlaubt. In der Praxis hat sich jedoch gezeigt, dass mit Hilfe von Kriging-Modellen (in der Fachliteratur häufig als Gaussian Processes bezeichnet [26]) in vielen Fällen die besten Ergebnisse erzielt werden können. Aus diesem Grund wurde in den in diesem Projekt angestellten Untersuchungen auf den Einsatz weiterer Surrogat-Modelle weitestgehend verzichtet. Dies stellt jedoch keinen Nachteil dar, da in der späteren Anwendung ein Austausch des Surrogat-Modells sehr einfach vorgenommen werden kann. Weitere verfügbare Modelle liegen in den erstellten Implementierungen vor, und können daher direkt eingesetzt werden. In den angestellten Untersuchungen hat sich der Einsatz der sog. Efficient Global Optimization [13] bewährt, ein Verfahren, welches Kriging als Surrogat-Modell nutzt, und eine globale Optimierung des Suchraumes durch Verwendung des sog. Expected Improvement Infill-Kriteriums erlaubt. Dabei wird neben der Lösungsqualität auch die Unsicherheiten der zu untersuchenden Parameter berücksichtigt.

Auf Seite der Feature-Methoden sind insbesondere Verfahren zur Dimensionsreduktion relevant, als auch transformierende und konstruierende Verfahren zu nennen, wie z.B. Slow Feature Analysis [32] und Genetic Programming [25].

Die behandelten Inhalte wurden auf zahlreichen nationalen und internationalen Fachtagungen und Konferenzen vorgestellt (s. Tabelle 1). Auf nationaler Ebene existiert der Workshop Computational Intelligence der Gesellschaft Mess- und Automatisierungstechnik, der jährlich in Dortmund stattfindet. Auf internationaler Ebene gibt es die jährlich stattfindende Genetic and Evolutionary Computation Conference (GECCO), als auch die zweijährlich stattfindende Conference on Parallel Problem Solving from Nature (PPSN). Weiterhin findet jährlich der Congress on Evolutionary Computation (CEC) statt, welcher allerdings alle zwei Jahre im Rahmen des World Congress on Computational Intelligence (WCCI) ausgetragen wird. Im Rahmen der WCCI werden ebenfalls renommierte Konferenzen wie etwa die International Joint Conference on Neural Networks (IJCNN) in einer Konferenz zusammengefasst.

Eine Liste der Fachtagungen, auf denen SOMA-Projektbeteiligte Vorträge hielten, ist nachfolgend aufgeführt:

Literaturrecherche Es wurde eine umfassende Literaturrecherche durchgeführt, die sowohl die veröffentlichten Arbeiten auf Fachtagungen und Konferenzen, als auch Zeitschriftenbeiträge und Bücher umfasste. Außerdem wurde zu Projektbeginn eine Patentrecherche mittels Patent-Datenbanken vorgenommen, um patentrechtlich geschützte Werkzeuge und Verfahren zu berücksichtigen. Nach abschließender Erkenntnis sind die im Projekt SOMA eingesetzten Verfahren unter freien Lizenzen (z.B. GNU Public Licence) verfügbar, bzw. liegen in für die Forschung frei verwendbaren Implementierungen vor (Open Source Lizenzen).

Tabelle 1: Liste von Fachtagungen mit Vorträgen der Projektmitarbeiter

Jahr	Konferenz / Workshop	Vortragender	Ort
2009	GMA CI Workshop	Konen	Witten-Bommerholz
	GECCO 2009	Konen	Montreal, Kanada
2010	BIOMA 2010	Konen	Ljubljana, Slovenien
	WCCI 2010	Koch	Barcelona, Spanien
	PPSN 2010	Koch	Krakau, Polen
	GMA CI Workshop	Koch	Witten-Bommerholz
2011	GECCO 2011	Koch	Dublin, Irland
	Dagstuhl Workshop „Organic Computing“	Konen	Dagstuhl
	Gesellschaft für Klassifikation	Koch	Frankfurt
	GMA CI Workshop	Konen	Dortmund
2012	PPSN 2012	Konen, Koch	Taormina, Italien
	GMA CI Workshop	Koch	Dortmund
2013	European Conference on Data Analysis 2010	Koch	Luxemburg
	GMA CI Workshop	Konen	Dortmund

Legende:

BIOMA	International Conference on Bioinspired Optimization Methods and their Applications
GECCO	International Conference on Genetic and Evolutionary Computation
GMA CI	Gesellschaft für Mess- und Automatisierungstechnik, Fachausschuss Computational Intelligence
PPSN	International Conference on Parallel Problem Solving from Nature
WCCI	World Congress on Computational Intelligence

1.5 Zusammenarbeit mit anderen Stellen

In dem Projekt SOMA sind zahlreiche Kooperationen mit Unternehmen und Arbeitsgruppen durchgeführt worden, sodass eine sichtbare Vernetzung des Zuwendungsempfängers und anderen Institutionen und Unternehmen entstanden ist.

Ein intensiver Austausch fand mit der Arbeitsgruppe von Prof. Bongards der Fachhochschule Köln statt. Zusammen mit den im ingenieurwissenschaftlichen Bereich tätigen Forschern sind interdisziplinär verschiedene Anwendungen bearbeitet worden, die immer wieder als Praxisbeispiele in Case Studies und Publikationen genutzt werden konnten. Gute Kontakte bestehen auch zu dem Lehrstuhl für Computergestützte Statistik von Claus Weihs an der TU Dortmund und dem Lehrstuhl von Thomas Zielke an der FH Düsseldorf. Mit der Arbeitsgruppe von Prof. Bartz-Beielstein besteht eine gute Zusammenarbeit innerhalb der Fachhochschule Köln. Im Rahmen des Projektes ist die Forschungsstelle *Computational*

Intelligence, Optimierung und Data Mining (CIOP) an der Fachhochschule Köln gegründet worden (Webseite: `http://www.gociop.de`), die auch nach Abschluss des Projektes SOMA Bestand hat. Im Rahmen der Forschungsstelle werden regelmäßig Kolloquien mit Vorträgen z. T. internationaler Forscher gehalten. Die Forschungsstelle dient zudem der besseren Vernetzung der Lehrstühle an der Fakultät für Informatik und Ingenieurwissenschaften untereinander. Außerdem ist die Forschungsstelle eine Anlaufsstelle für Nachwuchswissenschaftler im Rahmen von studentischen Projekten und Bachelor- und Masterarbeiten.

Mit der Ruhr-Universität Bochum besteht ein intensiver Kontakt zu Prof. Dr. Laurenz Wiskott, einem der Leiter des Instituts für Neuroinformatik. In dieser Kooperation werden Themen wie Feature-Entwurfsalgorithmen auf Basis der Slow Feature Analysis sowie Verfahren zur Feature-Extraktion betrachtet.

Der Projektmitarbeiter Patrick Koch begann im November 2009 sein Promotionsvorhaben an der Universität Leiden, welches er kurz nach dem Projektende von SOMA im Oktober 2013 erfolgreich abschließen konnte [14]. Neben der Promotion des Mitarbeiters bestehen mit der Universität Leiden weitere aktive Kontakte. Zum Zeitpunkt der Berichterstellung befand sich noch eine weitere Publikation in Revision.

Weiterhin bestehen gute Kontakte zu regionalen Industrieunternehmen und weiteren international angesehenen Forschergemeinschaften.

2 Eingehende Darstellung

In diesem Abschnitt werden die erzielten Ergebnisse im Detail vorgestellt. An den Stellen an denen Ergebnisse bereits auf Konferenzen und Fachzeitschriften veröffentlicht worden sind, werden die entsprechenden Literaturhinweise angegeben.

