Full Model Selection mit vielen unabhängigen Variablen: Ein Beispiel für SAS-Tuning bei komplexen rechenintensiven Aufgaben

Rainer Kaluscha
Institut für Rehabilitationsmedizinische Forschung an der Universität Ulm
Wuhrstr. 2/1
Bad Buchau
rainer.kaluscha@uni-ulm.de

Silke Jankowiak
Institut für Rehabilitationsmedizinische Forschung an der Universität Ulm
Wuhrstr. 2/1
Bad Buchau
silke.jankowiak@uni-ulm.de

Gert Krischak
Institut für Rehabilitationsmedizinische Forschung an der Universität Ulm
Wuhrstr. 2/1
Bad Buchau
gert.krischak@uni-ulm.de

Zusammenfassung

Bei komplexen Auswertungen können erhebliche Rechenzeiten entstehen. Möchte man nicht tagelang auf Ergebnisse warten, lassen sich häufig – entsprechende Hardware vorausgesetzt - durch Parallelisierung und kleine Optimierungen erhebliche Zeitgewinne erzielen. Dies ermöglicht dann, z.B. im Rahmen von Sensitivitätsanalysen oder Robustheitsprüfungen, auch bei komplexen Modellen und großen Datensätzen, mehrere Varianten durchzuspielen und zu bewerten.

Schlüsselwörter: Variablenselektion, Logistische Regression, Rehabilitation, Arbeitsmarkt, Rechenzeit, Tuning, Optimierung

1 Einleitung

Der Einsatz moderner Computer erlaubt die Verwendung komplexer Rechenmodelle und großer Datensätze. Bei der Modellierung ist aber oft nicht a-priori klar, welche Variablen als potentielle Confounder in die Rechenmodelle aufgenommen werden sollen. Neben inhaltlichen Überlegungen kommen dann oft Heuristiken wie *forward selection* oder *backward elimination* zum Einsatz. Verschiedene Heuristiken kommen aber bei gleichen Daten nicht zwangsläufig zum gleichen Modell und es bleibt fraglich, ob sie immer das „beste“ Modell liefern.

Ferner kommen für manche Einflussgrößen oder die Zielgröße unterschiedliche Operationalisierungen in Betracht, so dass im Rahmen von Sensitivitätsanalysen mehrere Varianten zu prüfen sind.
Hier kann es schnell zu einer *kombinatorischen Explosion* kommen, da bei *k* Einflussvariablen für jede die binäre Entscheidung zu treffen ist, ob sie im Modell verwendet wird oder nicht. Es gibt also 2^k mögliche Modellvarianten. Aufgrund des exponentiellen Wachstums wird die Zahl der Modelle bereits bei relativ wenigen Variablen ziemlich groß; bei zehn Variablen sind schon mehr als 1.000 Modellvarianten ($2^{10} = 1.024$) möglich. Bestehen für eine oder mehrere Variablen auch noch verschiedene Möglichkeiten der Operationalisierung, muss sogar noch mit der Anzahl dieser Möglichkeiten multipliziert werden.
Nun stellt sich die Frage, wie sich diese vielen Modellvarianten am effektivsten durchrechnen lassen, damit Rechenzeit und Auswerteaufwand minimiert werden können; insbesondere, wenn z.B. aufgrund von Verfeinerungen der Stichprobe mehrere Durchläufe erforderlich sind. Dies wird im folgendem an einem Beispiel aus der Rehabilitationsforschung erläutert.

2 Problemstellung

Bei Rehabilitationsmaßnahmen der Rentenversicherung ist die anschließende berufliche Wiedereingliederung der Betroffenen ein wichtiges Ziel. Dabei spielen aber nicht nur medizinische Parameter, sondern auch externe Einflussgrößen wie der Arbeitsmarkt eine Rolle. Für die Untersuchung und Gewichtung der Einflussgrößen untereinander stand ein anonymisierter Routinedatensatz der Deutschen Rentenversicherung mit Angaben zu 400.000 Rehabilitationsmaßnahmen [1] zur Verfügung. Dieser wurde mit Arbeitsmarktdaten der Bundesagentur für Arbeit [2] angereichert (monatliche Arbeitslosenquoten auf Bundes- und Länderebene). Mittels eines logistischen Regressionsmodells wurde anschließend versucht, die berufliche Wiedereingliederung als binäre Zielgröße vorherzusagen.
Dabei kommen zwölf routinemäßig erhobene Patientenmerkmale als potentielle Confounder (unabhängige Variablen) in Betracht, d.h. es sind $2^{12} = 4.096$ Modellvarianten möglich. Jede dieser Modellvarianten soll zudem mit jeweils zehn verschiedenen Arbeitsmarktindikatoren kombiniert werden, um „gute“ von „schlechten“ Indikatoren zu trennen. Damit sind insgesamt 40.960 Modellvarianten zu berechnen und zu beurteilen. Die Prüfung aller Modellvarianten (*full model selection)* erlaubt Aussagen zur Robustheit der Ergebnisse gegenüber unterschiedlichen Strategien bei der Variablenselektion. So ist es durchaus möglich, dass eine Variable in einem Modell als signifikante Einflussgröße identifiziert wird, bei Hinzufügen oder Weglassen anderer Variablen diese Signifikanz aber wieder verliert.

