

eXamen.press

eXamen.press ist eine Reihe, die Theorie und Praxis aus allen Bereichen der Informatik für die Hochschulausbildung vermittelt.

Jürgen Paetz

Soft Computing in der Bioinformatik

Eine grundlegende Einführung und Übersicht

Mit 43 Abbildungen und 5 Tabellen

 Springer

Jürgen Paetz

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen
Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
<http://dnb.ddb.de> abrufbar.

ISSN 1614-5216

ISBN-10 3-540-29886-X Springer Berlin Heidelberg New York

ISBN-13 978-3-540-29886-1 Springer Berlin Heidelberg New York

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland vom 9. September 1965 in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtsgesetzes.

Springer ist ein Unternehmen von Springer Science+Business Media
springer.de

© Springer-Verlag Berlin Heidelberg 2006
Printed in Germany

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürfen. Text und Abbildungen wurden mit größter Sorgfalt erarbeitet. Verlag und Autor können jedoch für eventuell verbliebene fehlerhafte Angaben und deren Folgen weder eine juristische Verantwortung noch irgendeine Haftung übernehmen.

Satz: Druckfertige Daten des Autors
Herstellung: LE- \LaTeX , Jelonek, Schmidt & Vöckler GbR, Leipzig
Umschlaggestaltung: KünkelLopka Werbeagentur, Heidelberg
Gedruckt auf säurefreiem Papier 33/3142 YL - 5 4 3 2 1 0

Allen Lesern – für Forschung, Anwendung und Lehre

Vorwort

Dieses Buch behandelt die Grundlagen des in der Forschung und Praxis bedeutenden Gebiets „Soft Computing“. Der Stoff aus den Themengebieten Datenanalyse, Neuronale Netze, Fuzzy-Logik, Maschinelles Lernen, evolutionäre Strategien und naturanaloge Algorithmen ist für eine bis zu vierstündige Vorlesung mit Übungen konzipiert. Teile einer zweistündigen Vorlesung „Soft Computing zur Wissensextraktion“ und einer dreistündigen Vorlesung „Soft Computing für Bioinformatiker“ wurden für dieses Buch überarbeitet und ergänzt. Das Buch enthält auch eine Einführung in die Matlab-Arbeitsumgebung, die für die Übungen verwendet werden kann. Im Anhang findet der Leser die wichtigsten Begriffe der Wahrscheinlichkeitsrechnung. Besonders interessant in diesem Buch ist die Darstellung der Anwendungsmöglichkeiten der Verfahren des Soft Computing innerhalb der Bioinformatik (Kapitel 1, 5, 8, 11, 18, 25, 31) und die Anwendung biologischer Prinzipien im Rahmen des Soft Computing (Kap. 26-30, 32-34), die selbst wiederum auf Probleme der Bioinformatik angewendet werden können. Auch wurden die Chemieinformatik und einige medizinische Anwendungen berücksichtigt. Speziell zu diesen Anwendungsthemen wurden (neben anderen) über 80 aktuelle Originalarbeiten (davon über 60 aus den Jahren ab 2000) eingearbeitet, so dass ein weiterführendes Studium dieser Arbeiten ermöglicht wird. Aufgrund der Detailfülle der Originalarbeiten, sowohl aus informatorischer als auch aus biologischer Sicht, können viele Arbeiten nur in verkürzter Form wiedergegeben werden. Insgesamt wird dem Leser ein Angebot von über 250 Literaturangaben und über 60 Aktivierungselementen gemacht. Dennoch können viele Themen nur im Sinne des Buches einführend und überblicksartig behandelt werden. Dem interessierten Leser sei das ergänzende Studium ausgewählter Originalliteratur daher empfohlen. In jedem Kapitel werden zu Beginn die Lehrziele genannt und am Ende wird der Lehrstoff kurz repetiert. Zum Abschluss sei allen gedankt, die das Buchprojekt in verschiedenster Art und Weise unterstützt haben.