2.1 Verwendung der Zuwendung und erzielte Ergebnisse

Die im Projekt SOMA erzielten Ergebnisse sind im Folgenden detailliert aufgeführt. Die Gliederung orientiert sich dabei an den in dem Projektplan definierten Arbeitspaketen.

Bereitstellung und Test CI-Verfahren Im Rahmen des Projektes ist das Open-Source Framework *Tuned Data Mining in R* (TDMR) entwickelt worden [21], [22], [23], [24]. Das Framework dient zur vereinfachten Durchführung und Etablierung von Prognoseverfahren mit einer integrierten Optimierung der freien Parameter. Das Framework ist unter der GPL Lizenz verfügbar gemacht worden. Mittlerweile ist für die R-Version 3.0.1 eine aktuelle Version 1.0.1 von TDMR verfügbar. Es wurde darauf geachtet, dass eine relativ einfache Benutzbarkeit und Dokumentation des Frameworks gegeben ist, sodass auch für den fachfremden Nutzer eine schnelle Einarbeitung möglich ist.

In TDMR sind verschiedene Prognoseverfahren enthalten. Dazu zählen unter anderem Random Forest, Support Vector Machines und Naive Bayes. TDMR erlaubt es zusätzlich weitere Prognoseverfahren sehr einfach in das Framework

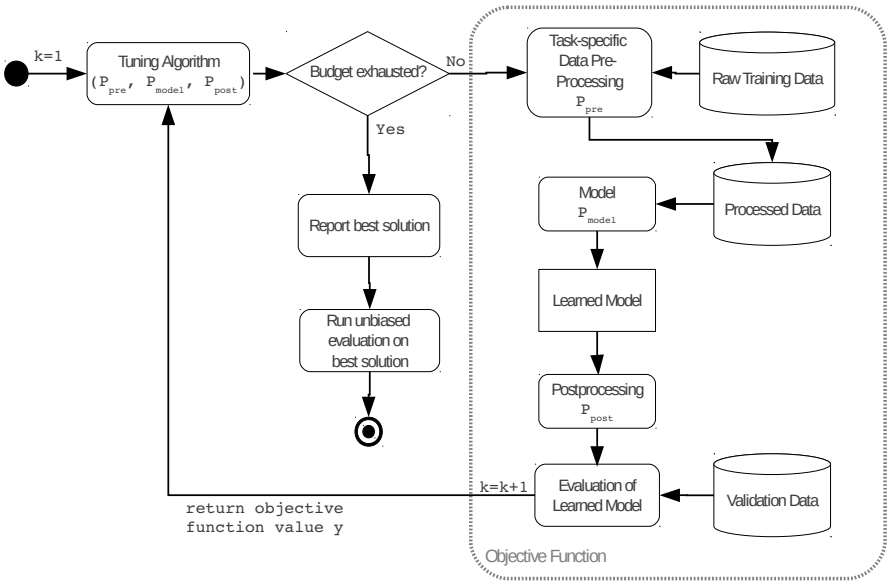

Abbildung 2: Schematischer Ablauf eines Tuned Data Mining Prozesses.

zu integrieren, sodass die durch die Sprache R bereitgestellten Verfahren leicht ergänzt werden können.

Die Optimierung von Parametern für ein Prognoseverfahren ist in Abb. 2 schematisch dargestellt. Dabei wird von einer umfassenden Optimierung der Prozessparameter ausgegangen. Die Optimierung erfolgt durch ein frei wählbares Optimierverfahren. Im Projekt SOMA wurde eine Reihe unterschiedlicher Optimierer in Betracht gezogen. Dazu zählen unter anderem:

- Latin Hypercube Sampling[1]
- CMA-ES von Hansen und Ostermeier [10, 11]
- BFGS von Broyden [4], Fletcher [7], Goldfarb [8] und Shanno [29]
- SPO von Bartz-Beielstein u.a. [1]

In dem gewählten Optimierverfahren oder *Tuning Algorithm* werden zunächst die Modellparameter und Parameter für die Vorverarbeitungsschritte festgelegt. Anschließend wird das Prognosemodell mit dem durch den Optimierer festgelegten Parametereinstellungen erstellt und ausgewertet (Objective Function). Die Zielfunktion beinhaltet dabei den kompletten Data Mining Prozess, beginnend von der Auswahl der Trainingsdaten, als auch die Evaluation auf unabhängigen Testdaten. Die Ergebnisse des Lernverfahrens werden anschließend zurück an

[1] Obwohl es sich bei LHS um keine wirkliche Optimierung im klassischen Sinn handelt, wurde das Verfahren in TDMR integriert. Es diente in empirischen Studien als Referenzverfahren.

den Optimierer übergeben und daraus ausgehend die weiteren Entscheidungen getroffen. Das Verfahren verläuft sequentiell, bis die vorab definierte Budgetgrenze – etwa die Dauer einer gewissen Zeit überschritten ist, oder die Anzahl an Funktionsauswertungen erreicht wurde. Bei Erreichen des Abbruchkriteriums, wird die beste Parametereinstellung an den Nutzer zurück gegeben und erneut anhand unabhängiger Daten ausgewertet, sodass ein mehrfach validiertes Ergebnis vorliegt. In diesem Punkt ist es möglich und auch in TDMR vorgesehen, anhand der existierenden Ergebnisse weitere Analyseverfahren einzusetzen (z.B. Sensitivitätsanalysen oder Estimation of Distribution (EDA) Methoden).

SPO-Metastrategien und alternative Optimierer Im Projekt SOMA wurde auf eine möglichst effiziente Optimierung der vorliegenden Modellparameter geachtet. Es zählte zu den wesentlichen Zielsetzungen, dass eine Optimierung der Parameter für kleine und mittlere Unternehmen gewährleistet werden kann. Diese Unternehmen verfügen i.d.R. nicht über eine Hochleistungsrechenarchitektur und verzichten schon alleine aus diesem Grund oftmals auf eine systematische Optimierung. Im Projekt SOMA wurden aktuelle Optimierverfahren eingesetzt, die es ermöglichen Parameter in wenigen Schritten zu optimieren.

In der Sprache *R* liegen bereits verschiedene Implementierungen von Optimierverfahren vor. Dazu zählen die Verfahren BFGS, LHS, CMA-ES und SPO. Nichtsdestotrotz wurde im Projekt SOMA eine Anbindung an ein weiteres Verfahren geschaffen und in das Framework TDMR integriert. Dabei handelt es sich um das in Java implementierte Verfahren CMA-ES, welches zwar bereits in einer anderen freien Implementierung in R vorlag, welche allerdings in einigen Testläufen unabsehbares Verhalten zeigte. Aus diesen Gründen wurde eine Anbindung an die originale Java-Software von Hansen [11] geschaffen, sodass nun eine stabile, und korrekte Implementierung dieses wichtigen Verfahrens vorliegt.