3 Technische Realisierung

Für die Auswertungen stand ein leistungsfähiger Computeserver (16 CPUs, 256 GB RAM, SuSE Linux Enterprise 11, SAS 9.3) zur Verfügung. Der benötigte SAS-Pro-

grammcode für die Modellvarianten wurde mittels eigener Utilities aus einer Schablone generiert.

Neben Unix-Shell-Skripten kam dabei ein kleines C-Programm zum Einsatz, das eine Dezimalzahl in eine Bitfolge wandelt und aus einer vorgegebenen Variablenliste die den gesetzten Bits entsprechenden Variablen auswählt. Diese werden dann in das MODEL- bzw. CLASS-Statement eingesetzt. So ergibt z.B. Modellnr. *3334* die Bitfolge *110100000110* (niedrigstes Bit rechts), d.h. die Variablen 2, 3, 9, 11, 12 (hier: einstellige Hauptdiagnose nach ICD-10, Bundesland, Arbeitsfähigkeit bei Entlassung, Alter, Vorjahresentgelt) werden in die Schablone (hier durch spitze Klammern symbolisiert) eingesetzt:

```
PROC LOGISTIC data=FDZ.ERWERB;
   CLASS <ICD11 BLAND AFENTL>;
   MODEL ERWERB = <ICD11 BLAND AFENTL_jahre entgelt0>
                  <ggfs. Arbeitsmarktindikator>;
RUN;
```

Dabei muss noch unterschieden werden, welche Variablen nur im MODEL-Statement und welche auch im CLASS-Statement einzusetzen sind, z.B. über Groß-/Kleinschreibung gesteuert.

Setzt man alle Kombinationen der Variablen in die Codeschablone ein, ergeben sich 4.096 Code-Stücke. Diese wurden dann einmal ohne Arbeitsmarktindikatoren sowie mit neun unterschiedlichen Arbeitsmarktindikatoren vervielfältigt. Da a-priori nicht offenkundig war, ob die Arbeitslosenquote auf Bundes- oder Landesebene aussagekräftiger ist und zu welchem Zeitpunkt (bei Beginn der Rehabilitationsmaßnahme, bei Wiedereingliederung ein Jahr danach oder ein gleitender Durchschnitt über diesen Zeitraum) sie am aussagekräftigsten ist, sollten für diese Einflussgröße unterschiedliche Operationalisierungsmöglichkeiten durchgespielt werden. Zusätzlich wurde auch die auf den zehnjährigen Landesdurchschnitt standardisierte relative Landesquote als weiterer potentieller Arbeitsmarktindikator betrachtet.

4 Optimierung der Laufzeit

Berechnet man alle 40.960 Modelle ohne besondere Optimierungen innerhalb eines SAS-Jobs, ergibt sich eine Gesamtlaufzeit von zwei Tagen (48,2h). Dies entspricht einer durchschnittlichen Rechenzeit von 4,2s pro Modell, was angesichts der komplexen Modelle und der großen Fallzahl erstaunlich schnell ist.

Eine erste mögliche Optimierung ist die Verlagerung der SAS WORK-Library auf eine RAM-Disk, da SAS dort für jedes Modell einige größere temporäre Files erzeugt. Dies verkürzt die Laufzeit zwar auf 44,9h; der Zeitbedarf bleibt aber immer noch unbefriedigend hoch.

Anscheinend lassen sich die Operationen bei der Berechnung eines Modelles schlecht automatisiert parallelisieren, so dass SAS praktisch nur einen der sechzehn verfügbaren CPUs des Computeservers benutzt.