Obertshausen,
Dezember 2005

Jürgen Paetz

Inhaltsverzeichnis

1	Grundlagen der Datenanalyse – Bioinformatik	
1.1	Datenanalyse	3
1.2	Literaturhinweise	5
1.3	Clusterung, Klassifikation und Regelgenerierung.....	7
2	Übersicht und Einordnung	
2.1	Statistische und intelligente Datenanalyse – Soft Computing	15
2.2	Neuronale Netze	19
2.3	Fuzzy-Technologie	20
2.4	Maschinelles Lernen	21
2.5	Evolutionäre Strategien	21
2.6	Naturalanaloge Algorithmen	23
3	Neuronale Netze: Grundlagen	
3.1	Literaturübersicht	29
3.2	Das Neuronale Netz	31
3.3	Das Neuron	33
3.4	Die Lernregel	35
3.5	Das Perzeptron	37
3.6	Die dritte Generation.....	39
4	Backpropagation	
4.1	Das Backpropagation-Lernen	45
4.2	Probleme beim Backpropagation-Lernen	47
4.3	Varianten und Eigenschaften	49
4.4	Regelgenerierung mit Feedforward-Netzen	50
4.5	Dekompositionelle Regelextraktion	52
5	Backpropagation-Netze in der Bioinformatik	
5.1	Vorbereitung	59
5.2	Strukturvorhersage	62
5.3	Vorhersage von Spaltungstellen – Antivirale Medikamente.....	62
5.4	Drug Design	63
5.5	Weitere Anwendungen	65
6	RBF-Netze und Varianten	
6.1	RBF-Netz nach Poggio und Girosi.....	71
6.2	Erste und zweite Variante – Regularisierung und Glättung	73

6.3	Dritte und vierte Variante – Generalisiertes RBF-Netz und RBF-MD-Netz	74
6.4	Fünfte Variante – PNN, RCE- und DDA-Netze	75
6.5	Sechste Variante – Überwachtes Wachsendes Neuronales Gas	80
6.6	Siebte Variante – Elliptische Basisfunktionen	81
6.7	Anwendungen – SVM	81
7	Regelgenerierung aus RBF-Netzen	
7.1	Naive Idee der Regelgenerierung	85
7.2	RecBF-DDA-Netz	86
8	RBF-Netze in der Bioinformatik	
8.1	Klassifikation von Genomsignaturen	91
8.2	Proteinklassifikation	92
8.3	Klassifikation von Phytoplanktonarten	92
8.4	Vorhersage des Siede- und Flammpunkts	93
8.5	Vorhersage der Retention einer Flüssig-Chromatographie	93
8.6	Bio-Basisfunktionen	94
9	Kohonen-Netze und Varianten	
9.1	Selbstorganisierende Karten	99
9.2	Lernende Vektorquantisierung	101
9.3	U-Matrix-Methode	101
9.4	Wachsende Zellstrukturen	103
9.5	(Wachsende) Neuronale Gase	104
9.6	Nicht-stationäre Erweiterungen	105
10	Regelgenerierung mit Kohonen-Netzen	
10.1	Sig*-Methode	109
10.2	Diskussion	111
11	Kohonen-Netze in der Bioinformatik	
11.1	Clusterung von Aminosäuren	115
11.2	Anwendungen im Drug Design	117
11.3	SOM und U-Matrix-Methode zur Genexpressionsanalyse	118
11.4	Projektion eines 3D-Protein-Modells mittels SOM	118
11.5	Abbau von PCB	119
11.6	Weitere Anwendungen	119
12	Fuzzy-Mengenlehre	
12.1	Literaturübersicht	123

12.2	Fuzzy-Mengen	124
12.3	Fuzzy-Mengenoperationen	128
12.4	Fuzzy-Relationen	129
13	Das Extensionsprinzip und Fuzzy-Zahlen	
13.1	Das Extensionsprinzip	135
13.2	Fuzzy-Zahlen	135
14	Fuzzy-Datenanalyse	
14.1	Fuzzy-Clustering	139
14.2	Fuzzy-Klassifikation	140
14.3	Weitere Verfahren	141
15	Fuzzy-Logik	
15.1	Fuzzy-Inferenz	145
15.2	Linguistische Variablen.....	146
16	Fuzzy-Systeme	
16.1	Schema eines Fuzzy-Systems	151
16.2	Regelmodelle.....	152
16.3	Fuzzy-Inferenz	153
16.4	Diskussion	154
17	Neuro-Fuzzy-Systeme	
17.1	Einführung	159
17.2	Das Neuro-Fuzzy-System nach Jang (ANFIS)	162
17.3	Das Neuro-Fuzzy-System nach Carpenter et al. (FUZZY-ARTMAP)	163
17.4	Das Neuro-Fuzzy-System nach Nauck und Kruse (NEF-CLASS)	164
17.5	Das Neuro-Fuzzy-System nach Huber und Berthold (Fuzzy-RecBF-DDA).....	167
18	Fuzzy-Technologie in der Bioinformatik	
18.1	Sekundärstrukturvorhersage.....	175
18.2	Clustering von Daten zur Genexpressionsanalyse mit FUZZY-ART	176
18.3	Genexpressionsanalyse mit FUZZY-ARTMAP	176
18.4	Motiv-Extraktion mit Neuro-Fuzzy-Optimierung	177
18.5	Vererbung von Eigenschaften.....	178
18.6	Fuzzifizierung von Polynukleotiden	179
18.7	Weitere Anwendungen	181