Das modellbasierte Optimierverfahren Sequential Parameter Optimization Toolbox (SPOT) ist um einige wesentliche Bestandteile erweitert worden. Dazu zählen insbesondere die Weiterentwicklung der verfügbaren Meta-Modelle, als auch die Integration neuer Funktionen in dem Paket selbst. Für verrauschte Optimierungsaufgaben wurde z.B. die Anbindung an Forrester Re-interpolierungsmethode erzeugt. Weiterhin sind insbesondere die Erweiterungen für parallele Umgebungen zu nennen, mit denen es ermöglicht wurde, Tuning-Läufe auf parallelen Architekturen durchzuführen. Hierbei war es notwendig, Probleme wie mehrfacher Datei-Zugriff zu verhindern, ohne dass ein Abstürzen der Software verursacht wird. Weiterhin sind verschiedene Bericht-Funktionen in die Software integriert worden, so gibt es z.B. die Möglichkeit automatisiert eine Sensitivitätsanalyse zu erzeugen (spotReportSens, s. Abb. 3). Mithilfe der Sensitivitätsanalyse ist es möglich, direkt den Einfluss der Einstellungen der Modellparameter auf die Prognosequalität abzulesen. In Abb. 3 deutet sich z.B. für hohe Einstellungen des Parameters “XPERC” ein verschlechterter Einfluss auf das Klassifikationsergebnis ab, während zu niedrige Einstellungen allerdings auch negativ zu bewerten sind. Somit kann die Sensitivität der jeweiligen Modellparameter direkt aus dem Bericht abgelesen werden, was dem Anwender eine Einstellung und Auswahl der Modellparameter sehr stark erleichtert.

Hinsichtlich der Optimierverfahren wurde in mehreren Studien gezeigt [6], [16], [19], [23], [24], dass mit SPOT sehr robuste Ergebnisse erzielt werden können.

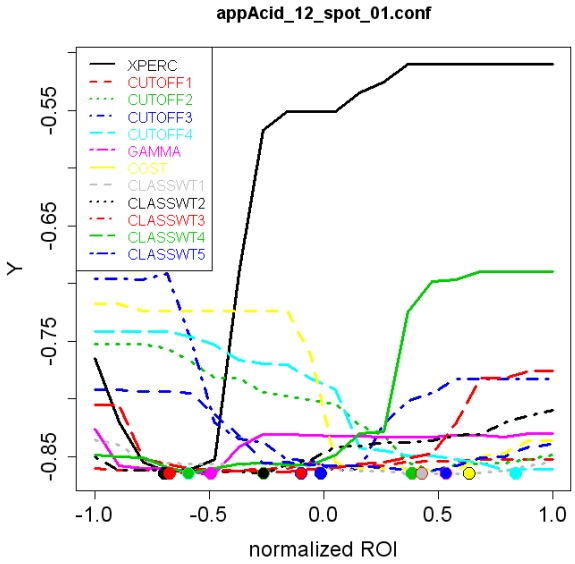

Abbildung 3: SPOT Report Sens: Sensitivitätsanalyse der Parameter.

Trotz des in Data Mining Anwendungen üblichen hohen Rauschens konnten in der Praxis nutzbare Modellparameter ermittelt werden. Dabei lag ein besonderer Fokus auf der effizienten Optimierung mit einem stark limitierten Budget (definiert durch die Anzahl an Funktionsauswertungen, bzw. der Rechenzeit der Optimierung). Hier erwiesen sich Kriging-Modelle zur Parameteroptimierung als hilfreich, da sie auch komplexe Fitness-Landschaften gut abbilden können. Kriging, auch bekannt als Gaussian Processes [26] ist in der rechnergestützten Optimierung erstmalig unter dem von Sacks u.a. [27] propagierten *Design and Analysis of Computer Experiments* untersucht worden und wurde später von Jones u.a. [13] unter dem Paradigma *Efficient Global Optimization* erfolgreich um das sog. *Expected Improvement* Infill-Kriterium erweitert. Das Expected Improvement berücksichtigt sowohl die Ergebnisse der Parametereinstellungen, als auch die Modellunsicherheit (vgl. Abb. 4). Damit ist es möglich eine globale Optimierung des Suchraumes mit automatischer Explorations-Balance durchzuführen.

Ein besonderes Augenmerk lag auch auf der Verkürzung der Optimierungszeiten. Obwohl mit SPOT und Kriging bereits die Anzahl an Auswertungen drastisch reduziert werden konnten, nimmt den Hauptanteil der Rechenzeit noch die Modellbildung während der Optimierung in Anspruch. Daher untersuchten Koch und Konen [17], [18] die Auswirkung einer reduzierten Trainingsmenge während der Optimierung auf die gefundenen Modellparameter. Hierbei zeigte sich, dass bereits mit geringen Teilmengen der zur Verfügung stehenden Trainingsdaten gute Parameter-Einstellungen gefunden werden können, unter einer wesentlich verbesserten Rechenzeit. Die Verbesserung der Rechenzeit ist dabei abhängig vom verwendeten Prognosemodell.

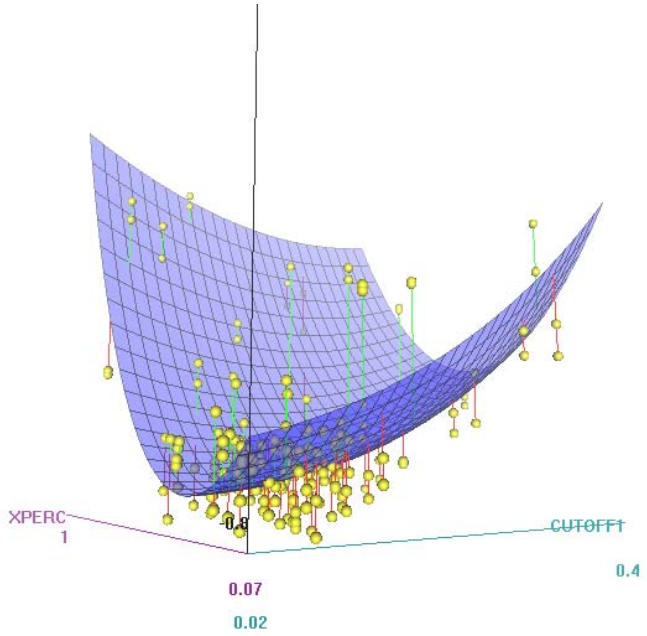

Abbildung 4: Geglätteter Fit der approximierten Funktion unter Berücksichtigung der Modellunsicherheiten.

Feature Entwurfsalgorithmen Es sind verschiedene Verfahren zur Feature Extraktion und zum Feature Entwurf untersucht und getestet worden. Für die Feature Extraktion wurden verschiedene Verfahren zur Selektion von Merkmalen aus dem vorhandenen Datensatz analysiert. Dazu zählen unter anderem:

- Feature Vorwärtsselektion / Feature Rückwärtsselektion
- Filter-Verfahren (z.B. Korrelationsfilter, Entropie, etc.)
- Random Forest Importance
- Genetischer Algorithmen

Eine Übersicht und Taxonomie zu diesen Verfahren geben Guyon und Ellisseeff [9]. Obwohl gezeigt werden kann, dass z.B. mit Genetischen Algorithmen eine optimale Feature-Menge ermittelt werden kann, haben diese Verfahren den Nachteil, dass sie sehr viele Modell-Auswertungen erfordern. Aus Effizienz-Gründen ist daher von diesen Verfahren abzuraten. Einfache Filter hingegen bestimmen sehr schnell eine reduzierte Feature-Menge, haben allerdings den Nachteil, dass sie zum Einen oftmals nicht mit diskreten Merkmalen umgehen können, zum Anderen durch einige wenige verrauschte Features beeinflusst werden.

