



Miriam Mathea (Autor)

**Schätzen der
Klassenzugehörigkeitswahrscheinlichkeit zur
Definition des Arbeitsbereichs von
chemieinformatischen Klassifikationsmodellen**

Schätzen der Klassenzugehörigkeitswahrscheinlichkeit
zur Definition des Arbeitsbereichs von
chemieinformatischen Klassifikationsmodellen

Miriam Mathea



Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

<https://cuvillier.de/de/shop/publications/7732>

Copyright:

Cuvillier Verlag, Inhaberin Annette Jentsch-Cuvillier, Nonnenstieg 8, 37075 Göttingen,
Germany

Telefon: +49 (0)551 54724-0, E-Mail: info@cuvillier.de, Website: <https://cuvillier.de>



Inhaltsverzeichnis

Veröffentlichungen der Dissertation	III
Danksagung	V
Abkürzungsverzeichnis.....	VI
Inhaltsverzeichnis	IX
1 Theoretische Grundlagen	1
1.1 Quantitative-Struktur-Wirkungs-Beziehungen.....	1
1.2 Moleküldeskriptoren	2
1.2.1 Einleitung.....	2
1.2.2 Fingerabdruck-Deskriptoren (engl.: Fingerprints).....	2
1.2.3 Topologische Deskriptoren	3
1.3 Einführung in die Multivariate Datenanalyse	3
1.3.1 Einleitung.....	3
1.3.2 Datenvorbehandlung	5
1.3.3 Parametrische Methoden	5
1.3.4 Nichtparametrische Methoden	6
1.3.5 Das Dilemma zwischen Vorhersagegenauigkeit und Interpretierbarkeit.....	6
1.4 Regression.....	7
1.4.1 Einfache Lineare Regression.....	7
1.4.2 Modellvalidierung	7
1.4.3 Multiple Lineare Regression	12
1.4.4 Shrinkage-Methoden.....	14
1.4.5 Dimensions-Reduktions-Methoden.....	16
1.5 Klassifikation.....	20
1.5.1 Einleitung.....	20



1.5.2	Beurteilung der Klassifikationsgüte des Modells.....	20
1.5.3	k-Nächste Nachbarn (engl.: k-Nearest Neighbor (KNN)).....	22
1.5.4	Random Forests (RF)	25
1.5.5	Support Vector Machines (SVM).....	27
1.5.6	Neuronale Netze (engl.: Neural Networks (NN))	29
1.5.7	Bayes-Klassifikator (engl.: Naive Bayesian Classifier (NBC))	30
1.5.8	Lineare Diskriminanz Analyse (engl.: Linear Discriminant Analysis (LDA))	31
1.5.9	Partial Linear Discriminant Analysis (PLSDA).....	32
1.5.10	Ensemble Methoden	33
1.6	Arbeitsbereich (AB) (engl.: Applicability Domain)	34
1.6.1	Einleitung.....	34
1.7	Kalibrierung von Wahrscheinlichkeitsschätzern.....	37
1.7.1	Einleitung.....	37
1.7.2	Kalibriermethoden.....	38
1.7.3	Zuverlässigkeits-Diagramme	41
1.8	Conformal Prediction (CP)	42
2	Zielsetzung der Arbeit.....	46
2.1	Einleitung	46
2.2	Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern	47
2.3	Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeits- schätzern versus CP	48
3	Methoden	50
3.1	Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern	50
3.1.1	Übersicht über die verwendeten Klassifikations- und Regressionstechniken sowie deren Hyperparametereinstellungen	50
3.1.2	Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer.....	51
3.1.3	Auswertung von Zuverlässigkeits-Diagrammen	54



3.1.4	Bewertung der Exaktheit von Klassenzugehörigkeits-Wahrscheinlichkeits- schätzern.....	54
3.1.5	Modellvalidierung	57
3.1.6	Datensätze und molekulare Deskriptoren.....	57
3.1.7	Simulationsaufbau.....	58
3.1.8	Festlegung einer Zuverlässigkeitsgrenze.....	59
3.2	Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeits- schätzern versus CP	65
3.2.1	Übersicht über die verwendeten Klassifikationstechniken sowie deren Hyperparametereinstellungen	65
3.2.2	Modellvalidierung	65
3.2.3	Datensätze und molekulare Deskriptoren.....	66
3.2.4	Einhaltung des Signifikanzlevels mit dem R package „conformal“	66
3.2.5	Einhaltung des Signifikanzlevels mit Klassenzugehörigkeits- Wahrscheinlichkeitsschätzern.....	68
4	Ergebnisse	70
4.1	Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern	70
4.1.1	Visuelle Analyse der Zuverlässigkeits-Diagramme und Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken vor und nach Kalibrierung	70
4.1.2	Vorversuch: Einfluss der Variablenanzahl des Datensatzes auf den Fehler sowie Beurteilung der Fehlermaße	78
4.1.3	Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits- Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels Simulationsstudien.....	83
4.1.4	Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits- Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels realer Datensätze	98



4.1.5	Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels Simulationsstudien	103
4.1.6	Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels realer Datensätze.....	107
4.2	Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern versus CP	110
5	Diskussion.....	112
5.1	Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeits-schätzern.....	112
5.1.1	Visuelle Analyse der Zuverlässigkeits-Diagramme und Histogramme von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerten unterschiedlicher Klassifikations-und Regressionstechniken vor und nach Kalibrierung	112
5.1.2	Einfluss der Variablenanzahl des Datensatzes auf den Fehler sowie Beurteilung der Fehlermaße	115
5.1.3	Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels Simulationsstudien und Realdatensätzen.....	118
5.1.4	Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits-Wahrscheinlichkeitsschätzer der betrachteten Klassifikations-und Regressionstechniken mittels Simulationsstudien und Realdatensätzen	123
5.2	Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeitsschätzern versus CP	124
6	Zusammenfassung und Schlussfolgerung	127
7	Ausblick	129
8	References	XII
9	Anhang	XXIII

9.1	Charakterisierung von Klassenzugehörigkeits-Wahrscheinlichkeits- schätzern.....	XXIII
9.1.1	Visuelle Analyse der Zuverlässigkeits-Diagramme und Histogramme von Klassenzugehörigkeits-Wahrscheinlichkeitsschätzwerten unterschiedlicher Klassifikations- und Regressionstechniken vor und nach Kalibrierung	XXIII
9.1.2	Vorversuch: Einfluss der Variablenanzahl des Datensatzes auf den Fehler sowie Beurteilung der Fehlermaße	XXVII
9.1.3	Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits- Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels Simulationsstudien.....	XXXIX
9.1.4	Analyse potentieller Einflussfaktoren der Klassenzugehörigkeits- Wahrscheinlichkeitsschätzer unterschiedlicher Klassifikations- und Regressionstechniken mittels realer Datensätze	CLIX
9.1.5	Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits- Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels Simulationsstudien	CLXX
9.1.6	Analyse des Einflusses von Hetero-Ensembles auf die Klassenzugehörigkeits- Wahrscheinlichkeitsschätzer der betrachteten Klassifikations- und Regressionsmethoden mittels realer Datensätze.....	CLXXVIII
9.1.7	MOE Deskriptoren.....	CLXXIX
9.2	Vergleich: Definition des AB mit Klassenzugehörigkeits-Wahrscheinlichkeits- schätzern versus CP	CLXXXIII