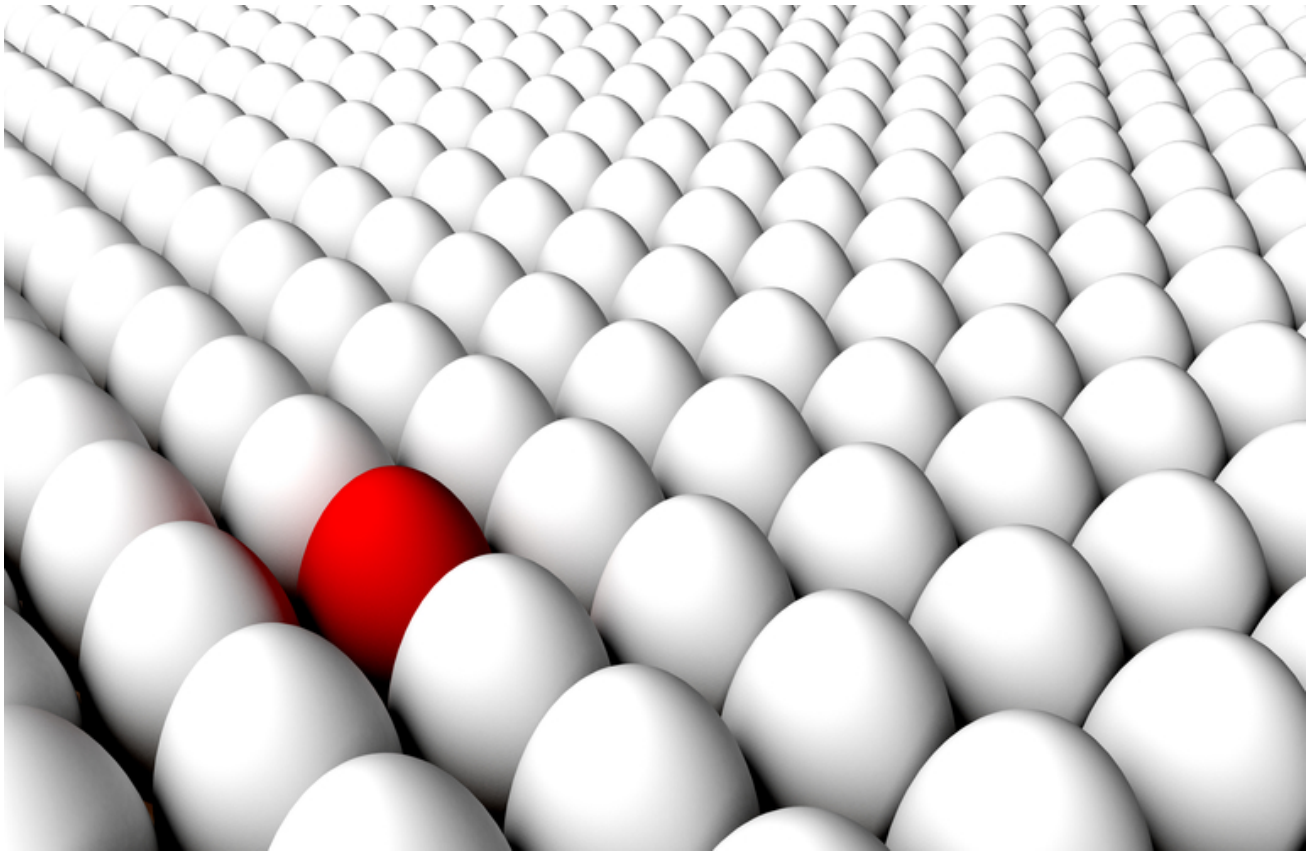
Software-Entwicklung

Warum nicht jeder Ausreißer ein Ausreißer ist

Anomalie-Erkennung mit Machine Learning

von Torben Ott



Eine Anomalie oder ein Ausreißer (engl. „outlier“) bezeichnet einen Datenpunkt, dessen Eigenschaften so stark von der Norm abweichen, dass sich der Verdacht ergibt, er würde durch einen besonderen Mechanismus generiert. Diese zugegebenermaßen etwas zirkuläre Definition zeigt bereits die Schwierigkeiten auf, die diesem Thema innewohnen. Doch für das unternehmerische Handeln sind gerade die anomalen Ereignisse enorm wertvoll: Von der betrügerischen Zahlungstransaktion und dem häufig ausfallenden „Montags“-Equipment auf der negativen Seite bis zum besonders zahlungsfreudigen Kunden auf der positiven Seite signalisieren Anomalien Handlungsbedarf für ein Unternehmen. Gerade in den von der Masse abweichenden Daten lassen sich die interessanten Geschäftsvorfälle finden und neue Erkenntnisse aufdecken.