Eine bessere Alternative ist hier der Random Forest Importance Selektor (vgl. Abb. 5), mit dem es möglich ist anhand der Modellvorhersage eine gute Auswahl an Features zu treffen. Direkter Nachteil des Verfahrens ist die bedingte Abhängigkeit von dem Random Forest Prognosemodell. Es hat sich jedoch in

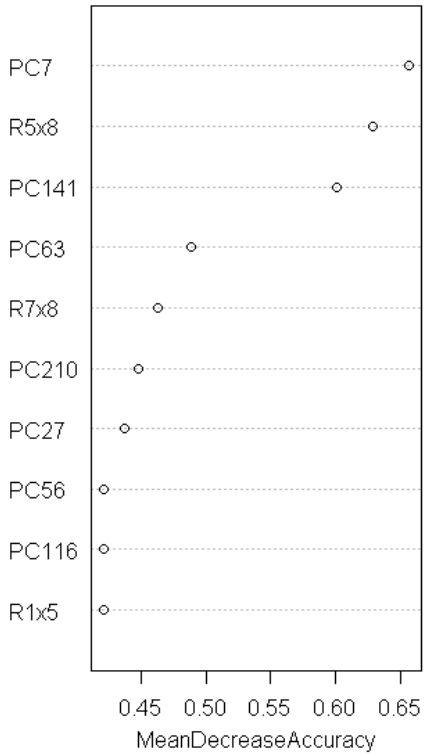

Abbildung 5: Beispiel für die Random Forest Importance. Die Features werden je nach Bedeutung eingeordnet (je weiter rechts, desto wichtiger ist das Merkmal). Die Bedeutung der Features für das Klassifikationsergebnis wird anhand eines Qualitätskriteriums (hier: MeanDecreaseAccuracy) berechnet, woraufhin durch das Optimierverfahren eine reduzierte Feature-Menge ausgewählt werden kann.

einer experimentellen Studie [24] gezeigt, dass mit diesem Verfahren in der Praxis eine sehr gute Feature Auswahl erzielt werden kann (vgl. Tab. 2).

Tabelle 2: Klassifikationsgenauigkeit (in %) für eine Industrieanwendung (Vorhersage der Säurekonzentration in Biogasanlagen) unter Berücksichtigung verschiedener Feature-Entwurfsalgorithmen. Verglichen werden die Bildung neuer Features über Hauptkomponentenanalyse (PCA) sowie Monome, als auch die Feature-Selektion mittels Random Forest Importance (FS-RFI) und Genetischen Algorithmen (FS-GA). An dem Beispiel zeigt sich, dass die Verwendung der entsprechenden Feature-Entwurfsverfahren kombiniert mit einer Feature Selektion basierend auf dem Random Forest Prognosemodell das beste Ergebnis liefern.

	PCA	Monome	FS-RFI	FS-GA	Klass.genauigkeit
1	X	X	X	-	$(89.95 \pm 0.41)\%$
2	X	X	-	X	$(89.47 \pm 0.52)\%$
3	X	-	X	-	$(86.72 \pm 0.77)\%$
4	-	X	X	-	$(83.38 \pm 0.78)\%$
5	-	-	X	-	$(82.90 \pm 1.35)\%$
6	X	X	-	-	$(82.60 \pm 0.92)\%$
7	-	-	-	-	$(82.59 \pm 0.42)\%$

Zum Feature Entwurf wurden verschiedene Verfahren eingesetzt. Einerseits ist für Zeitreihendaten ein allgemeines Feature-Entwurfsverfahren entwickelt worden, welches es ermöglicht normale Klassifikations- oder Regressionsmodelle für Zeitreihendaten anzuwenden. Koch u.a. [16, 15] untersuchten dieses allgemeine Verfahren für eine Anwendung in der Wasserwirtschaft und erzielten verbesserte Ergebnisse gegenüber einem speziellen Zeitreihen-Prognosemodell. In einer weiteren Studie wurde die Hauptkomponentenanalyse (Principal Component Analysis, PCA) zum Feature-Entwurf eingesetzt. In einer weiteren Prognose-Anwendung konnte mit Hilfe der PCA das bisherige Referenzverfahren sowohl hinsichtlich der Vorhersagegenauigkeit, als auch der benötigten Trainingszeit verbessert werden [24].

Die Anwendung von Slow Feature Analysis (SFA) kann bei als Zeitreihen vorliegenden Daten zu vielversprechenden Feature Sets führen. Dies wurde für das Anwendungsbeispiel der Gestenerkennung gezeigt [20]. SFA war dabei in der Lage die vorherzusagenden Gesten teilweise genauer zu prognostizieren als das bekannte Referenzverfahren Random Forest. Dabei erfordert SFA nur einen Bruchteil an Speicheraufwand und Rechenzeit um eine Prognose durchzuführen und ist daher insbesondere für die Anwendung auf mobilen Geräten interessant. Das bestehende SFA Toolkit in Matlab (sfa-tk) wurde für die Klassifikation grundlegend weiterentwickelt. Im Rahmen einer durchgeführten Case Study im Projekt SOMA ist außerdem eine weitere Implementierung von SFA in der Sprache R entstanden, welche im Rahmen der GPL Lizenz auf dem R-Projektserver CRAN

verfügbar gemacht worden ist (`http://cran.r-project.org/web/packages/rSFA/index.html`).

Tabelle 3: Fehlerraten (in Prozent) mit der Slow Feature Analysis für ein Gestenerkennungsproblem (Auswahl aus fünf Gesten: Kreis, Wurf, Frisbee, Bowling, 'z'-Geste). Dick gedruckte Werte sind am Besten. Die Erkennung der Gesten wurde in einer Vergleichsstudie mit dem Random Forest Klassifikator und dem Gauss-Klassifikator verglichen.

Klassifikator	Min.	Durchschn.	Max.	Std.Abw.
SFA	1.68	**2.03**	**2.24**	0.18
Random Forest	**1.54**	2.09	2.37	0.30
Gauss	13.55	14.02	14.39	0.22

In einer mit der Technischen Universität Dortmund behandelten Studie ist untersucht worden, ob es mittels Genetischer Programmierung (GP) möglich ist, verbesserte Kernel-Funktionen für Support Vector Machines zu bestimmen. Als Ergebnis konnten Standard-Kernel-Funktionen mit GP gefunden werden, allerdings konnten keine verbesserten Ergebnisse mit diesem Verfahren erzielt werden. Die Methode GP wird jedoch aktuell in einer sehr aktive Forschungsgemeinschaft behandelt, sodass in Zukunft nach Anpassung der Variationsoperatoren evtl. bessere Lösungen möglich sind. Aus diesem Grund sind die Ergebnisse der Studie in der Fachzeitschrift Evolutionary Intelligence festgehalten worden [15].

Case Studies In dem Projekt SOMA wurden verschiedene Case Studies im Master-Studiengang *Automation & IT* der FH Köln durchgeführt. Durch die internationale Beschaffenheit des Studiengangs, konnten insbesondere gemischte Gruppen mit nationalen und internationalen Studierenden für die Case Studies gewonnen werden.

Die folgende Liste gibt einen kurzen Überblick über die im Projekt SOMA veranstalteten Case Studies und die behandelten Fragestellungen:

WS 2009/2010 Predicting Fill Levels of Stormwater Overflow Tanks. Teilnehmer: Michael Tamutan, Thomas Ludwig, Aldo Sedeño

WS 2010/2011 Predicting Ammonium Concentrations in Wastewater Treatment Plants. Teilnehmer: Velasco Diego, Maxim Shatskiy

WS 2011/2012 Extensions for Tuned Data Mining. Teilnehmer: Martin Zaefferer, Fasika Ayodele, Ashwin Kumar, Prawyn Jebakumar

WS 2012/2013 Building and analyzing SVM ensembles with Bagging and AdaBoost on big data sets. Teilnehmer: Ricardo Ramos Guerra, Jörg Stork

Die durchgeführten Case Studies waren sehr erfolgreich und konnten für die weitere Aufgabenverteilung und Forschung im Projekt SOMA genutzt werden.

