Anonymisierungsverfahren

Verfahren zur Anonymisierung von Einzeldaten

Son Pham

Ludwig-Maximillian-Universität München Institut für Statistik

19. Juni 2015

Inhaltsverzeichnis

1. 1. Einleitung
   - Gesetzliche Grundlage
   - Arten der Anonymität

Inhaltsverzeichnis

1. 1. Einleitung
   - Gesetzliche Grundlage
   - Arten der Anonymität
2. 2. Anonymisierungsverfahren
   - Unterscheidung der Verfahren
   - Verfahren zur Informationsreduktion
   - Verfahren zur Informationsveränderung

Inhaltsverzeichnis

1. 1. Einleitung
   - Gesetzliche Grundlage
   - Arten der Anonymität
2. 2. Anonymisierungsverfahren
   - Unterscheidung der Verfahren
   - Verfahren zur Informationsreduktion
   - Verfahren zur Informationsveränderung
3. 3. Mikroaggregation
   - Grundidee
   - Arten der Mikroaggregation
   - Mikroaggregation in R
   - Bewertung der Mikroaggregation

Inhaltsverzeichnis

1. 1. Einleitung
   - Gesetzliche Grundlage
   - Arten der Anonymität
2. 2. Anonymisierungsverfahren
   - Unterscheidung der Verfahren
   - Verfahren zur Informationsreduktion
   - Verfahren zur Informationsveränderung
3. 3. Mikroaggregation
   - Grundidee
   - Arten der Mikroaggregation
   - Mikroaggregation in R
   - Bewertung der Mikroaggregation
4. 4. SAFE-Methode
   - Grundidee
   - SAFE-Methode

Gesetzliche Grundlage

Bundesstatistikgesetz (BStatG) §16 Geheimhaltung

- bis 1987 kaum möglich Daten weiterzugeben (Absatz 1)

- 1987 Überarbeitung des BStatG mit Einführung des Absatz 6
  - Möglichkeit Daten weiterzugeben neu geregelt
  - Weitergabe nur zu wissenschaftlichen Zwecken
  - Voraussetzung ist die faktische Anonymität der Daten

Arten der Anonymität

Absolute Anonymität

- Deanonymisierung der Daten unmöglich
- als Public Use Files (PUF) für jeden zugänglich
- Problem: es ist nie auszuschließen dass man die Daten Deanonymisieren kann

Arten der Anonymität

Faktische Anonymität

- Deanonymisierung der Daten nur möglich wenn Aufwand von Zeit,Kosten und/oder Arbeitskraft unverhältnismäßig groß ist
- Weitergabe der Daten nur an wissenschaftliche Institute
- Hauptziel: Verringerung der Zuordnungsmöglichkeit bei Erhalt des statistischen
  Informationsgehalts

Arten der Anonymität

Probleme:

- faktische Anonymität ist für jeden Datensatz unterschiedlich
  - Aufwand/Nutzen einer Deanonymisierung
  - Wer nutzt die Daten? (Statistische Ämter/Externe)
  - Welches Zusatzwissen kann erlangt werden?
- nicht jeder Datensatz eignet sich zur faktischen Anonymisierung
- Analysemöglichkeiten beschränkt

Arten der Anonymität

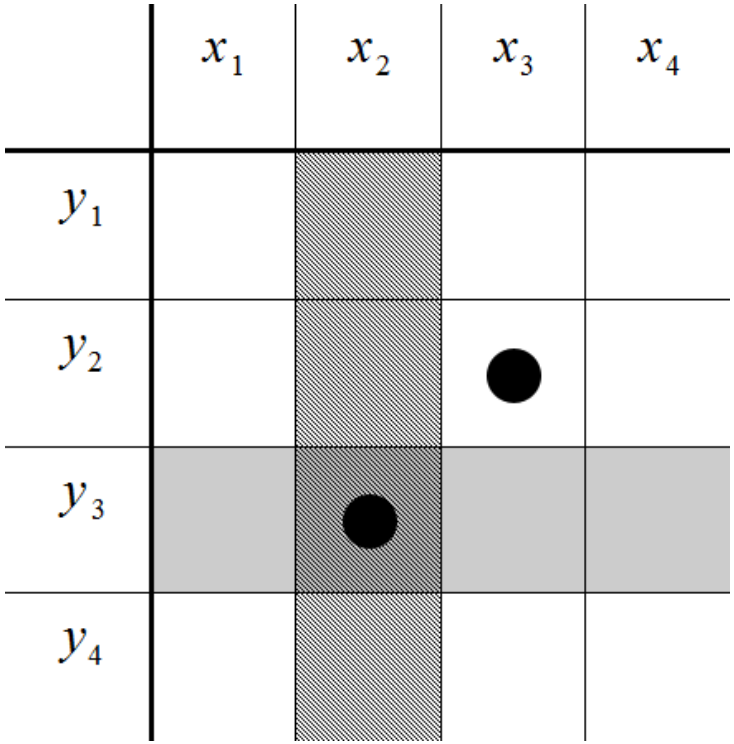
Formale Anonymität

- nur Entfernung der direkten Identifikatoren
- Merkmalsumfang, regionale und fachliche Gliederung bleiben erhalten
- Fernrechnen: Verarbeitung externer Syntax vor Ort

