CAT / PIKE Breast Cancer Detection

Volkan Korunan

03.02.2020

Kann ... // Algorithmus x // exact ... // (Kunden)-Problem ... berechnen / lösen?

Ja, ein neuronales Netz kann die Klassifikation mit einer Genauigkeit von 98% durchführen.

CAT bzw. PIKE

- **Problem (P):** Welches Problem ist für die Lösung des Kunden von größter Bedeutung?
  Kann mit Hilfe eines neuronalen Netzes zuverlässig eine Unterscheidung zwischen gesundem und Krebsgewebe durchgeführt werden. Die Klassifizierung soll anhand von Eigenschaften, welche aus einer CSV Tabelle die Inofromation entnommen werden und anschließend angewendet wird.

- **Intervention (I):** Welche Bibliotheken und Algorithmen werden hauptsächlich verwendet? (z.B. `pandas` für Finanzdaten)
  Für das Aufbauen und Trainieren des neuronalen Netzes wurde die Tensorflow Bibliothek in der Version 2.10.0 mit der Keras High-Level Schnittstelle verwendet. Für die Datenanalyse und Preprocessing wurde weiterhin pandas, numpy, matplotlib und einige Funktionen aus dem Scikit-Learn Paket verwendet.

- **Kontrollintervention (K):** Welche alternative Methode oder Algorithmus könnte ebenfalls angewendet werden? (z.B. `scikit-learn` für Klassifikationsaufgaben)
  Es wäre möglich ander Bibliothek wie XGBoost oder TensorFlow zu verwenden, um Machine-Learning-Modelle zu erstellen. Diese Bibliotheken bieten unterschiedliche Ansätze, die in bestimmten Situationen bessere Ergebnisse liefern können. Zudem könnte man ein anderen Ansatz nutzen statt eine Binäreklassifikation, könnte man auch eine Entscheidungsbaum im Betracht ziehen.

- **Ergebnismaß (E):** Was ist das Ziel der Analyse? (z.B. einen Prädiktor oder Klassifikator erstellen)
  Eine Klassifizierung mit hoher Genauigkeit (98%) soll erreicht werden.

Die Suche nach der besten Evidenz

1. Problemdefinition
   Die Problemdefinition besteht darin, eine Methode zu finden, die mithilfe eines neuronalen Netzes Proben von Brustgewebe, die durch Feinnadelbiopsie gewonnen wurden, zuverlässig als gesund (gutartig) oder krebsartig (bösartig) klassifizieren kann. Das Ziel ist es, anhand der Zellmerkmale im Datensatz die Diagnosegenauigkeit zu verbessern und eine automatisierte Unterscheidung zu ermöglichen.

2. Definition einer relevanten, suchbaren Frage
   Kann die Klassifizierung mittels eines neuronalen Netzes mit hoher Genauigkeit (98%) durchgeführt werden?

3. Auswahl der wahrscheinlichsten Quelle für die Evidenz
   Nach dem Training wird die Performance des Modells an einer Auswahl der vorhandenen Daten (Testset, 20%) evaluiert.

4. Erstellung einer Suchstrategie
   Verschiedene Architekturen für neuronale Netze (Anzahl der Schichten, Anzahl der Einheiten, Aktivierung, etc.) werden untersucht, um eine möglichst hohe Genauigkeit auf einem Validierungsset zu erreichen. Um Overfitting der Trainingsdaten zu vermeiden, werden mit Early-Stopping und Dropout zwei Strategien während des Trainings untersucht.

5. Zusammenstellung der Evidenzausbeute

   - Falls die Ausbeute gering ist, Auswahl einer alternativen Quelle und Wiederholung der Schritte
     - **Alternative Algorithmen:** Einsatz von Random Forests, Gradient Boosting (XGBoost) oder Support Vector Machines als Alternativen.
     - **Feature Engineering:** Zusätzliche Merkmale könnten extrahiert oder die relevanten Features anders kombiniert werden.
     - **Hyperparameter-Tuning:** Feinabstimmung der Modell-Hyperparameter, z.B. Lernrate, Anzahl der Schichten oder Regularisierungsparameter.
     - **Datensätze erweitern:** Wenn man andere Informationen mit einfließen würde , könnte es zu besseren Klassifikation führen.

6. Anwendung der Evidenz

   - Das Modell kann für die Vorhersage bzw. Diagnose von Brustkrebs eingesetzt werden.
   - Eine regelmäßige Überprüfung der Modellleistung auf neuen Daten ist empfohlen, um die Vorhersagequalität langfristig zu gewährleisten.