**Vergleich von Data-Mining-Verfahren zur Outlier Detection**

Jonas M. Brauer, 01.05.2023

**Abstract**

Beim Data-Mining sammelt man viele Daten und untersucht diese u.a. auf Abhängigkeitsbeziehungen zwischen Variablen. Man versucht, Korrelationen zu entdecken, da sich hinter diesen eine mögliche Kausalität befinden kann. Dabei ist es störend, wenn ein „Rauschen“ in Gestalt von Ausreißern vorhanden sind, das die Sicht auf diese Dinge trübt. In einem anderen Fall will man gerade an diese Ausreißer heran, um außergewöhnliche Vorgänge zu entdecken. Dies kann z.B. die Bekämpfung von Kreditkartenbetrug sein, indem man in den Daten nach ungewöhnlichen Vorgängen sucht. Ob man jetzt Ausreißer loswerden oder genauer betrachten will, in beiden Fällen muss man sie irgendwie erkennen. Das ist bei Daten mit einer (univariat), zwei (multivariat) oder sogar drei Dimensionen auch von Menschenhand möglich, aber spätestens bei vieldimensionierten Daten muss, auch zur effizienten Automatisierung, ein Verfahren her. Davon gibt es einige, die über die Zeit hinweg entwickelt wurden. Doch sind sie alle gleich fähig? In diesem Vortrag geht es zu Beginn erst einmal um die Vorstellung einiger univariater statistischer Kennzahlen und Verfahren zur Ausreißer-Erkennung sowie im weiteren Verlauf um multivariate Verfahren und deren Vergleich untereinander. Am Ende folgt eine anschauliche interaktive Vorstellung einer Python-Implementierung zur Ausreißer-Erkennung. Im Zuge der Erarbeitung dieses Themas wurden relevante wissenschaftliche Paper zurate gezogen und sich ein breites Grundwissen in der Stochastik/Statistik angeeignet. Letzteres hat sich als sehr hilfreich herausgestellt, da viele Konzepte in den wissenschaftlichen Quellen aufeinander aufbauen und zum Verständnis komplexerer Sachverhalte vonnöten sind. Die univariate Statistik hat einige Konzepte/Kennzahlen, auf die in der multivariaten Statistik aufgebaut und welche dort verwendet werden. So ist hilfreich zu wissen, dass der Mittelwert und die empirische Standardabweichung einer Normalverteilung leicht von Ausreißern verfälscht werden können, es aber mit dem Median bzw. dem MAD (der Median aller Abstände zum Median) Ersatzkennzahlen gibt, die beide mit einem Breakdown-Value von 50% gegenüber einem Anteil von 50% an Ausreißern an den Gesamtdaten immun sind. Ein multivariates Verfahren zur Erkennung von Ausreißern verwendet die Mahalanobis-Distanz. Diese verwendet wiederum die nicht robusten Kennzahlen Mittelwert und empirische Kovarianzmatrix. Dadurch kann der Maskierungseffekt auftreten, bei dem durch einen verzerrenden Einfluss von Ausreißern auf die Mahalanobis-Distanz, tendenziell weniger Ausreißer erkannt werden. Abhilfe schafft ein Verfahren, dass die Robuste Distanz verwendet. Dieses nutzt auch Mittelwert und empirische Kovarianzmatrix, aber errechnet auf Grundlage mithilfe der MCD von Ausreißern bereinigten Daten. Die MCD ist die Minimum-Covariance-Determinante. Hierbei handelt es sich um einen robusten Schätzer multivariater Beobachtungen, der Ausreißer aussortiert, indem eine festgelegte Anzahl von Beobachtungen ausgewählt wird, deren Kovarianz-Matrix die kleinstmögliche Determinante besitzt. Die Robuste und die Mahalanobis-Distanz haben gemein, dass sie nur für unimodale, symmetrische Verteilungen verwendbar sind. Beide geben die Distanz von Datenpunkten zu dem Schwerpunkt einer Auswahl oder der gesamten Menge an Datenpunkten an. Bei multimodalen Verteilungen kann es mehrere Häufungspunkte und damit auch lokale Ausreißer, die sich eventuell im Schwerpunkt aller Datenpunkte, aber trotzdem außerhalb von Häufungspunkten, befinden. Diese würden von den distanzbasierten Verfahren nicht korrekt als Ausreißer identifiziert werden, da diese nur globale Ausreißer finden. Für diesen Fall gibt es dichtebasierte Verfahren, wie jenes, das den Local-Outlier-Factor verwendet. Dieser identifiziert Punkte als Ausreißer, die sich in einem Bereich geringer Punkte-Dichte befinden. Weiterhin gibt es ein ähnliches aber entfernungsbasiertes Verfahren, nämlich das der „k nächsten Nachbarn“. Hierbei wird für jeden Datenpunkt seine k-nächsten-Nachbarn und die einzelnen Entfernungen zwischen ihm und diesen bestimmt. Die n Punkte mit der größten Maximalentfernung werden als Ausreißer definiert. Ab wann ein Datenpunkt sich genügend von den restlichen Daten unterscheidet, um als Ausreißer zu gelten, unterscheidet sich bei den verschiedenen Verfahren. Es ist Definitionssache, was als Ausreißer gilt und nicht immer eindeutig. Zwischen einem Punkt der eindeutig als Ausreißer gilt und einem der eindeutig kein Ausreißer ist, gibt es oft unendlich mögliche Zwischenpunkte. Um Lage und Streuung Ausreißer-frei zu bestimmen, aber gleichzeitig Datenpunkte nicht diskret entweder in die Berechnung einzubeziehen oder auszulassen, kann die Stahel-Donoho-Outlyingness verwendet werden. Sie ermittelt für jeden Datenpunkt eine Gewichtung, um die „Ausreißerhaftigkeit“ von Punkten kontinuierlich zu bestimmen. Mithilfe dieser Gewichte, können aus den Datenpunkten anschließend ein gewichtetes Mittel und eine gewichtete Kovarianzmatrix berechnet werden.

