Finding decision scores by evolutionary strategies

Claus-Dieter Volko

Referat des Artikels von Jürgen Paetz: Finding optimal decision scores by evolutionary strategies, *Artificial Intelligence in Medicine*, 32 (2004), S. 85-95.

1 Worum geht es?

In der medizinischen Diagnostik werden für bestimmte Krankheiten häufig Punktesysteme (*Scores*) verwendet, die in Tabellenform vorliegen. Aus diesen Tabellen kann abgelesen werden, unter welchen Umständen die jeweiligen Stadien vorliegen. Die Stadieneinteilung soll die Prognose des Verlaufs erleichtern und dem Arzt bei der Wahl der richtigen Therapiemaßnahmen helfen.

Klassische Scores werden von Experten erstellt und auf *"Consensus meetings"* verabschiedet. Sie stellen also Kompromisse zwischen den subjektiven Erfahrungen mehrerer anerkannter Fachleute dar. Aus diesem Grund stehen Scores im Verdacht, nicht optimal zu sein und manchmal Fehlprognosen zu liefern.

Der Autor vermutet, dass man einen besseren Score erhalten könne, indem man, ausgehend von konkreten Patientendaten, mit Hilfe einer evolutionären Strategie vorgehe, d.h. immer wieder neue Scores generieren lasse und schließlich den tauglichsten *"fittest"* Score auswähle.

Er demonstriert seinen Ansatz am Beispiel des septischen Schocks mit abdominaler Ursache, einer schweren systemischen Reaktion auf eine Infektion des Magen-Darm-Trakts. Der durch evolutionäre Strategien generierte Score, genannt "RRT-Score", schneidet im Vergleich mit mehreren klassischen, "expertengetriebenen" Scores tatsächlich gleich gut oder besser ab.

2 Klassische Scores in der Intensivmedizin

Für Intensivpatienten sind mehrere klassische Scores im Einsatz. Dazu gehören:

(a) Sepsis-bezogene Organausfallsbeurteilung (*Sepsis-related organ failure assessment, SOFA*): SOFA bewertet Störungen einzelner Organfunktionen mit ganzen Zahlen. Die Summe dieser Werte für einzelne Organe wird als "SOFA-Score" bezeichnet. Da SOFA die beste Klassfikationsleistung der Scores, welche wir hier nennen werden, aufweist, eine etwas detailliertere Beschreibung.

Folgende Variablen müssen überprüft werden: PaO_2 (der arterielle Sauerstoffpartialdruck), FiO_2 (der Anteil des Sauerstoffs an der gesamten eingetmeten Luft), Blutplättchen, Bilirubin, MAP (*mean arterial pressure*, durchschnittlicher arterieller Blutdruck, errechnet aus dem systolischen und dem diastolischen Blutdruck), die Katecholamine Dopamin (Dopa), Dobutamin (Dobu), Epinephrin (Adrenalin, Epi) und Norepinephrin (Noradrenalin, Nepi), Kreatinin und die Urinausscheidung. Weiters fließt in den SOFA die Glasgow-Koma-Skala (GCS) ein. Es handelt sich dabei um ein Punktesystem, mit dem der Bewusstseinszustand bewertet werden kann. Sie ist eher subjektiver Natur, weil Experten die Bewusstseinszustände auf Grund von Beobachtungen beurteilen. Daher wird sie in den folgenden Analysen weggelassen. Der maximale Punktestand von SOFA ist 24 (einschließlich GCS) oder 20 (ohne GCS). Ohne GCS müssen zwölf Einzelvariablen gemessen werden.

(b) Akute physiologische und chronische Gesundheits-Evaluation (*Acute physiological and chronic health evaluation, APACHE II*): APACHE II (als Verbesserung von *APACHE I*) ist ein Punktesystem für die Prognose des Zustands von Intensivpatienten hinsichtlich akuter Störungen, ihres Alters und des gesamten Gesundheitsstatus (ganzzahlige Werte von 0 bis 71).

(c) Vereinfachter akuter physiologischer Punktestand (*Simplified acute physiology score, SAPS II*): SAPS II ist eine Verbesserung von SAPS I, einem variablenreduzierten APACHE I. Nur 13 statt 34 Variablen werden verwendet.