Die im WS 2009/2010 behandelten Arbeiten zum Thema "Predicting Fill Levels of Stormwater Overflow Tanks" dienten als Vorarbeit für anschließende Publikationen in diesem Bereich [16, 19]. Die Prognose von Füllständen in

Regenüberlaufbecken ist nur ein erfolgreiches Beispiel für die sehr gute Modellqualität der in dem Forschungsvorhaben SOMA optimierten Prognosemodelle.

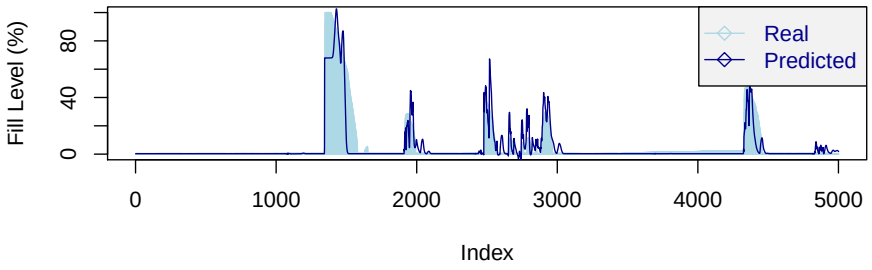

Abbildung 6: Prognose von Füllständen in Regenüberlaufbecken. Die helle blaue Fläche gibt die tatsächlichen Füllstände an, während die dunkle Linie die prognostizierten Füllstände beschreibt. Die Abbildung zeigt ein mittels TDMR und SPOT optimiertes Regressionsmodell auf Basis von Support Vector Regression.

Aus der im WS 2011/2012 durchgeführten Case Study zum Thema "Extensions for Tuned Data Mining" konnten Teile der entwickelten Software zur Ergänzung und Erweiterung des TDMR Frameworks genutzt und übernommen werden. Die in diesem Rahmen erstellte Implementierung der Slow Feature Analysis in der Sprache R wurde zudem als eigenständiges Software-Paket auf dem freien Projekt-Server CRAN zur Verfügung gestellt.

Das in der Case Study *Building and analyzing SVM ensembles with Bagging and AdaBoost on big data sets* behandelte Thema ist ein bedeutendes Anwendungsgebiet, und wurde als Publikation im Rahmen der European Conference on Data Analysis als Konferenzbeitrag eingereicht [30]. Der Votrag fand großen Anklang bei der internationalen Forchergemeinde, sodass auch nach Abschluss des Projektes SOMA weitere Arbeiten angedacht sind.

Bachelorarbeiten Der Student Markus Thill schrieb seine Bachelor-Arbeit zum Thema Reinforcement Learning mit N-Tupel-Systemen für das Brettspiel "Vier Gewinnt". Die erreichten Ergebnisse übertrafen das bisherige Referenzverfahren (Temporal Difference Learning). Damit gelang es in kürzerer Zeit in über 90% der Fälle gegen einen optimal spielenden Agenten zu gewinnen, sofern das Spiel einen Gewinn zulässt. Eine Veröffentlichung in diesem Bereich wurde auf der internationalen Fachkonferenz Parallel Problem Solving from Nature (PPSN) in Taormina, Italien vorgestellt. Markus Thill wurde außerdem für seine überzeugende Arbeit mit dem Opitz-Förderpreis (1. Platz) ausgezeichnet.

Masterarbeiten Die Studentin Kristine Hein untersuchte das Problem der Gestenerkennung mittels der im Projekt SOMA behandelten Methode zur Merkmalsgenerierung *Slow Feature Analysis* [12]. Dabei konnten für den erzeugten Klassifikator vergleichbare Prognoseraten wie für ein auf Random Forests basierendes Modell erzielt werden. Dies kann als ein großer Erfolg gewertet werden, denn das auf SFA basierende Prognosemodell ist in seiner Anwendung deutlich schneller sowohl im Training, als auch in der Anwendung, was insbesondere bei zeitkritischen Anwendung in integrierten Systemen eine Rolle spielen kann. Kristine Hein wurde als Auszeichnung für ihre hervorragende Arbeit mit dem begehrten Opitz-Förderpreis 2011 ausgezeichnet.

Aufbauend auf den Ergebnissen von Frau Hein untersuchte der Student Daniel Bertram die beschleunigungsbasierten 3D- Gestendaten auf einem Smartphone und erzielte erstaunlich stabile Ergebnisse bei der Erkennung auf diesen Geräten [2]. Herr Bertram wurde außerdem für seine Masterarbeit mit dem CBC-Förderpreis 2013 ausgezeichnet.

Die Studentin Renée Schulz führte die erfolgreichen Arbeiten im Bereich Gestenerkennung fort [28] und arbeitete neben dem Interface auf Basis der Nintendo Wii auch mit Microsofts XBox Kinect, welches sensorisch weiterentwickelt ist und komplexere Bewegungserkennungen erlaubt. Frau Schulz wurde mit dem Ferchau-Förderpreis 2013 ausgezeichnet.

Meilensteine

MS1: Abschluss der ersten Case Studies. Abhängig von der Evaluation der Case Studies und der Hinweise auf Erfolg bzw. Misserfolg einzelner Verfahren wird die Schwerpunktsetzung in den darauffolgenden Arbeitspaketen gesteuert.

MS2: Das Framework für den Einsatz verschiedener Modellierungsverfahren in Verbindung mit SPO ist fertiggestellt.

MS3: Abschluss der letzten Case Studies. Abhängig von der Evaluation der Case Studies und der Hinweise auf Erfolg bzw. Misserfolg einzelner Verfahren wird die Schwerpunktsetzung in den darauffolgenden Arbeitspaketen gesteuert.

MS4: Abschluss der methodischen Arbeiten zu Modul E, es liegen erste Anwendungsresultate vor. Abhängig von diesen Resultaten wird mindestens ein geeignetes Referenzprojekt ausgewählt, das in Arbeitspaket C3 (Dokumentation von Referenzprojekten) für die Fachöffentlichkeit aufbereitet wird.

MS5: Abschluss der methodischen Arbeiten zu Modul S, es liegen erste Anwendungsresultate vor. Abhängig von diesen Resultaten wird mindestens ein geeignetes Referenzprojekt ausgewählt, das in Arbeitspaket C3 (Dokumentation von Referenzprojekten) für die Fachöffentlichkeit aufbereitet wird

Die Meilensteine konnten wie geplant eingehalten werden. Die durchgeführten Case Studies waren in ihren Ergebnissen sehr erfolgreich und zielführend zur Aufgabenstellung. So konnten die Ergebnisse der Case Studies zum Teil direkt für weiterführende Publikationen genutzt werden. Beispielsweise wurden die Untersuchungen der im Wintersemester 2012/2013 durchgeführten Case Study im Rahmen einer europäischen Fachkonferenz (European Conference on Data Analysis) vorgestellt und stießen insgesamt auf großes Interesse bei den Forschern.

Aus der im Wintersemester 2011/2012 durchgeführten Case Study entstand außerdem die Open Source Software *rSFA*, welche auch im Rahmen des Data Mining Frameworks TDMR zur Feature-Gewinnung genutzt wird.