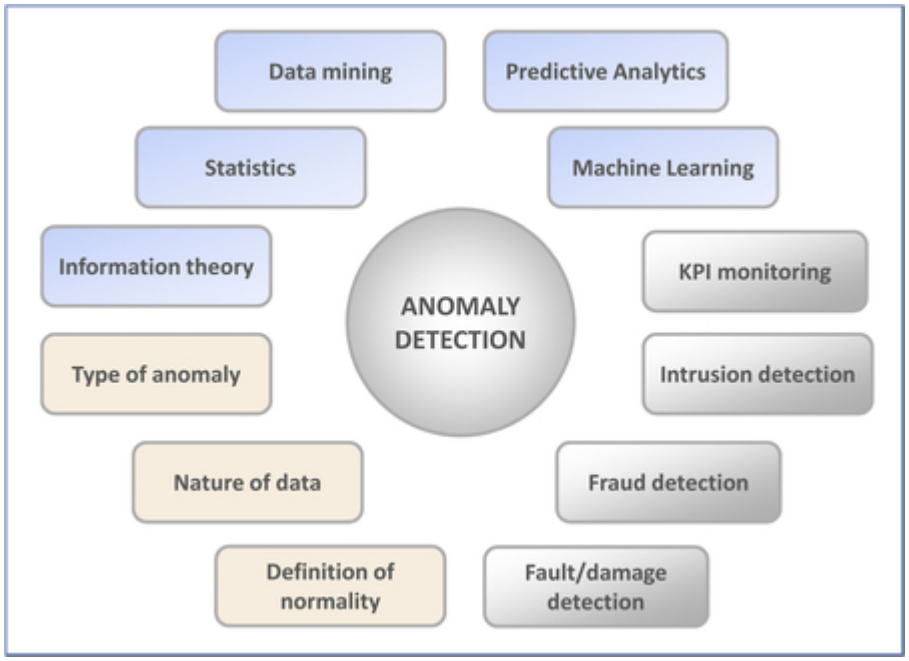


Abb. 1: Schlüsselelemente der Anomalie-Erkennung: Methoden (blau), Definitionen (orange) und Anwendungsgebiete (grau). Adaptiert aus [CBK09]

Bei kleineren Datenmengen und geringen Anforderungen an die Zeitkritikalität können derartige Überprüfungen gerade noch „per Hand“ durchgeführt werden. Große Datenmengen hingegen und schnelle Abfertigungszeiten erfordern die Unterstützung durch intelligente Algorithmen. Diese Algorithmen sind es auch, die Entscheidungen darüber, was als anomal gilt, objektivieren und damit einer quantitativen Analyse überhaupt erst zugänglich machen.

Dieser Beitrag soll einen Überblick über die verschiedenen Facetten dieser Thematik **(siehe Abbildung 1)** geben. Neben grundlegenden Eigenschaften von Daten-Anomalien und Anwendungsfällen der Anomalie-Erkennung werden auch geeignete Zugänge aus dem Bereich des Machine Learning vorgestellt. Hierbei werden sowohl Möglichkeiten als auch Grenzen der verwendeten Verfahren aufgezeigt.

Alle Dashboards auf Grün? Warum intelligente automatische Lösungen nötig sind

Neben den Begriffen „Anomalie“ und „Ausreißer“ ist im Englischen auch der Begriff *Novelty Detection* für die Erkennung außergewöhnlicher Datenpunkte geläufig und wird weitestgehend synonym verwendet [Pim14]. Im engeren Sinne geht es bei der Novelty Detection um den Vergleich eines Datenpunktes mit einer als normal bekannten Gesamtheit, während *Outlier Detection* die Identifikation von Ausreißern in einer gemischt normal/anomalen Gesamtheit bezeichnet.

Als Novelty Detection erfährt der Vorgang ein wenig mehr jene Wertschätzung, die ihm eigentlich zusteht: Darin steckt nicht nur die Abwehr ungeliebter Gefahren, sondern auch die Gelegenheit, den Wert von Daten für die Business Intelligence zu steigern und sie dem Entscheider oder der Entscheiderin zugänglich zu machen.