**Literatur**

Zuriana, A. B., Rosmayati, M., Akbar, A., Mustafa, M. D.. (2006). „A Comparative Study for Outlier Detection Techniques in Data Mining“. IEEE, 1-4244-0023-6/06, S. 1-2

Hubert, M., Rousseeuw, P. J.. (2011). “Robust statistics for outlier detection”. John Wiley & Sons, Inc., WIREs Data Mining Knowl Discov, 2011, 1, 73–79, DOI: 10.1002/widm.2, S. 1-3.

Ismiguzel, Idil: Outlier Detection with Simple and Advanced Techniques: A tutorial on how to detect outliers using standard deviation, interquartile range, isolation forest, DBSCAN, and local outlier factor, in: Towards Data Science, 2022, https://towardsdatascience.com/detecting-outliers-with-simple-and-advanced-techniques-cb3b2db60d03, (abgerufen am 30.04.23)

Petrovskiy, M. I.. (2003). “Outlier Detection Algorithms in Data Mining Systems”. Programming and Computer Software, Vol. 29, No. 4, 2003, pp. 228–237. Translated from Programmirovanie, Vol. 29, No. 4, 2003. 0361-7688/03/2904-0228. S. 2-3

Hubert, M., Debruyne, M., Rousseeuw, P. J.. (2017). “Minimum Covariance Determinant and Extensions”. arXiv:1709.07045v1, S. 3-4.

scikit learn: sklearn.covariance.MinCovDet, in: scikit learn, o. D., https://scikit-learn.org/stable/modules/generated/sklearn.covariance.MinCovDet.html (abgerufen am 01.05.23)

Ming-Chao Guo, Shi-Jie Pan, Wen-Min Li, Fei Gao, Su-Juan Qin, Xiao-Ling Yu, Xuan-Wen Zhang, Qiao-Yan Wen. (2023). “Quantum Algorithm for Unsupervised Anomaly Detection”. arXiv:2304.08710v1, S. 1-2