Alle weiteren Meilensteine konnten wie geplant abgeschlossen werden. Als wesentlicher Bestandteil des Projektes SOMA ist das Framework TDMR [21] für den Einsatz der verschiedenen Modellierungsverfahren der Öffentlichkeit zugänglich gemacht worden. Auch die Meilensteine M4 und M5 wurden erfüllt und sind durch die zahlreichen Veröffentlichungen im Projekt SOMA der Fachwelt zugänglich gemacht worden.

2.2 Zahlenmäßiger Nachweis

Das Projektbudget wurde entsprechend der im Projektantrag beschriebenen Planung verausgabt. Änderungen dieser Planung waren entweder nicht notwendig, oder wurden dem Projektträger frühzeitig mitgeteilt und genehmigt. Der im Projekt SOMA finanziell verfügbare Rahmen wurde somit planmäßig eingehalten.

Wie im Finanzierungsplan veranschlagt, wurde der wesentliche Anteil des Projektbudgets für das wissenschaftliche Personal verausgabt. Innerhalb der Personalausgaben kam es geringen Verschiebungen innerhalb der Positionen 817 *Beschäftigungsentgelte E1-E11 (wissenschaftliche Hilfskräfte)* zu der Position 812 *Entgeltgruppe E12-E15 (wissenschaftliche Mitarbeiter)*. Mit dieser Verschiebung ist es ermöglicht worden, den Projektmitarbeiter Patrick Koch über den gesamten Projektzeitraum zu beschäftigen.

Insgesamt wurde das Projekt auf Antrag im Mai 2013 mit einer kostenneutralen Übertragung von Mitteln in Höhe von 30.244€ in das Jahr 2013 bis zum 31.06.2013 verlängert. Diese Verlängerung konnte durch Umwidmung von Personalkosten für wissenschaftliche Hilfskräfte erzielt werden. Diese Umwidmung war ein notwendiger Schritt, der vom Projektträger genehmigt worden ist, da die Programmiertätigkeiten, die bis dahin von studentischen und wissenschaftlichen Hilfskräften durchgeführt wurden, zunehmend schwieriger und komplexer geworden sind. Die in SOMA eingearbeiteten Hilfskräfte, die bisher über die Position Mittel für Beschäftigungsentgelte finanziert wurden und nahezu seit Projektbeginn für SOMA arbeiteten, hatten inzwischen anderweitige Tätigkeiten aufgenommen und standen deshalb nicht mehr für SOMA zur Verfügung. Eine Einarbeitung von neuen Hilfskräften in die umfangreiche Programmcodeumgebung von SOMA wäre mit hohen Aufwänden sowohl für die Hilfkräfte als auch für den die Hilfskräfte betreuenden Doktoranden verbunden. In Anbetracht der geringen Restlaufzeit des Projekts war demzufolge das Verhältnis von Aufwand zu Ertrag bei der Beschäftigung neuer Hilfskräfte sehr ungünstig. Stattdessen wurden die geplanten Arbeiten vom Doktoranden selbst ausgeführt.

Auf Seiten der Investitionen (Position 850) wurde gegenüber dem Projektantrag eine Änderung vorgenommen, die dem Projektträger aber ebenfalls schnellstmöglich mitgeteilt worden ist. Zur Verbesserung der Lehre in den Case Studies und der Präsentation auf Fachtagungen und Konferenzen wurde in dem Projekt ein mobiles Gerät angeschafft. Dazu wurden insgesamt 1589€ aus Position 812 auf Position 850 umgewidmet. Die Genehmigung dieser Umwidmung wurde vom Projektträger im Mai schriftlich mitgeteilt.

2.3 Nutzen und Verwertbarkeit

Die in diesem Forschungsvorhaben untersuchten Methoden umfassen überwachte Lernverfahren, Verfahren zur Merkmalsextraktion und Merkmalsgewinnung sowie die Kombination dieser Verfahren mit effizienten Optimieralgorithmen. Aufgrund der einfachen vorgefertigten Struktur innerhalb der Software TDMR besteht damit insbesondere für kleine und mittlere Unternehmen (KMUs) ein großes Nutzenpotential. In dem Projekt SOMA wurde darauf Wert gelegt, dass keine tiefergehende Einarbeitung in die Thematik notwendig ist, um eine erste Prognoseaufgabe durchzuführen und die unterliegenden Modellparameter integrativ zu optimieren. Hier ist außerdem wichtig, dass keine besondere Hochleistungsrechenarchitektur für derartige Vorhaben notwendig ist, denn durch die effiziente Implementierung und Kupplung mit Sub-Sampling basierten Ansätzen, werden für die meisten anfallenden Aufgaben in KMUs keine Hochleistungsrechner benötigt.

Die in dem Forschungsvorhaben entwickelten Softwarebibliotheken können gerne von jedem Interessenten genutzt werden und sind bereits in Form von Open-Source Software der Allgemeinheit zugänglich gemacht worden.

2.4 Fortschritt anderer Stellen

Dem Zuwendungsempfänger (ZE) sind weitere in ähnlichen Bereichen forschenden Stellen bekannt. Dieses Wissen konnte genutzt werden, um das eigene Vorhaben zu unterstützen und zu erweitern.

Mitarbeiter vom Lehrstuhl von Prof. Claus Weihs an der TU Dortmund führten ebenfalls Untersuchungen im Bereich der Optimierung unter limitierten Budgets mit Surrogat-Modellen durch. Diese Untersuchungen waren dem ZE frühzeitig bekannt. Es bestehen jedoch wesentliche Unterschiede hinsichtlich der Zielsetzung der unterliegenden Anwendungen. Anhand der vorliegenden Ergebnisse konnten Vergleiche mit den Studien der TU Dortmund unternommen werden. Insgesamt wurden ähnliche Ergebnisse wie in dem Projekt SOMA beobachtet. Die Ähnlichkeit der Forschungsrichtungen konnten für einen Ausbau der Kooperation mit der TU Dortmund genutzt werden, ohne dass die eigenen Ziele des Projektes bzw. der Verwertbarkeit gefährdet waren oder sind. Die Daten sind abgeglichen worden und in die Literatur hinzugefügt worden, um einen vollständigen Blick auf die untersuchten Bereiche zu gewährleisten.

Der an der Ruhr-Universität Bochum forschende Prof. Dr. Laurenz Wiskott gilt als Mitentwickler der in dem Projekt SOMA eingesetzten Slow Feature Analysis. Durch diesen bestehenden Kontakt zu der Ruhr-Universität Bochum konnten Teile der bestehenden Software genutzt und weiter entwickelt werden.

An der FH Düsseldorf fanden einige Forschungsarbeiten im Bereich Deep Learning und Deep Neural Networks unter der Leitung von Prof. Dr. Thomas Zielke statt. Mit dem Lehrstuhl besteht ein guter Kontakt, sodass vergleichende Arbeiten möglich waren und durchgeführt wurden.

2.5 Publikationen im Projekt

Insgesamt sind aus dem Projekt SOMA zahlreiche Publikationen sowohl auf nationaler, als auch auf internationaler Ebene hervorgegangen. Neben der Veröffentlichung

der Artikel in der wissenschaftlichen Gemeinschaft konnte somit auch die Sichtbarkeit der Arbeitsgruppe und der Fachhochschule Köln international gestärkt werden. Wie aus nachfolgender Aufstellung ersichtlich, handelt es sich um insgesamt 10 technische Berichte, 5 Workshop-Beiträge, 10 Konferenzbeiträge und 3 Artikel in Fachzeitschriften.

