

Intelligente Systeme

Aufgabe 3 - Verhaltensklassifikation

Verhaltensklassifikation

Ausgangssituation	2
Annahmen / Limitierungen	2
Inbetriebnahme der Software	3
Allgemeiner Ablauf der Software	3
Parametereinstellungen	3
Erklärung zur Klassifikation	4
Ergebnisse und Evaluation	4

Ausgangssituation

Es werden zwei Fischarten - Marlin und Segelfisch - betrachtet, die sich relativ stark in ihrem Aussehen aber auch in ihrem Angriffsverhalten bei der Jagd nach Sardinen ähneln. Durch eine Software sollen Angriffssequenzen einem der beiden Fische zugeordnet werden.

Biologen haben innerhalb eines Forschungsprojektes das Jagdverhalten beider Arten beobachtet, untersucht und dabei immer wiederkehrende, atomare Aktionen erkannt, die in ihrer Zusammensetzung einem Angriff entsprechen. Es ist zu beachten, dass ein Angriff nicht aus einer bestimmten Sequenz von Aktionen besteht, sondern eher aus einer Folge dieser Aktionen.

Die Software soll die Angriffssequenzen unter Beachtung der Bayes'schen Klassifikation und der Markov-Kette erster Ordnung zuordnen.

Annahmen / Limitierungen

Die Formel der Bayes'schen Klassifikation wurde wie folgt vereinfacht:

$$\begin{aligned}c_{\max} &= \operatorname{argmax}_{c \in C} P(c | b) \\&= \operatorname{argmax}_{c \in C} [P(b|c) * P(c) / P(b)] \\&= \operatorname{argmax}_{c \in C} [P(b|c) * P(c)] \\&= \operatorname{argmax}_{c \in C} [P(b|c)]\end{aligned}$$

$P(b)$ kann in der Berechnung weggelassen werden, weil dieser Wert keinen Einfluss am Endergebnis hat. $P(c)$ kann ebenfalls weggelassen werden, da zur Vereinfachung angenommen werden darf, dass beide Fischarten gleich häufig auftreten, nämlich jeweils zu 50%.

Wir nehmen an, dass der Benutzer der Software nur Dateien übergibt, die nur korrekte Angriffssequenzen beinhaltet.

Eine weitere Annahme unserer Seite ist, dass eine Angriffssequenz, die von unserer Software gleich wahrscheinlich zum Segelfisch und Marlinfisch zugeordnet wird, zu beiden zugeordnet werden darf. Das bedeutet, diese Angriffssequenz wird doppelt zugeordnet, da sich unsere Software nicht entscheiden kann, zu welchem Fisch die Sequenz gehört.

Theoretisch müssten wir alle Wahrscheinlichkeiten zur Berechnung in ihren Logarithmus umwandeln, da mit sehr kleinen Wahrscheinlichkeiten das Problem auftritt, dass die Berechnungen ungenau werden. Wir verzichten bei unserer Rechnung, die Wahrscheinlichkeiten umzuwandeln, da die Werte nicht zu klein sind.

Inbetriebnahme der Software

Der Software müssen zum Start vier Argumente übergeben werden:

1. Der Pfad zur Datei, in dem die Trainingsdaten für den Marlin Fisch gespeichert sind.
2. Der Pfad zur Datei, in dem die Testdaten für den Marlin Fisch gespeichert sind.
3. Der Pfad zur Datei, in dem die Trainingsdaten für den Segelfisch gespeichert sind.
4. Der Pfad zur Datei, in dem die Testdaten für den Segelfisch gespeichert sind.

Weitere Einstellungen muss der Benutzer nicht vornehmen. Die Software visualisiert in der Konsole eine kleine Tabelle, die mittels der Markov-Kette erster Ordnung erstellt wurde. Am Ende der Ausführung des Programms werden die Prozentsätze der korrekt klassifizierten Sequenzen (Marlin Fisch, Segelfisch, beide vereinigt) wiedergegeben.

Allgemeiner Ablauf der Software

Wurde die Software korrekt gestartet, so werden zuerst die Daten aus den Dateien für die weitere Verarbeitung importiert. Anschließend werden für beide Fischarten deren Tabellen durch die Markov-Kette erster Ordnung erstellt. Je Fischart werden zwei Tabellen erstellt, einmal mit der absoluten und einmal mit der relativen Häufigkeit. Die Tabellen werden anhand des Trainingsdatensatzes des jeweiligen Fisches befüllt und bestimmt.

Mit Hilfe der erstellten Tabellen werden die Angriffssequenzen, die sich in den jeweiligen Testdateien der Fischarten befinden, zugeordnet. Anschließend wird berechnet, wie viele Angriffssequenzen korrekt zugeordnet wurden und in Prozent auf der Konsole ausgegeben.

