DOKUMENTATION MICROAGGREGATIONUTIL

ALLGEMEINE BESCHREIBUNG

Die Klasse *MicroAggregationUtil* stellt eine praktische Implementierungen der von *J. Höhne* im Artikel *„Anonymisierungsverfahren für Paneldaten“* theoretisch beschriebenen *eindimensionalen Mikroaggregationsverfahren mit fester und variabler Gruppengröße* bereit. Die Implementierungen sollen mit dieser Dokumentation anschaulich erläutert werden. Zusätzlich wurde natürlich auch darauf geachtet, dass der Quellcode selbst ausreichend dokumentiert ist.

Das Grundprinzip von Mikroaggregationsverfahren ist die Gruppierung von möglichst ähnlichen Merkmalsträgern und deren Vereinheitlichung durch das Ersetzen der Merkmalswerte durch den Durchschnittswert der Gruppe. Durch diese Vereinheitlichung der Merkmalswerte wird das Risiko einer eindeutigen Zuordnung gesenkt und gleichzeitig wegen der Veränderung der Merkmalswerte der Nutzen einer eventuellen Reidentifikation von Einheiten reduziert. [1]

SCHNITTSTELLEN

Eine Instanz der Klasse erhält man mittels `MicroAggregationUtil.getInstance()` danach stehen einem die beiden Methoden

- `performOneDimensionalMicroAggregationWithFixedGroupSize()`
- `performOneDimensionalMicroAggregationWithVariableGroupSize()`

zur Verfügung.

ARBEITSWEISE

An einem konkreten Beispiel soll nun die Arbeitsweise des Algorithmus für die Mikroaggregation für mit variabler Gruppengröße beschrieben werden. Dieser nutzt (nach dem Finden der optimalen Gruppengröße) den Algorithmus für eine feste Gruppengröße, weswegen auf eine separate Beschreibung verzichtet wird.

[1] (Höhne, 2008)

BEISPIELDATEN

Die Arbeitsweise soll anhand folgender Beispieldaten erläutert werden:

356,00
670,00
815,00
132,00
613,00
916,00
538,00
348,00
3,00
396,00
401,00

SCHRITT 1: SORTIEREN

Im ersten Schritt wird die Spalte absteigend sortiert:

916,00
815,00
670,00
613,00
538,00
401,00
396,00
356,00
348,00
132,00
3,00

SCHRITT 2: OPTIMALE GRUPPENGRÖßE ERMITTELN

In diesem Schritt werden ausgehend von der übergebenen initialen Gruppengröße m, alle Gruppengrößen im Intervall $M := [m, 2m-1]$ getestet. Laut Höhne soll $m \geq 3$ gewählt werden. Für $m = 3$ ergibt sich somit beispielsweise das Intervall $M = [3,5]$, es werden also die Gruppengrößen 3,4, und 5 überprüft.

SCHRITT 2.1: SPALTE ZERLEGEN

Eine Gruppengröße gilt als optimal wenn die gruppeninterne Varianz minimal ist. Um dies zu ermitteln wird die Spalte nacheinander in Gruppen der Größe $m \in M$ zerlegt. Für $m = 3$ ergeben sich damit beispielsweise die Mikroaggregationsgruppen:

916,00
815,00
670,00

613,00
538,00
401,00

396,00
356,00
348,00

132,00
3,00

SCHRITT 2.2: KENNZAHLEN ERMITTELN

Für jede Mikroaggregationsgruppe G werden dann folgende Kennzahlen ermittelt:

MITTELWERT

Zunächst wird die Gruppengröße von G bestimmt:

$$groupSize = |G|$$

Dann werden alle Elemente $g \in G$ aufsummiert:

$$groupValueSum = \sum_{value \in G} value$$

Der Mittelwert ergibt sich dann aus:

$$groupMeanValue = \frac{groupValueSum}{groupSize}$$

STANDARDABWEICHUNG

Diese ergibt sich aus:

$$groupStandardDeviatation = \sqrt{\sum_{value \in G} \frac{(value - meanValue)^2}{groupSize}}$$

VARIANZ

Die Varianz ist das Quadrat der Standardabweichung:

$$groupVariance = groupStandardDeviatation^2$$

Für die erste Mikroaggregationsgruppe aus dem Beispiel ergeben sich dann folgenden Kennzahlen:

- $groupSize = 3$
- $groupValueSum = 916 + 815 + 670 = 2401$
- $groupMeanValue = \frac{2401}{3} = 800.33$
- $groupStandardDeviatation = 100.96$
- $groupVariance = 100.96^2 = 10193.56$

SCHRITT 2.3: GRUPPENGRÖßE PRÜFEN

In diesem Schritt wird überprüft ob die aktuelle Gruppengröße eine Verbesserung bringt. Dazu werden die einzelnen gruppeninternen Varianzen aufsummiert und überprüft ob diese Summe kleiner als die Summe für die vorherige Gruppengröße ist. Initialisiert wird die Summe dabei mit dem $positven\ Unendlichen$, d.h. das Überprüfen der ersten Gruppengröße bringt somit immer eine Verbesserung.

$$groupVarianceSum = \sum groupVariance$$

$$optimalGroupSize \leftarrow \min(groupVarianceSum)$$

Hat man die optimale Gruppengröße ermittelt kann man das Anonymisierungsverfahren für feste Gruppengrößen verwenden.