(d) Mehrfacher Organ-Fehlfunktion-Punktestand (*Multiple organ dysfunction score, MODS*): MODS bewertet die Zustände von Organen (Lungen, Leber, Niere, Gerinnung, Herz, neurologisches System) mit ganzen Zahlen.

Wenn eine Variable, die für die Berechnung eines Punktesystems benötigt wird, nicht gemessen wurde, nahm der Autor in der folgenden Analyse an, dass ihr Wert unkritisch war, und wies ihr null Punkte zu. Dies ist eine allgemein übliche, vernünftige medizinische Annahme, obwohl die Resultate vielleicht wegen dieser Annahme nicht immer exakt sind.

Beim septischen Schock mit abdominaler Ursache ist die Klassifikationsleistung zum Zeitpunkt der Aufnahme gering. Wenn man eine ROC-Analyse durchführt, also die Sensitivität auf die Y-Achse und die Spezifität auf die X-Achse aufträgt, beträgt die Fläche unter der Kurve (*area under the curve*, AUC) zu diesem Zeitpunkt nur etwa 0,5, was einer Zufallsverteilung entspricht. Die Patienten

verweilen lange im Krankenhaus, im Schnitt 19,3 Tage. Da Ärzte während des Krankenhausaufenthalts durch ihre therapeutischen Maßnahmen den Zustand des Patienten verändern, kann der Ausgang der Erkrankung und Therapie normalerweise erst in den letzten drei Tagen vor der Entlassung bzw. vor dem Exitus klassifiziert werden. Aus diesem Grund wurde die Klassifikationsleistung der verschiedenen klassischen Scores und des neuen, durch eine evolutionäre Strategie generierten Scores nur während der letzten drei Tage miteinander vergleichen. In diesem Zeitraum findet sich eine AUC von 0,79 für MODS, 0,79 für SAPS II, 0,80 für APACHE II (nur akute Störungen) und 0,89 für SOFA.

3 Prinzip der Evolutionsstrategien

Zuerst müssen die Merkmale (synonym Variablen) ausgewählt werden, die für die Stadieneinteilung der Krankheit relevant sind. Dies geschieht mit einem Merkmalauswahlalgorithmus. Es gibt viele Algorithmen dieser Sorte. In dieser konkreten Arbeit wurde "Importance calculation for neuro-fuzzy networks" verwendet, ein Algorithmus, den Paetz in einer früheren Arbeit vorgestellt hatte. Aus dieser früheren Arbeit war bereits bekannt, dass die optimalen Merkmale für die Klassifikation des septischen Schocks der systolische Blutdruck, der diastolische Blutdruck und die Thrombozytenzahl sind. Diese drei Variablen geben dem neuen Score seinen Namen - "RRT-Score", wobei "RR" für den Blutdruck ("Riva-Rocci") und "T" für die Thrombozytenzahl steht.

Jeder Parameter kann Werte innerhalb eines bestimmten Intervalls annehmen, und die einzelnen Krankheitsstadien unterscheiden sich voneinander durch die Intervalle, in welchen die einzelnen Parameter jeweils liegen müssen. Bei der Optimierung des Score besteht das Ziel also einerseits darin, die Intervallgrenzen so festzulegen, dass eine größtmögliche Aussagekraft resultiert. Andererseits geht es darum, sich möglichst an die optimale Anzahl von Schritten pro Merkmal und damit an die optimale Anzahl von Krankheitsstadien anzunähern.

Der evolutionäre Algorithmus geht nach Vorbild der biologischen Evolution vor. Zuerst werden verschiedene Scores generiert und von diesen die 50 Prozent tauglichsten ausgewählt. Dann werden diese Parameter durch zufällige Änderungen dieser Individuen ("Mutationen") und Kombinationen von Individuen ("Rekombinationen") optimiert. Diese drei Schritte - Selektion, Mutation und Rekombination - werden nun solange wiederholt, bis das Abbruchkriterium erfüllt ist. Dieses Kriterium besteht meistens entweder im Erreichen einer bestimmten Generationenzahl oder einer lediglich minimalen Differenz zwischen den Tauglichkeits-

graden von zwei aufeinander folgenden Generationen. Im konkreten Projekt wurde das Erreichen der 25. Generation als Abbruchkriterium gewählt.