Unterscheidung der Verfahren

- Informationsreduktion oder Informationsveränderung
- metrische Variablen oder kategoriale Variablen
- alle Anonymisierungsverfahren verändern Einzeldaten

Unterscheidung der Verfahren



Bearbeitung...

- aller Merkmale
- eines Merkmales
- aller Merkmale eines Merkmalträgers
- von Einzelwerten

Merkmalsträgerbezogene Verfahren

- Entfernen auffälliger Merkmalsträger (Personen- und Haushaltsdaten)
- Systematische Einschränkung der Grundgesamtheit (Unternehmensdaten)
- (Sub-)Stichprobenziehung (Mikrozensus)

Merkmalsbezogene Verfahren

- Entfernen/Ersetzen/Zusammenfassen von Merkmalen
  - Konstruktion neuer Merkmale aus Kombination alter Merkmale
  - Bildung von Verhältniszahlen als Kennziffern
- Vergröberung von Merkmalsausprägungen
  - Bildung von Gruppen (zum Beispiel Umsatzklassen)
  - Rundung von metrischen Variablen
  - Zusammenfassung existierender Kategorien

Ausprägungsbezogene Verfahren

- Ersetzen von Einzelwerten mit fehlenden Werten
- Anwendung bei seltenen Werten oder seltenen Kombinationen von Werten

Verfahren für kategoriale Variablen

- Vertauschung von Merkmalsausprägungen zwischen Merkmalträgern (Swapping)
- diskrete Merkmale durch die Definition von Übergangswahrscheinlichkeiten randomisiert (Post-Randomisierung)

Verfahren für metrische Variablen

- Swapping
- Schätzung bestimmter Werte auf Basis von Regressionsmodellen (Imputationsverfahren)
- Addition/Multiplikation der Originalwerte mit Zufallszahlen (Stochastische Überlagerung)
- Erzeugung eines simulierten Datensatzes der die empirischen Verteilung der Originaldaten approximiert (Simulationsverfahren)

Ausprägungsbezogene Verfahren

- Klonen/Zerlegen von auffälligen Merkmalsträgern
- Beschränkung des Wertebereichs wobei Werte über/unter dem Bereich liegen auf das Maximum/Minimum des Bereichs festgesetzt werden (Censoring)
- Replacement-Verfahren

Grundidee

- Zusammenfassung von ähnlichen Werten zu Gruppen und Ersetzen der Werte durch das arithmetische Mittel der Gruppe
- Gruppe mit mindestens 3 Werten
- verschiedene Ansätze:
  - deterministisch: Zusammenfassung ähnlicher Werte
  - stochastisch: zufällige Gruppenbildung
  - Gruppen für alle Variablen gleich oder verschieden

Deterministische/Abstandsorientierte Mikroaggregation

Gemeinsame Mikroaggregation

- Daten werden nach einer dominierenden Variable absteigend sortiert und immer 3 Werte zu einer Gruppe zusammengefasst
- Daten werden nach einer Hilfsvariable sortiert
- Gruppen mit Hilfe der euklidischen Distanz gebildet

Deterministische/Abstandsorientierte Mikroaggregation

Getrennte Mikroaggregation

- Sortierung des Datensatzes nach zu gruppierender Variablen
- Zusammenfassen von 3 bis 5 Werten in Gruppen
- Vorgang für alle metrischen Variablen wiederholen

- Problem: in Bereichen mit hoher Dichte an Datenpunkten ist nur ein geringer Abstand zwischen den Mitteln benachbarter Gruppen

Deterministische/Abstandsorientierte Mikroaggregation

Gruppierte Mikroaggregation

- Gruppierung der Variablen nach Korrelation zwischen den Variablen
- gemeinsame Mikroaggregation innerhalb der Gruppen
- verbindet Vorteile der gemeinsamen und der getrennten Mikroaggregation

Modifikation der Verfahren möglich in dem man die Größe der Gruppen variabel gestaltet mit zwischen 3 bis 5 Werten.