Verteilt man den generierten SAS-Code nun auf sechzehn Dateien, kann in sechzehn SAS-Jobs parallel gerechnet werden, so dass alle verfügbaren sechzehn CPUs ausgelastet werden. Dabei sollte durch entsprechende Kommandozeilenparameter jedem SAS-Job auf der RAM-Disk eine eigene WORK-Library zugewiesen und die USER-Library nur read-only geöffnet werden, um Konflikten und Wartezeiten durch gesperrte Dateien vorzubeugen:

```
sas -WORKINIT -WORK /ramdisk/01 –RSASUSER -NONEWS job01.sas &
sas -WORKINIT -WORK /ramdisk/02 -RSASUSER -NONEWS job02.sas &
sas -WORKINIT -WORK /ramdisk/03 -RSASUSER -NONEWS job03.sas &
```

Dies führt nun zu der gewünschten deutlichen Reduktion der Laufzeit: Das Ergebnis liegt statt nach zwei Tagen bereits nach knapp drei Stunden vor!

Setzt man Laufzeit (171min) und Gesamtrechenzeit (2.178min User und 312min System) in Beziehung, zeigt sich, dass durchschnittlich 14,5 von 16 CPUs parallel arbeiten konnten, d.h. die Laufzeit skaliert sehr gut mit der Anzahl der verfügbaren CPUs. Allerdings muss man sich um die Parallelisierung selbst kümmern und händisch bzw. per Batch-File mehrere SAS-Jobs erzeugen.

5 Analyse der Modellvarianten

Anschließend wurden mit Hilfe von Unix Utilities [3] die c-Statistik und Akaikes Information Criterion (AIC) zur Ermittlung der Modellgüte aus dem SAS-Output extrahiert und so die „besten“ Modelle bestimmt. Dies waren übrigens mit c-Werten von 0,81 die „großen“ Modellvarianten mit elf bzw. zwölf Variablen und einem der Arbeitsmarktindikatoren.

Ferner wurde für jede der zwölf Variablen sowie die Arbeitsmarktindikatoren gezählt, in wie vielen Modellvarianten sie statistisch signifikant wurden (Wald’sches Chi-Quadrat in der Effektanalyse).

Dies erlaubt interessante inhaltliche Rückschlüsse: so erwiesen sich die auf Landesebene abgeleiteten Indikatoren der Arbeitslosenquote denen auf Bundesebene abgeleiteten Indikatoren als überlegen. Erstere wurden nämlich in allen 4.096 Modellvarianten, in denen sie eingeschlossen wurden, auch statistisch signifikant, während letztere nur in etwa der Hälfte der Fälle im Modell blieben. Zudem waren auch die aus dem gleitenden Zwölfmonatsdurchschnitt gebildeten Indikatoren den punktuellen Messungen der Arbeitslosigkeit unmittelbar vor und ein Jahr nach der Rehabilitationsmaßnahme überlegen.

Auch bei den anderen Variablen ergaben sich durch Prüfung der Modellvarianten interessante Ergebnisse: Während etwa das Alter stets im Modell blieb, wurde das Geschlecht in etwa einem Fünftel der Varianten nicht signifikant; sein Einfluss auf die Zielgröße erscheint also als weniger robust.

6 Schlussfolgerung

Bei aufwändigen Problemstellungen können technische Optimierungen lohnen und helfen, das Potential moderner Hardware auszuschöpfen. Die erzielte Verkürzung der Laufzeit ermöglicht eine breitere inhaltliche Analyse, wenn z.B. zur Prüfung der Sensitivität oder Robustheit viele Modellvarianten betrachtet werden müssen. Somit können letztlich belastbarere Aussagen getroffen werden.

Literatur

[1] Forschungsdatenzentrum der Rentenversicherung: „Scientific Use File: Abgeschlossene Rehabilitation im Versicherungsverlauf 2002 – 2009 (SUFRSDLV09B)“. http://www.fdz-rv.de

[2] Bundesagentur für Arbeit: Aktuelle Daten - Arbeitslosigkeit und Grundsicherung für Arbeitsuchende nach Ländern. http://statistik.arbeitsagentur.de

[3] Kaluscha R: Datenmanagement mit Oracle, SAS, Perl und Unix Utilities: Werkzeuge für alle Fälle. KSFE 2007. http://de.saswiki.org/images/7/77/11.KSFE-2007-Kaluscha-Datenmanagement-Werkzeuge.pdf