Für das Beispiel ergibt sich eine optimale Gruppengröße von 3.

SCHRITT 3: WERTE ANONYMISIEREN

Dieser Schritt ist für beide Mikroaggregationsverfahren mit fester und variabler Gruppengröße identisch. Zunächst werden nach dem Teilen der Gruppe (vgl. Schritt 2.1) die in Schritt 2.2 beschrieben Kennzahlen berechnet. Mittels dieser wird dann Anonymisierung der Werte durchgeführt.

SCHRITT 3.1: MIKROAGGREGATIONSGRUPPE TEILEN

Hat man eine Spalte in Mikroaggregationsgruppen der Größe m zerlegt, wird eine einzelne Mikroaggregationsgruppe in diesem Schritt wiederum in zwei möglichst gleich große Untergruppen geteilt: in größere und kleinere Werte (absteigend sortiert). Dazu wird zunächst die Größe der Untergruppe mit den größeren Werten bestimmt:

$$biggerValuesGroupSize = rint(groupSize/2)$$

Wobei *rint()* den ganzzahligen Anteil berechnet. Der Wert $biggerValuesGroupSize$ kann lediglich bei einer $groupSize < 2$ den Wert 0 annehmen, weswegen er in diesem Fall auf 1 gesetzt wird und die damit verbundene 1-elementige Restgruppe wie eine Untergruppe größerer Werte behandelt wird.

Die Größer der Untergruppe der kleineren Werte ergibt sich folglich aus:

$$smallerValuesGroupSize = groupSize - biggerValuesGroupSize$$

Hat man die Mikroaggregationsgruppe in Untergruppen der entsprechenden Größe geteilt, werden diese wie von Höhne beschrieben im nächsten Schritt unterschiedlich anonymisiert.

Das Teilen der ersten Mikroaggregationsgruppe aus dem Beispiel ergibt folgende Untergruppen:

916,00
815,00

670,00

SCHRITT 3.2.1 GRÖßERE WERTE ANONYMISEREN

Die Untergruppe der größeren Werte wird folgendermaßen werden alle durch einen anonymisierten Wert ersetzt der sich wie folgt berechnet.

$$meanValue + \sqrt{\frac{groupSize - biggerValuesGroupSize}{biggerValuesGroupSize}} * standardDeviatation$$

Für das Beispiel ergibt sich hier 800.33 als anonymisierter Wert.

SCHRITT 3.2.1 KLEINERE WERTE ANONYMISIEREN

Die Formel zur Ersetzung der Gruppe die die kleineren Werte beinhaltet lautet:

$$meanValue - \sqrt{\frac{biggerValuesGroupSize}{groupSize - biggerValuesGroupSize}} * standardDeviatation$$

Für das Beispiel ergibt sich der Wert 657.55.

SCHRITT 4: WERTE ZUSAMMENFÜHREN

Nachdem eine Spalte zunächst in Mikroaggregationsgruppen und diese dann wiederum in Untergruppen für größere und kleinere Werte, müssen diese natürlich wiederum zusammengeführt werden.

In unserem Beispiel sieht dies so aus:

800,33
800,33
657,54
517,33
517,33
393,20
366,66
366,66
336,97
158,71
67,5

Einzelne Mikroaggregationsgruppen wurden durch unterschiedliche Farben hervorgehoben. Innerhalb der Mikroaggregationsgruppen stellen die dunkleren Zellen die größeren Werte und die helleren Zellen entsprechend die kleineren Werten dar.

SCHRITT 5: WERTE MISCHEN

Dieser Schritt wird zwar von Höhne nicht explizit erwähnt, jedoch wären Daten in einem Zustand wie nach Schritt 4 eventuell sehr leicht wieder deanonymisierbar, weswegen die Spalte noch einmal durchgemischt wird.

ERGEBNIS

Abschließend die Ein- und Ausgangsdaten für das Beispiel im direkten Vergleich:

356	517,33
670	800,33
815	657,54
132	158,71
613	366,66
916	517,33
538	393,2
348	366,66
3	67,5
396	336,97
401	800,33

LITERATURVERZEICHNIS

Höhne, J. (2008). Anonymisierungsverfahren für Paneldaten. *Wirtschafts- und Sozialstatistisches Archiv, Band 2, Heft3*, S. 259-267.

UMSETZUNG

Pascal Wasem
Matrikelnumer 3479498
PI Master
Pascal.Wasem@googlemail.com