Wie werden die Scores bewertet? Zur Bewertung der Scores wird das Optimierungs-Kriterium durch eine *"Fitness function"* überprüft. Im Falle des vom Autor durchgeführten Projekts handelt es sich bei diesem Kriterium um die Leistung des neuen Score als Schwellenwertklassifikator: Ein optimaler Score soll in der Lage sein, mit einer hohen Wahrscheinlichkeit zu prognostizieren, ob der Patient überleben wird. Genauer gesagt, muss es eine Schwelle ϵ mit folgender Eigenschaft geben: Wenn der Punktestand höchstens ϵ beträgt, dann wird der Patient mit hoher Wahrscheinlichkeit sterben; wenn der Punktestand aber größer als ϵ ist, wird der Patient mit hoher Wahrscheinlichkeit überleben. Die optimale Schwelle für den Score wird berechnet, indem der Punktestand jedes Datensatzes, der als "überlebt" oder "verstorben" klassifiziert wird, überprüft wird. (Alle Datensätze sind entweder als "überlebt" oder als "verstorben" klassifiziert.) Es werden alle τ von 0 bis zum Maximum weniger eins durchprobiert und dasjenige τ, welches die höchste Klassifikationsleistung aufweist, genommen.

Die Scores werden als Tupel von Parametern $(n_1, ..., n_r; b_1, ..., b_{1,n_1}; ...; b_{r,1}, ..., b_{r,n_r})$ kodiert. Hierbei repräsentiert n_k die Anzahl der Schritte des Merkmals k, und $b_{k,1}$ bis b_{k,n_k} sind die Intervallgrenzen der einzelnen Schritte.

Wie sind Mutation und Rekombination in diesem konkreten Projekt implementiert worden? Es werden drei Mutationsoperatoren verwendet: Inkrementieren eines einzelnen Wertes n_i, Dekremenetieren von n_i und Veränderen einer einzelnen Intervallgrenzen $b_{i,j}$. Der Rekombinationsoperator geht so vor, dass er jedes Merkmal des rekombinanten Scores entweder aus einem der beiden miteinander zu rekombinierenden Scores ohne Änderungen übernimmt (*"copy"*) oder durch Mischen (*"merge"*) neu generiert. Das Mischen läuft wie folgt ab: Das n_i des rekombinanten Scores wird als Mittelwert der n_i der miteinander zu rekombinierenden Scores errechnet; die $b_{i,j}$ werden zufällig den beiden miteinander zu rekombinierenden Scores in einer geeigneten Anzahl entnommen.

Im konkreten Projekt wurde es so gehandhabt, dass in jeder Generation 35 Prozent der Veränderungen durch Mutation oder Rekombination in einer Änderung der Intervallgrenzen bestehen, je 18 Prozent in einer Zunahme bzw. Abnahme der Schrittzahl eines Merkmals und die restlichen rund 29 Prozent in Rekombinationen.

4 Resultate

Zur Generierung des RRT-Score wurden die Daten von 282 Patienten mit septischem Schock mit abdominaler Ursache genommen, die in 62 deutschen Spitälern zwischen 1997 und 2001 retrospektiv gesammelt worden waren. Sechzehn Patienten wurden ausgelassen, weil ihre Thrombozytenzahl nicht bekannt war. Nach dieser Vorbearbeitung blieben 608 dreidimensionale tägliche Datensätze von den übrigen 266 Patienten. Es gab in diesen Datensätzen keine fehlenden Werte. 289 Datensätze gehörten der Klasse "überlebt" und 319 der Klasse "verstorben" an. Der systolische und der diastolische Blutdruck waren normalverteilt; der Durchschnitt betrug 121,85 (Standardabweichung 22,61) bzw. 60,25 (Standardabweichung 13,61). Der Graph der Thrombozytenzahl enthielt einen Überhang auf der rechten Seite mit einem Mittel von 253,99 (Standardabweichung 186,18).

Fünf Experimente wurden mit denselben Parametereinstellungen durchgeführt und der Fortschritt der Anpassungszunahme verfolgt, indem die durchschnittliche Klassifikationsleistung aller Individuen pro Generation für jedes Experiment ausgegeben wurde. Die Scores wurden an Hand jeweils der Hälfte der Datensätze berechnet, die Klassifikationsleistung mit der jeweils anderen Hälfte überprüft. Jeder Patient kam dabei nur entweder in den "Trainingsdaten" oder in den "Testdaten" vor.