E-prints und technische Berichte

- Guerra, R. R. and Stork, J.: Building and analyzing SVM ensembles with Bagging and AdaBoost on big data sets, Case Study Report, Cologne University of Applied Sciences, CIOP Technical Report 1/13, 2013.
- Konen, W. and Koch, P.: The TDMR Framework: Tuned Data Mining in R, Cologne University of Applied Sciences, CIOP Technical Report 02/12, 2012.
- Konen, W. and Koch, P.: The TDMR Tutorial: Examples for Tuned Data Mining in R, Cologne University of Applied Sciences, CIOP Technical Report 03/12, 2012.
- Thill, M.: Einsatz von N-Tupel-Systemen mit TD-Learning für strategische Brettspiele am Beispiel von Vier Gewinnt, Praxisprojektbericht, Cologne University of Applied Sciences, CIOP Technical Report 01/12, 2012.
- Konen, W.: SFA classification with few training data: Improvements with parametric bootstrap, Cologne University of Applied Sciences, CIOP Technical Report 09/11, 2011.
- Konen, W.: Der SFA-Algorithmus für Klassifikation, Cologne University of Applied Sciences, CIOP Technical Report 08/11, 2011.
- Hein, K.: Lernende Klassifikation beschleunigungsbasierter 3D-Gesten des Wii-Controllers, Cologne University of Applied Sciences, CIOP Technical Report 01/10, 2010.
- Flasch, O., Bartz-Beielstein, T., Davtyan, A., Koch, P., Konen, W., Oyetoyan, T.D., Tamutan, M.: Comparing CI Methods for Prediction Models in Environmental Engineering, Technical Report Cologne University of Applied Sciences, Germany, 2010
- Konen, W.: On the numeric stability of the SFA implementation sfa-tk. e-print published at `http://arxiv.org/abs/0912.1064`, 2009.
- Konen, W. and Koch, P.: How slow is slow? SFA detects signals that are slower than the driving force. e-print published at `http://arxiv.org/abs/0911.4397`, 2009.

Workshop-Beiträge

- Koch, P. and Konen, W.: Subsampling strategies in SVM ensembles. In: Hoffmann, F., Hüllermeier, E. (Eds.): Proceedings 23. Workshop Computational Intelligence, Dortmund. Universitätsverlag Karlsruhe, 2013.
- Konen, W.: Self-configuration from a Machine-Learning Perspective. e-print published at `http://arxiv.org/abs/1105.1951` and Dagstuhl Preprint Archive, Workshop 11181 Organic Computing – Design of Self-Organizing Systems, 2011.

- Konen, W., Koch, P., Flasch, O. and Bartz-Beielstein, T.: Parameter-Tuned Data Mining: A General Framework. In: F. Hoffmann, E. Hüllermeier (eds.), Proceedings 20. Workshop Computational Intelligence, Dortmund. Universitätsverlag Karlsruhe, 2010.
- Koch, P., Flasch, O., Konen, W. and Bartz-Beielstein, T. (2010): Optimization of Support Vector Regression Models for Stormwater Prediction. In: Hoffmann, F. and Hüllermeier, E. (ed.): Proceedings 20. Workshop Computational Intelligence, Dortmund. Universitätsverlag Karlsruhe, 2010.
- Flasch, O., Bartz-Beielstein, T., Koch, P., Konen, W.: Genetic Programming Applied to Predictive Control in Environmental Engineering. In: F. Hoffmann, E. Hüllermeier (eds.), Proceedings 19. Workshop Computational Intelligence, Dortmund. Universitätsverlag Karlsruhe, 2009.

Konferenzbeiträge

- Stork, J., Ramos, R. R., Koch, P. and Konen, W: SVM ensembles are better when different kernel types are combined. Proceedings of the European Conference on Data Analysis (ECDA), Luxembourg, p. 1–10, 2014. Submitted.
- Thill, M., Koch, P. and Konen, W.: Reinforcement learning with n-tuples on the game Connect-4. In: C. Coello Coello, V. Cutello et al. (eds.), PPSN'2012: 12th International Conference on Parallel Problem Solving From Nature, Taormina, Springer, pages 195–204, 2012.
- Koch, P. and Konen, W.: Efficient sampling and handling of variance in tuning data mining models. In: C. Coello Coello, V. Cutello et al. (eds.), PPSN'2012: 12th International Conference on Parallel Problem Solving From Nature, Taormina, Springer, pages 184–194, 2012.
- Konen, W., Koch, P., Flasch, O., Bartz-Beielstein, T., Friese, M. and Naujoks, B.: Tuned Data Mining: A Benchmark Study on Different Tuners, Proc. GECCO 2011, Dublin, July 2011.
- Bartz-Beielstein, T., Friese, M., Zaefferer, M., Naujoks, B., Flasch, O., Konen, W. and Koch, P.: Noisy optimization with sequential parameter optimization and optimal computational budget allocation In Proceedings of Genetic and Evolutionary Computation Conference, pages 119–120, 2011.
- Koch, P., Konen, W. and Hein, K., Gesture Recognition on Few Training Data using Slow Feature Analysis and Parametric Bootstrap. In P. Sobrevilla (ed.), Proc. IEEE World Congress on Computational Intelligence (WCCI), Barcelona, 2010.
- Koch, P., Konen, W., Flasch, O., Bartz-Beielstein, T.: Optimizing Support Vector Machines for Stormwater Prediction. In: R. Schaefer (ed.), Proc. 11th International Conference on Parallel Problem Solving From Nature (PPSN), Krakow, 2010.
- Flasch, O., Bartz-Beielstein, T., Davtyan, A., Koch, P. and Konen, W.: Comparing SPO-tuned GP and NARX Prediction Models for Stormwater Tank Fill Level Prediction. In P. Sobrevilla (ed.), Proc. IEEE World Congress on Computational Intelligence (WCCI), Barcelona, 2010.
- Ziegenhirt, J., Bartz-Beielstein, T., Flasch, O., Konen, W. and Zaefferer, M.: Optimization of Biogas Production with Computational Intelligence – A Comparative Study. In P. Sobrevilla (ed.), In Proceedings of the IEEE World Congress on Computational Intelligence (WCCI), Barcelona, 2010.

– Konen, W. and Koch, P.: How slow is slow? SFA detects signals that are slower than the driving force, In: B. Filipic, J. Silc (eds.), Proc. 4th Int. Conf. on Bioinspired Optimization Methods and their Applications, In Proceedings of the Conference on Bioinspired Optimization Methods and their Applications (BIOMA) 2010, Ljubljana, Slovenia, 2010.

Zeitschriftenbeiträge

– Koch, P., Wagner, T., Emmerich, M. T. M., Baeck, T., and Konen, W.: Efficient multi-criteria optimization on noisy machine learning problems. Applied Soft Computing, 2014, (under review).
– Koch, P., Bischl, B., Flasch, O., Bartz-Beielstein, T., Weihs, C. and Konen, W.: Tuning and evolution of support vector kernels. Evolutionary Intelligence, 5(3):153-170, 2012.
– Konen, W. and Koch, P., The slowness principle: SFA can detect different slow components in nonstationary time series. In: Jurij Šilc and Bogdan Filipič (eds.) International Journal of Innovative Computing and Applications (IJICA), 2010.

Bachelor-Arbeiten

– Thill, M.: Reinforcement Learning mit N-Tupel-Systemen für Vier Gewinnt, Bachelor Thesis, Fachhochschule Köln, 2012. Preisträger (1. Platz) beim Opitz-Innovationspreis 2013.

