[Anhang] [ESM - Online only]

Anhang I: Modellselektion

Einige Klassifikatoren besitzen spezielle Parameter, die vor einer Nutzung des Verfahrens durch den Anwender vorgegeben werden müssen. Beispielweise muss für den k-NN-Klassifikator die Zahl der zu betrachteten Nachbarobjekte spezifiziert werden. Zur Kalibrierung solcher Parameter wurde ein vollautomatischer Ansatz verfolgt: Auf Basis von Literaturempfehlungen wurde je Parameter eine Kandidatenliste gebräuchlicher Einstellungen erstellt. Anschließend wurden sämtliche Alternativen je Durchlauf und Datensatz empirisch mittels einer fünffachen Kreuzvalidierung (Izenman 2008, S. 121) auf den Trainingsdaten evaluiert und die Einstellung mit höchstem AUC-Wert ausgewählt. Mit dieser Einstellung wurde ein endgültiges Klassifikationsmodell auf den kompletten Trainingsdaten erstellt und zur Vorhersage der Testdaten verwendet. Für einen Klassifikator mit z. B. zwei Verfahrensparametern und jeweils drei alternativen Einstellungen ergeben sich bei dieser Vorgehensweise: (3*3 Parameterkombinationen*5-facher Kreuzvalidierung*10 zufälligen Trainings-/Testmengen)+10 endgültigen Klassifikationsmodellen = 460 zu erstellende und bewertende Klassifikationsmodelle je Datensatz. Hierdurch wird für jeden Klassifikator und jede zufällige gezogene Trainingsstichprobe die bestmögliche Parametrisierung aus den gegebenen Alternativen ausgewählt. Die entsprechenden Kandidatenlisten zeigt Tab. 6.

[Tab06]

Tab. 6 In der Modellselektion untersuchte Parametereinstellungen

Klassifikator (vgl. Tab. 1)	Anzahl freier Parameter	Parameter	Untersuchte Einstellungen
NBayes	0		
LDA	Diese Verfahren bedürfen prinzipiell keiner Parametereinstellung. Aufgrund numerischer Probleme kann das zugrundeliegende Optimierungsproblem aber z. T. bei hochdimensionalen Datensätzen mit korrelierten Merkmalen nicht gelöst werden. Daher wurde eine rückwärtsgerichtete Merkmalsselektionsheuristik eingesetzt.		
QDA			
LogReg			
K-NN	1	Anzahl der nächsten Nachbarn	[1;3;5;7;9]
C4.5	1	Konfidenz der Pruningstrategie zum Zurückschneiden des Entscheidungsbaumes	[0,1; 0,2; 0,25; 0,3]
CART	0		
LSVM	1	Regularisierungskonstante	$2^{[-6, -5, \ldots, 16]*}$
RBFSVM	>=3	Regularisierungskonstante	Für eine nichtlineare Projektion der Daten wurde die radiale Basisfunktion eingesetzt. Diese besitzt einen freien Parameter, sodass insg. zwei Einstellungen zu treffen waren. Für die simultane Optimierung beider Parameter wurde ein heuristischer Suchalgorithmus implementiert. Dieser hat im Durchschnitt 21 verschiedene Einstellungen evaluiert, um ein lokales Optimum zu finden.
		Kernfunktion für die nichtlineare Abbildung der Daten	
		Individuelle Parameter der Kernfunktion.	
RVM	0		
Bagging	1	Anzahl der Basisklassifikatoren im Ensemble	[5; 25; 50]
RF	2	Anzahl der Entscheidungsbäume im Ensemble	[50; 100; 250; 500]
		Anzahl der zufällig betrachteten Attribute je Ebene eines Entscheidungsbaumes	$[0{,}5;\ 1;\ 2] \cdot \sqrt{M}$ M = Anzahl Merkmale
SGB	1	Anzahl der Boosting-Iterationen	[5; 10; 25]

* Es ist üblich, eine exponentiell-skalierte Kandidatenliste zu verwenden, um einen möglichst großen Wertebereich zu evaluieren (Hsu et al. 2003, S. 5).