Modulkonzept zu Detektionstheorie

Förderkennzeichen:	01IS17073
Vorhabenbezeichnung:	Verbundprojekt: COSY-Entwicklung von sieben Praxis-Versuchen zum Thema Datenanalyse und Maschinelles Lernen an zwei Hochschulstandorten
Zugehörigkeit:	Modulkonzept Versuch 4 - Umweltdaten

Allgemeines:

Die Detektionstheorie beschäftigt sich mit der Erkennung von Auffälligkeiten, zum Beispiel Ausreißern, in einem Datensatz. Generell sind Ausreißer Datenpunkte, welche bezüglich ihrer Werte oder der Art ihres Auftretens von den erwarteten Daten abweichen. In der Regel wird zwischen punktuellen, kontextuellen und kollektiven Ausreißern unterschieden.

Punktuelle Ausreißer beschreiben einzelne Datenpunkte, die für sich gesehen vom Rest der Daten abweichen. Dies kann bei gemittelten täglichen Außentemperaturen zum Beispiel ein Wert von -999 Grad Celsius sein.

Kontextuelle Ausreißer unterscheiden sich nicht generell vom Rest der Daten sondern nur in einem bestimmten Kontext. Bei den Außentemperaturen wäre das zum Beispiel ein Wert von 30 Grad Celsius im Januar. Dies wäre kein punktueller Ausreißer, da ein solcher Wert im Sommer auftreten kann. Im Kontext der Jahreszeit kann der Wert jedoch als kontextueller Ausreißer klassifiziert werden, da 30 Grad Celsius im Winter auf einen Messfehler schließen lassen.

Kollektive Ausreißer sind Datenpunkte die für sich gesehen keine Ausreißer sind, aber in Verbindung mit anderen Daten auffällig sind. Bezüglich der Außentemperaturen wäre dies der Fall wenn an mehreren Tagen hintereinander exakt der gleiche Temperaturwert erfasst worden wäre. Diese Werte können für sich gesehen absolut plausibel sein, allerdings ist das mehrmalige Auftreten der exakt gleichen Temperatur an mehreren Folgetagen extrem unwahrscheinlich.

Zur Visualisierung punktueller Ausreißer sollen die Studierenden die Daten als Histogramplot darstellen und mit einer Normalverteilungskurve vergleichen. Punktuelle Ausreißer können dadurch sehr einfach als Balken weit außerhalb des Bereichs der Normalverteilung identifiziert werden.

Für kontextuelle Ausreißer soll eine Zeitreihenzerlegung (Time Series Decomposition) durchgeführt werden. Bei dieser Methode wird jeder Wert eines Datensatzes in drei Komponenten (Trend T_t, Saisonwert S_t, Restfehler I_t) zerlegt, die entweder addiert (additive Methode) oder multipliziert (multiplikative Methode) den ursprünglichen Messwert ergeben. Da die multiplikative Methode besser zu Zeitreihen mit einem exponentiellen Wachstum passt, soll hier die additive Methode verwendet werden. Die Zerlegung ergibt sich also wie folgt:

$$y_t = T_t + S_t + I_t$$

Die einzelnen Komponenten ergeben sich wie folgt:

$$T_t = y * \begin{bmatrix} p-1 \text{ mal} & \begin{cases} \frac{1}{2p} \\ \frac{1}{p} \\ ... \\ \frac{1}{p} \\ \frac{1}{2p} \end{cases} \end{bmatrix}$$

$$S_t = \frac{1}{p} \cdot \sum_{i=t-\frac{p}{2}}^{t+\frac{p}{2}} T_i$$

$$I_t = y_t - T_t - S_t$$

mit p = Fenstergröße

Diese Methode bezieht den Kontext in Form der Saisonkomponente und des Trends in die Betrachtung ein und eignet sich daher für die Detektion von kontextuellen Ausreißern.

Zur Erkennung von kollektiven Ausreißern soll die Varianz in sich bewegenden Zeitfenstern berechnet werden. Dabei zeigt ein Varianzwert nahe 0 eine Häufung identischer Temperaturwerte, die einen kollektiven Ausreißer darstellen.

Die Datenvorverarbeitung wird in diesem Modulkonzept nicht behandelt, stattdessen werden fertig vorverarbeitete Datensätze bereitgestellt, an denen ohne weitere Arbeitsschritte Ausreißerdetektion durchgeführt werden kann. Versuch 4 eignet sich besonders für die Vermittlung von Ausreißerdetektion, da in diesem Versuch eine Vielzahl von Datensätzen vorliegen die alle Typen von Ausreißern enthalten.

Aufgaben:

Den Studierenden werden Beispieldatensätze für verschiedenen Ausreißertypen zur Verfügung gestellt. Anhand dieser Datensätze sollen die Studierenden mit verschiedenen Methoden Ausreißer in den Datensätzen erkennen und visualisieren. Nacheinander werden die Methoden erklärt und sollen dann anhand der jeweiligen Daten angewendet werden. Die Teilschritte dabei sind:

- Einlesen von Daten
- Detektion punktueller Ausreißer durch
  - Plotten eines Histograms und der zugehörigen Normalverteilung
  - Erkennen punktueller Ausreißer durch Interpretation des Plots
- Detektion kontextueller Ausreißer durch
  - Implementierung und Ausführung einer Zeitreihenzerlegung
  - Erkennen kontextueller Ausreißer durch Betrachtung der Restfehlerkomponente
- Detektion kollektiver Ausreißer durch
  - Berechnung der Varianz innerhalb eines sich über die Daten bewegenden Zeitfensters

– Erkennen kollektiver Ausreißer durch Betrachtung der Varianzen

Ziele:

In diesem Modul soll Studierenden die Notwendigkeit der Ausreißererkennung und Datenbereinigung bewusst gemacht werden und sie sollen verschiedene Methoden kennen lernen wie dies bewerkstelligt werden kann. Konkret wird dies am Beispiel der zur Verfügung gestellten Temperaturdaten veranschaulicht, wodurch die Studierenden auch die Besonderheiten bei der Analyse von Zeitreihendaten kennen lernen.
Lernziele sind im einzelnen:

- Die Studierenden verstehen die Notwendigkeit der Ausreißerdetektion
- Die Studierenden kennen und verstehen verschiedene Verfahren um verschiedene Typen von Ausreißern zu erkennen
- Die Studierenden kennen die Besonderheiten von Zeitreihendaten
- Die Studierenden sind in der Lage Daten zu visualisieren und zu interpretieren.
- Die Studierenden kennen und verstehen die zur Ausreißerdetektion benötigten statistischen Modelle.

Literatur:

- A Fuller, Wayne. (2019). Introduction to statistical time series / Wayne A. Fuller. SERBIULA (sistema Librum 2.0)