Parametereinstellungen

In unserer Software gibt es ein paar Parameter, die leicht und schnell geändert werden können:

1. `matrix_size_x`: Mit diesem Parameter wird angegeben, wie viele Spalteneinträge die Matrix bzw. die Tabelle hat.
2. `matrix_size_y`: Mit diesem Parameter wird angegeben, wie viele Zeileneinträge die Matrix bzw. die Tabelle hat.
3. `matrix_params`: Mit diesem Parameter (eine Liste) wird angegeben, was im Header-Kopf der Zeile und der Spalte steht. In der Liste sind folgende Parameter enthalten:
 - `approach` (Annäherung an den Sardinenschwarm)
 - `bill_use` (Einsatz des Rostrums)
 - `prey_contact` (Direkter Kontakt des Rostrums mit der Beute)
 - `open_mouth` (Öffnen des Mauls)
 - `ingest` (Fressen einer Sardine)
 - `leave` (Beenden des Angriffs)

Erklärung zur Klassifikation

In diesem Abschnitt geht es darum, wie die einzelnen Angriffsequenzen klassifiziert werden - ob die jeweilige Sequenz zum Marlin Fisch gehört oder zum Segelfisch.

Zuerst werden durch den jeweiligen Trainingssatz der Art der jeweilige Klassifikator trainiert. Es werden durch die Markov-Kette erster Ordnung zwei Matrizen erstellt, in den die absolute bzw. die relative Häufigkeit (gerundet) dargestellt ist, wie wahrscheinlich es ist, dass die Aktion x aus der Angriffsequenz ausgeführt wird, wenn die Aktion y zuvor ausgeführt wurde. An einem Beispiel erklärt, werden die Wahrscheinlichkeiten gespeichert, wie wahrscheinlich es ist, dass ein Fisch sein Maul öffnet, nachdem direkter Kontakt mit dem Rostrum mit der Beute bestand.

Mit Hilfe der erstellten Matrizen werden nun die Angriffsequenzen aus den jeweiligen Testdateien einer Art zugeordnet. Dabei wird die Regel der Bayes'schen Klassifikation angewendet, dessen Formel schon im Abschnitt "Annahmen / Limitierungen" erwähnt wurde. Einer Angriffsequenz wird die jeweilige Wahrscheinlichkeit zugeordnet, wie wahrscheinlich es ist, dass diese vom Marlin Fisch bzw. vom Segelfisch ausgeführt wurde. Die Sequenz wird dem Fisch zugeordnet, wo die Wahrscheinlichkeit höher ist.

Im Detail werden die Wahrscheinlichkeiten wie folgt berechnet: Für jede Aktion aus einer Sequenz wird die Wahrscheinlichkeit bestimmt, wie wahrscheinlich es ist, dass diese ausgeführt wird, wenn die davorliegende Aktion aus der Sequenz gegeben ist. Dabei muss für die Bestimmung nur in die jeweilige Matrix geschaut und der richtige Wert ausgelesen werden. Die ganzen, resultierenden Wahrscheinlichkeiten werden anschließend miteinander multipliziert, sodass man die Wahrscheinlichkeit erhält, wie wahrscheinlich es ist, dass die Angriffsequenz vom Marlin Fisch bzw. vom Segelfisch ausgeführt worden ist.

Ergebnisse und Evaluation

Die Sequenzen aus den Testdaten aus den jeweiligen Dateien der Fischart waren von vornherein korrekt zugeordnet. Es sollte hier nur überprüft werden, wie gut sich diese Angriffsequenzen zu der jeweiligen Fischart zurück ordnen ließ, wenn vorher mittels Trainingsdaten der Klassifikator trainiert worden ist.

Anhand der Matrizen, die mittels der Markov-Kette erster Ordnung erstellt worden ist, sind Unterschiede im Jagdverhalten zu erkennen, so öffnet beispielsweise der Segelfisch sein Maul viel häufiger, nachdem er den Schwarm von Sardinen erreicht hat, als der Marlin Fisch.

Das Ergebnis aus der Klassifikation ist etwas erstaunlich. Es können fast alle (ca 96,7%) Angriffsequenzen dem Segelfisch wieder korrekt zugeordnet werden, dem Marlinfisch allerdings nur ungefähr 63,3%. Eine Erklärung von unserer Seite, aber nicht bestätigte wäre, dass das Angriffsverhalten des Marlinfisches variabler ist als das vom Segelfisch und somit auch einige Sequenzen falsch zugeordnet werden.

Die vereinigte Wahrscheinlichkeit von korrekt zugeordneten Sequenzen beträgt 61,2%.