Stochastische Mikroaggregation

Zufällige Mikroaggregation

- Gruppenbildung der Werte zufällig

Bootstrap-Mikroaggregation

- für jeden Merkmalsträger werden 2 weitere gezogen (mit Zurücklegen) und bilden eine Gruppe
- Durchschnittswerte der Merkmalswerte in der Gruppe ersetzen Werte des ersten Merkmalträgers

Umfang des Pakets

- Abstandsorientierte Mikroaggregation getrennt
- Abstandsorientierte Mikroaggregation aller Variablen nach dominanter Variable
- Abstandsorientierte Mikroaggregation aller Variablen nach euklidischer Distanz

- stochastische Mikroaggregation getrennt
- stochastische Mikroaggregation aller Variablen

Umsetzung in R

- Selektion der metrischen Variablen
- Anzahl der 3-er Gruppen feststellen sowie einer Restgruppe
- Art der Sortierung feststellen und Datensatz danach sortieren
- eine Schleife welche eine Spalte durchläuft
- bei getrennter Mikroaggregation wird nach jedem Spaltendurchlauf neu sortiert
- eine Schleife die den Spaltendurchlauf für alle Spalten durchführt
- Ausgabe des neuen Dataframes mit row.names
- bei dem Verfahren mit euklidischer Distanz zusätzlich eine Funktion nötig

Simulation

- Datensatz mit 100 Beobachtungen und 10 metrischen und einem kategorialen Merkmal
- metrische Variablen normalverteilt mit Standardabweichung von 1 bis 10, sowie mean von 0 bis 9
- Anwenden der 5 verschiedenen Verfahren
- Vergleich der Varianzen und der Mittelwerte

Mittelwerte

	row.names	m.original	m.no.single	m.no.allcor	m.no.alleuklid	m.yes.all	m.yes.single
1	num1	-0.01281906	-0.01281906	-0.01281906	-0.01281906	-0.01281906	-0.01281906
2	num2	1.04395907	1.04395907	1.04395907	1.04395907	1.04395907	1.04395907
3	num3	1.93164819	1.93164819	1.93164819	1.93164819	1.93164819	1.93164819
4	num4	2.67068283	2.67068283	2.67068283	2.67068283	2.67068283	2.67068283
5	num5	4.12752315	4.12752315	4.12752315	4.12752315	4.12752315	4.12752315
6	num6	4.59998973	4.59998973	4.59998973	4.59998973	4.59998973	4.59998973
7	num7	5.93820786	5.93820786	5.93820786	5.93820786	5.93820786	5.93820786
8	num8	6.68458182	6.68458182	6.68458182	6.68458182	6.68458182	6.68458182
9	num9	7.03525721	7.03525721	7.03525721	7.03525721	7.03525721	7.03525721
10	num10	11.41507630	11.41507630	11.41507630	11.41507630	11.41507630	11.41507630

Varianzen

	row.names	v.original	v.no.single	v.no.allcor	v.no.alleuklid	v.yes.all	v.yes.single
1	num1	1.146730	1.135982	1.135982	0.3285895	0.5137251	0.325223
2	num2	3.671252	3.638995	1.308920	1.1793689	1.2832140	1.389636
3	num3	8.266919	8.246458	3.259178	3.6884756	2.9564949	2.777798
4	num4	15.561478	15.522074	4.540812	8.1881601	6.2009003	3.654645
5	num5	23.384914	23.044277	5.279720	13.1700270	7.4916385	7.827814
6	num6	38.146176	37.840718	12.531344	26.6238145	13.7816364	11.197220
7	num7	45.281311	44.900956	12.620604	30.0960462	15.8306208	16.559459
8	num8	70.259969	69.225229	20.759951	52.0966777	17.9623194	20.250172
9	num9	72.066152	71.402029	27.568990	61.2155545	34.7896012	26.938230
10	num10	121.828433	120.494048	37.413097	104.0327817	31.1806004	41.288126

Mögliche Erweiterungen des Pakets

- Option für Variable Gruppengrößen
- Option zur Festlegung der Gruppengröße
- Varianzerhaltung
- gruppierte Mikroaggregation
- komplexere Ausgabe

Vor- und Nachteile

- arithmetische Mittel bleiben erhalten
- Verringerung der Varianz, da Varianzen innerhalb der Gruppen eliminiert werden
- führt zur Verzerrung von Tests

Lösungsansatz

- Gruppen aus 4 Werten
- 2 Werte erhalten den Gruppendurchschnitt minus die Standardabweichung
- 2 Werte erhalten den Gruppendurchschnitt plus die Standardabweichung
- Varianzen sowie Mittelwerte bleiben erhalten