Master-Arbeiten

– Schulz, R.: Entwicklung und Vergleich von Verfahren zur Verbesserung der Gestenerkennung für den Einsatz in Natural User Interfaces. Master Thesis, Fachhochschule Köln, 2013. Preisträgerin beim Festo-Förderpreis 2013.
– Bertram, D.: Untersuchungen zur Varianzreduktion beschleunigungsbasierter 3D-Gestendaten, Master Thesis, Fachhochschule Köln, 2012. Preisträger (3. Platz) beim CBC-Förderpreis 2013.
– Hein, K: Gestenerkennung mit Slow Feature Analysis (SFA) – Klassifizierung von beschleunigungsbasierten 3D-Gesten des Wii-Controllers, Fachhochschule Köln, 2010. Preisträgerin (3. Platz) beim Opitz-Innovationspreis 2011.

Dissertationen

– Koch, P.: Efficient Tuning in Supervised Machine Learning. PhD Thesis, Universität Leiden, Niederlande, 2013.

Literaturverzeichnis

[1] T. Bartz-Beielstein, C.W.G. Lasarczyk, and M. Preuß. Sequential parameter optimization. In *IEEE Congress on Evolutionary Computation*, volume 1, pages 773–780. IEEE, 2005.

[2] Daniel Bertram. Untersuchungen zur varianzreduktion beschleunigungsbasierter 3d-gestendaten. Master thesis, Faculty of Computer Science and Engineering Science, Cologne University of Applied Science, Apr 2012.

[3] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Charles G Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593, 1965.

[5] C. Cortes and V. Vapnik. Support Vector Machine. *Machine Learning*, 20(3):273–297, 1995.

[6] O. Flasch, T. Bartz-Beielstein, A. Davtyan, P. Koch, and W. Konen. Comparing SPO-tuned GP and NARX prediction models for stormwater tank fill level prediction. In P. Sobrevilla, editor, *In Proceedings of the IEEE World Congress on Computational Intelligence (WCCI)*, 2010.

[7] R. Fletcher and M.J.D. Powell. A rapidly convergent descent method for minimization. *The Computer Journal*, 6(2):163–168, 1963.

[8] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[10] N. Hansen and A. Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of IEEE International Conference on Evolutionary Computation*, pages 312–317. IEEE, 1996.

[11] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[12] Kristine Hein. Lernende klassifikation beschleunigungsbasierter 3d-gesten des wii-controllers. CIOP Technical Report 01-10, Research Center CIOP (Computational Intelligence, Optimization and Data Mining), Cologne University of Applied Science, Faculty of Computer Science and Engineering Science, Jan 2010.

[13] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[14] P. Koch. *Efficient Tuning in Supervised Machine Learning*. PhD thesis, Universiteit Leiden, 2013.

[15] P. Koch, B. Bischl, O. Flasch, T. Bartz-Beielstein, C. Weihs, and W. Konen. Tuning and evolution of support vector kernels. *Evolutionary Intelligence*, 5(3):153–170, 2012.

[16] P. Koch, O. Flasch, W. Konen, and T. Bartz-Beielstein. Optimization of Support Vector Regression Models for Stormwater Prediction. In F. Hoffmann and E. Hüllermeier, editors, *In Proceedings of the 20th Workshop on Computational Intelligence*. Universitätsverlag Karlsruhe, 2010.

[17] P. Koch and W. Konen. Efficient sampling and handling of variance in tuning data mining models. In C. Coello Coello and V. Cutello, editors, *In Proceedings of the 12th International Conference on Parallel Problem Solving From Nature*, pages 184–194. Springer, 2012.

[18] P. Koch and W. Konen. Subsampling strategies in svm ensembles. In F. Hoffmann and E. Hüllermeier, editors, *Proceedings 23. Workshop Computational Intelligence*. Universitätsverlag Karlsruhe, 2013.

[19] P. Koch, W. Konen, O. Flasch, and T. Bartz-Beielstein. Optimizing Support Vector Machines for stormwater prediction. In T. Bartz-Beielstein and M. Chiarandini, editors, *In Procedings of the Workshop on Experimental Methods for the Assessment of Computational Systems, held in conjunction with the PPSN 2010*, 2010.

[20] P. Koch, W. Konen, and K. Hein. Gesture recognition on few training data using Slow Feature Analysis and parametric bootstrap. In P. Sobrevilla, editor, *In Proceedings of the IEEE World Congress on Computational Intelligence (WCCI)*, 2010.

[21] W. Konen and P. Koch. The tdmr framework: Tuned data mining in r. Technical Report CIOP Technical Report 02/12, Cologne University of Applied Sciences, 2012.

[22] W. Konen and P. Koch. The TDMR Tutorial: Examples for Tuned Data Mining in R. Technical Report CIOP Technical Report 03/12, Cologne University of Applied Sciences, 2012.

[23] W. Konen, P. Koch, O. Flasch, and T. Bartz-Beielstein. Parameter-Tuned Data Mining: A General Framework. In F. Hoffmann and E. Hüllermeier, editors, *In Proceedings of the 20th Workshop on Computational Intelligence*. Universitätsverlag Karlsruhe, 2010.

[24] W. Konen, P. Koch, O. Flasch, T. Bartz-Beielstein, M. Friese, and B. Naujoks. Tuned Data Mining: A Benchmark Study on Different Tuners. In *In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, 2011.

[25] R. Poli, W.W.B. Langdon, N.F. McPhee, and J.R. Koza. *A Field Guide to Genetic Programming*. http://lulu.com, 2008.

[26] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[27] J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[28] Renée Schulz. Entwicklung und vergleich von verfahren zur verbesserung der gestenerkennung für den einsatz in natural user interfaces. Master thesis, Faculty of Computer Science and Engineering Science, Cologne University of Applied Science, Aug 2013.

[29] D.F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[30] J. Stork, R. Ramos, P. Koch, and W. Konen. SVM ensembles are better when different kernel types are combined. In B. Lausen, editor, *European Conference on Data Analysis (ECDA)*. (to appear), 2013.

[31] R. Storn and K. Price. Differential Evolution – A Simple and Efficient Heuristic for Global Optimization Over Continuous Spaces. *Journal of Global Optimization*, 11(4):341–359, 1997.

[32] L. Wiskott and T.J. Sejnowski. Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Computation*, 14(4):715–770, 2002.

Kontakt/Impressum

Diese Veröffentlichungen erscheinen im Rahmen der Schriftenreihe "CIplus". Alle Veröffentlichungen dieser Reihe können unter
www.ciplus-research.de
oder unter
http://opus.bsz-bw.de/fhk/index.php?la=de
abgerufen werden.

Köln, Januar 2012

Herausgeber / Editorship

Prof. Dr. Thomas Bartz-Beielstein,
Prof. Dr. Wolfgang Konen,
Prof. Dr. Horst Stenzel,
Dr. Boris Naujoks
Institute of Computer Science,
Faculty of Computer Science and Engineering Science,
Cologne University of Applied Sciences,
Steinmüllerallee 1,
51643 Gummersbach
url: www.ciplus-research.de

Schriftleitung und Ansprechpartner/ Contact editor's office

Prof. Dr. Thomas Bartz-Beielstein,
Institute of Computer Science,
Faculty of Computer Science and Engineering Science,
Cologne University of Applied Sciences,
Steinmüllerallee 1, 51643 Gummersbach
phone: +49 2261 8196 6391
url: http://www.gm.fh-koeln.de/~bartz/
eMail: thomas.bartz-beielstein@fh-koeln.de

ISSN (online) 2194-2870