Der tauglichste Score, der in diesen Experimenten generiert wurde, mit der geringsten Summe an Schritten klassifiziert 87,58 Prozent der Datensätze korrekt. Die Anzahl der Schritte für die drei Merkmale beträgt dabei 5, 8 bzw. 6. Die optimale Schwelle ϵ is gleich 6, d.h. die Faustregeln "Wenn RRT-Score kleiner 6, dann verstorben" und "Wenn RRT-Score größer oder gleich 6, dann überlebt" liefern in 87,58 Prozent der Fälle richtige Aussagen.

Die ROC-Analyse ergab eine AUC von 0,90, was in etwa gleich der des SOFA-Score (0,89) war. Somit ist die Klassifikationsleistung des neuen RRT-Scores ungefähr gleich gut wie die von SOFA bzw. besser als die anderer klassischer Scores.

Die Mortalität wurden in Relation zu der Punktezahl gruppiert. Patienten mit null bis zwei Punkten sind in einem sehr kritischen Zustand. Der Zustand von Patienten mit drei bis fünf Punkten ist noch kritisch, mit sechs bis neun Punkten bereits viel weniger kritisch. Mit 10 bis 13 Punkten sind Patienten in einem unkritischen Zustand. Zudem ist es bemerkenswert, dass die Mortalität mit fünf

Punkten 68,29 Prozent beträgt - 56 von 82 Datensätzen stammen von verstorbenen Patienten. Eine Mortalität von sechs Punkten weisen nur 19,75 Prozent auf - 16 von 81 Datensätzen stammen von verstorbenen Patienten. Dies ist eine um 48,54 Prozent geringere Mortalität als die Mortalität der Patienten mit einem RRT von fünf Punkten. Daher sollte ein Arzt sehr wachsam sein, wann immerder Punktestand auf fünf abnimmt.

Da das Punktesystem ausschließlich unter Verwendung von Daten der letzten drei Tage des Aufenthalts in der Intensivstation optimiert wurde, bleibt eine Frage offen: Kann das Punktesystem während des gesamten Krankenhausaufenthalts an allen Tagen verwendet werden? Um diese Fragestellung zu untersuchen, wurde übrprüft, ab welchem Tag der RRT-Score nur mehr Werte annimmt, die mit der endgültigen Prognose übereinstimmen, also größer gleich 6, wenn der Patient letzten Endes lebend entlassen wird, oder kleiner 6, wenn er im Krankenhaus versterben wird. Die Anzahl der Tage von diesem Zeitpunkt an bis zur Entlassung bzw. zum Exitus wird vom Autor als *"c-day"* bezeichnet. Es stellte sich heraus, dass der "c-day" bei überlebenden Patienten im Durchschnitt 11,0 bzw. bei verstorbenen Patienten im Durchschnitt 7,3 Tage betrug. Der RRT-Score ist also nicht nur in den letzten drei Tagen aussagekräftig.

Konklusiv lässt sich sagen, dass der Vergleich mit SOFA zeigte, dass der neue RRT-Score eine ähnliche Klassifikationsleistung aufweist. RRT ist leichter zu berechnenund zu verstehen als SOFA, denn SOFA benutzt Quotienten und die Boole'sche Oder-Funktion. Die Klassifikationsleistung von RRT ist viel besser als die von MODS, SAPS II und APACHE II.

Der Autor meint, dass ein Versuch, SOFA mit evolutionären Strategien zu optimieren, interessant wäre. Er hat diesen Versuch jedoch nicht durchgeführt, weil hierfür zu viele Daten fehtlen. Zwar waren von 266 Patienten systolischer Blutdruck, diastolischer Butdruck und Thrombozytenzahl bekannt, was für RRT ausreichte, jedoch fehlten andere Parameter, die für SOFA benötigt würden.

Abschließend sei gesagt, dass bereits SOFA kein rein expertengetriebenes Punktesystem ist, sondern schon in SOFA statistische Überlegungen eingeflossen sind. Das mag der Grund sein, warum RRT nur eine geringfügig verbesserte Klassifikationsleistung als SOFA aufweist, jedoch eine deutlich bessere als andere klassische Punktesysteme.

Der Autor ist der Meinung, dass evolutionäre Strategien zur Score-Generierung für medizinische Fragestellungen verstärkt verwendet werden sollen, und man

soll auch mögliche Anwendungen in anderen Gebieten, wie etwa in der Bioinformatik, untersuchen.