Grundidee

- Ziel: Eliminierung des Fallzahlproblems, Randsummenproblems, Dominanzproblems und Zuordnungsproblems
- Lösung: Verallgemeinerung der Merkmalsträger und Veränderung einzelner Werte
- Kombination aus Mikroaggregation und Einzeldatenanonymisierung

Begriffsklärung

X^0 = Datenbestand der metrischen Werte der Originaldaten
Zeile i = statistische Einheit; Spalte j = Merkmal

Z_j^0 = Zuordnungsmatrizen der kategorialen Variablen
j = 1,2,...,k ; k = Anzahl der kategorialen Merkmale
Z_j^0bestehen aus n Zeilen unds_j Spalten für jede Ausprägung

Kombination von kategorialen Variablen:
$Z_{i,jl} = Z_{i,j} \otimes Z_{i,l}$

Begriffsklärung

$\mathrm{T}_j^0 = (Z_j^0)X^0$
T_j^0enthält die Summen aller metrischen Werte, die durch die Spalten der Zuordnungsmatrix definiert sind.

$A_j^0 = (Z_j^0)Z_j^0$
die Anzahl der Merkmalsausprägungen ist auf der Hauptdiagonale abzulesen

Beispiel

$$X^0 = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 4 \\ 0 & 7 & 2 \end{pmatrix} \qquad Z_1^0 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$T_1^0 = \begin{pmatrix} 4 & 3 & 4 \\ 0 & 7 & 2 \end{pmatrix} \qquad A_1^0 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

Ansatz

X^a und Z_j^a bestehen aus je maximal n/3 Zeilen

H^a ist eine Diagonalmatrix mit ganzen Zahlen größer 3, wobei die Dimension der Matrix der Anzahl der Zeilen von X^a entspricht.
Die Werte auf der Diagonale geben an wie oft die Zeile existiert.

$$T_j^a = (Z_j^a)H^a X^a$$
$$A_j^a = (Z_j^a)H^a Z_j^a$$

G_j^0 ist eine Matrix mit Dimension wie T_j^0
Gibt an ob ein Dominanzproblem im Tabellenfeld ist.

Ansatz

suche nach X^a , Z_j^a und H^a
T_j^a und A_j^a möglichst ähnlich zum Original

$$F_H = \min\left(\max_{j\in(1,t)}\left(\max_{i\in(1,s_j)}\left(\left| a^a_{j,i,i} - a^o_{j,i,i}\right|\right)\right)\right)$$

$$F_T = \sum_{j=1}^{t}\sum_{i=1}^{s_j}\sum_{l=1}^{m}\left| t^a_{j,i,l} - t^o_{j,i,l}\right|(1 - g^o_{j,i,l})$$

Probleme

- Bestimmung von metrischen Werten (X), Zuordnungsmatrizen (Z) für maximal n/3 Werte sowie die Häufigkeitstabllen (H)

- Problem 1: Minimierungsaufgabe nach 2 Kriterien (F)
- Problem 2: Dimension der Echtdaten und Anzahl der Kategorien

Automatisiertes Lösungsverfahren

1. Bearbeitung der kategorialen Variablen - Bestimmung der Z und H um F_H zu minimieren
2. Zuordnung von X zu den erzeugten Z
3. Bearbeitung der metrischen Variablen in X und Bestimmung von zulässigen Lösungen (die alle Geheimhaltungskriterien erfüllen)
4. Veränderung von X sodass F_T minimiert wird, aber nur Lösungen die in Schritt 4 akzeptiert werden
5. Gruppierung und Durchschnittsbildung (Mikroaggregation)

Ausblick

- Analysepotential
- Erweiterung des R-Pakets zur Mikroaggregation

Vielen Danke für ihre Aufmerksamkeit!

"**Gesetz über die Statistik für Bundeszwecke BstatG**"
Bundesministerium der Justiz, juris GmbH,
https://www.destatis.de

"**Anonymising business micro data - results of a German project**"
Rainer Lenz, Martin Rosemann, Daniel Vorgrimler, Roland Sturm
https://www.destatis.de

"**Datenzugang | Anonymität von Mikrodaten**"
Statistische Ämter des Bundes und der Länder
http://www.forschungsdatenzentrum.de/anonymisierung.asp

"**Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten - Statistik und Wissenschaft Band 4**"
Gerd Ronning, Roland Sturm, Jörg Höhne, Rainer Lenz,Martin Rosemann, Michael Scheffler, Daniel Vorgrimler

"SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben"
Jörg Höhne, Berliner Statistik Monatsschrift 3/03
https://www.statistik-berlin-brandenburg.de/publikationen/aufsaetze/2003/MS-BE$_2 00303 - 01.pdf$