18.8	Virtuelles Screening mit Neuro-Fuzzy-Systemen	181
19	Maschinelles Lernen im Rahmen der KI und der Logik	
19.1	Literaturübersicht	187
19.2	Logik	189
19.3	PROLOG	191
19.4	Expertensysteme	192
19.5	Vorhersage von Faltungsklassen	192
20	Entscheidungsbäume	
20.1	Informationstheorie	198
20.2	Entscheidungsbaum-Lernen	199
21	Assoziationsregeln	
21.1	Assoziation und Generalisierung	205
21.2	Grundbegriffe und Anwendungen	206
21.3	A-priori-Algorithmus	208
21.4	Erweiterungen	209
22	Ausblick auf Zeitreihen	
22.1	Grundideen	215
22.2	Rekonstruktion einer Zeitreihe aus Microarray-Daten.....	217
23	Generalisierungsregeln	
23.1	Versionsraumlernen	221
23.2	AQ- und CN2-Verfahren	223
23.3	Hierarchische Generalisierung	224
23.4	Heuristische Generalisierungsregeln zur Klassifikation	224
24	Weitere Verfahren des Maschinellen Lernens	
24.1	Analoges Schließen – Case Based Reasoning	233
24.2	Rough Set-Theorie	236
25	Maschinelles Lernen in der Bioinformatik	
25.1	Single Nucleotide Polymorphisms	241
25.2	Kombination von Sekundärstrukturvorhersagen	242
25.3	Entscheidungsregeln von Spoligotyping-Daten	243
25.4	Regelerzeugung von Genexpressionsdaten von Hefe	243
25.5	Protein-Protein-Interaktionen	244
25.6	Die Gen-Ontologie	244
25.7	Unterscheidung zwischen Drugs und Non-Drugs	245
25.8	Auffinden relevanter Molekülstrukturen	245

26	Überblick Optimierung, Genetik, Evolution	
26.1	Einführung	249
26.2	Grundlagen	250
27	Evolutionäre Strategien	
27.1	Literaturübersicht	255
27.2	Metropolis-Algorithmus und Simulated Annealing	256
27.3	Evolutionäre Strategien und Anwendungen	258
28	Genetische Algorithmen	
28.1	Mutation, Rekombination und Selektion	265
28.2	Classifier-Systeme	267
29	Genetisches Programmieren	
29.1	Idee des Genetischen Programmierens	271
29.2	Anwendungen	272
30	Formale Aspekte evolutionärer Strategien	
30.1	Theoretische Grundbegriffe	278
30.2	Das Schema-Theorem	279
30.3	Weiterführende Fragestellungen	281
31	Evolutionäre Strategien in der Bioinformatik	
31.1	Krebsvorhersage mit Simulated Annealing	285
31.2	Krebsvorhersage mit Genetischen Algorithmen	286
31.3	Multiples Alignment mit Genetischen Algorithmen	287
31.4	Rekonstruktion von Sequenzen	288
31.5	Optimierung im Drug Design Prozess	288
31.6	Evolutionäre Strategie im Molekularen Docking	290
31.7	Weitere Anwendungen	290
32	Evolutionäre Strategien in Fuzzy-Systemen	
32.1	Ansätze zur Regelloptimierung	295
32.2	Mutations- und Rekombinationsoperatoren	296
33	Naturalogische Algorithmen – Überblick	
33.1	Literaturübersicht	302
33.2	DNA-Computing	303
33.3	Membrane-Computing	304
33.4	Künstliche Immunsysteme	307
33.5	Künstliches Leben	310
33.6	Schwarmalgorithmen	316

34	Ameisen und Ameisenkolonialgorithmen	
34.1	Natürliche Ameisen	321
34.2	Ameisenkolonialgorithmen	322
34.3	Anwendungen	326
35	Ausblick	
A	Wahrscheinlichkeitsrechnung – Übersicht	
B	Einführung in Matlab	
B.1	Das Toolbox-Konzept	341
B.2	Starten von Matlab	342
B.3	Die Hilfe	342
B.4	Der Editor	343
B.5	Die Arbeitsumgebung – Pfade	343
B.6	Funktionen und Skripte	343
B.7	Datenstrukturen	344
B.8	Kontrollstrukturen	345
B.9	Sonstige Funktionen / Demos	346
B.10	Daten laden und abspeichern	347
B.11	Datensammlungen	347
B.12	Entwurf von Experimentierskripten	349
B.13	Die Statistics-Toolbox	351
B.14	Die Neural Network-Toolbox	352
B.15	Graphical User Interface	353
B.16	Die Fuzzy-Toolbox	354
B.17	Die Optimization-Toolbox	354
B.18	Die Genetic Algorithm- and Direct Search-Toolbox	355
B.19	Die Bioinformatics-Toolbox und SimBiology	355
B.20	Der Compiler	356
B.21	Die Database-Toolbox	357
B.22	Die Web Server-Toolbox	357
	Literaturverzeichnis	359
	Sachverzeichnis	381