Die Daten, die für die Analyse zur Verfügung stehen, können naturgemäß vielfältig sein. Innerhalb des Bereichs der IT-Sicherheit wurde die Anomalie-Erkennung bereits sehr früh angewendet – gerade hier liegen große Datenmengen in Form von Logdaten vor. Ein Zugriff auf das IT-System aus dem Ausland zu ungewöhnlicher Zeit oder mit abweichendem Nutzungsmuster würde beispielsweise ein anomales Signal darstellen, das eine Analyse zutage fördern kann.

Eine weitere Facette der Anomalie-Überwachung findet man im Bereich des Internet of Things (IoT). Von Sensoren gelieferte Daten geben Auskunft über den Zustand von Maschinen, IT-Geräten und anderen Assets und erlauben die vorbeugende Wartung *(Predictive Maintenance)*. Auch das frühzeitige Reagieren auf von der Norm abweichende Zustände ist so möglich.

Eine ähnliche Methodik ergibt sich im Bereich der *Fraud Detection*. Dieser Methodik kommt vor allem im Zahlungsverkehr – aber nicht nur dort – eine zentrale Rolle zu. Ein illustratives Beispiel ist die gestohlene Kreditkarte. Das Kaufverhalten der kriminellen Person weicht in Bezug auf Umsatzhöhe, Ort und Frequenz so weit von der Norm ab, dass ein Alarm ausgelöst wird und der

betrügerische Umsatz verhindert werden kann. Auch im übrigen Dienstleistungsgeschäft ist das Erkennen von Betrugsversuchen eine wichtige Maßnahme.

Neben diesen „klassischen“ Anwendungsfällen ergibt sich mit der Digitalisierung sämtlicher Geschäftsprozesse auch zunehmend der Bedarf, Anomalien im Zeitverlauf betriebswirtschaftlich relevanter Key-Performance-Indikatoren (KPIs) zu erkennen. Hierzu zählen beispielsweise Einbrüche im Umsatz, Änderungen im Zahlungsverhalten von Kunden oder, im Online-Geschäft, auch sich verringernde Click-Through-Raten. Ein proaktives Reporting solcher Auffälligkeiten ergänzt die klassische Business Intelligence und kann Reaktionszeiten verringern.

Eine intelligente Automatisierung der Anomalie-Erkennung ermöglicht eine zunehmende Granularität der beobachteten Indikatoren. Verschiedene Kanäle, die in der klassischen BI lediglich en bloc betrachtet werden, sind so aufgeschlüsselt untersuchbar. Der Umsatz eines Unternehmens kann so beispielsweise auf der Ebene von Produktkategorien, Produkten und Verkaufskanälen überwacht werden – eine Aufgabe, die mit manueller Überprüfung nur unter großem Aufwand zu erfüllen ist.

Um die eingangs vorgestellte Definition einer Anomalie als „verdächtiger“ Datenpunkt zu konkretisieren, hilft im ersten Schritt ein Blick auf die verschiedenen Arten von Anomalien [CBK09]. Anhand dieser ersten Klassifizierung wird bereits deutlich, dass bei allen Anwendungen automatisierter Anomalie-Erkennung bereits vorab eine sorgfältige Definition der Zielsetzung unabdingbar ist.

Punktuelle Anomalie

Eine punktuelle Anomalie liegt vor, wenn ein einzelner Wert für sich allein stehend außergewöhnlich ist. Zu unterscheiden sind univariate und multivariate punktuelle Anomalien [Jol02]:

Eine **univariate Anomalie** äußert sich bereits in einer einzelnen Datendimension. Werden zum Beispiel die Schüler einer Grundschulklasse vermessen, so sticht eine Größe von 1,80 Metern hervor – der Verdacht liegt nahe, dass hier auch der Lehrer oder die Lehrerin vermessen wurde oder ein Aufzeichnungsfehler vorliegt. Dieser Datenpunkt wurde also durch einen anderen Mechanismus „generiert“. Univariate Anomalien fallen meist schnell auf, auch bei oberflächlicher Analyse.
Komplexere Anomalien äußern sich dagegen nur in der gemeinsamen Betrachtung mehrerer Dimensionen, sind also multivariat. Eine isolierte Betrachtung einer Dimension lässt derartige Ausreißer nicht erkennen. So würde bei der Vermessung der Schülerschaft einer Gesamtschule weder eine Größe von 1,70 m noch ein Gewicht von 25 kg sonderlich aus dem Rahmen fallen, da ein 10-jähriges Kind durchaus 25 kg wiegen kann und ein 16-jähriger Jugendlicher ebenso 1,70 m erreichen kann. Die Kombination beider Messungen (25 kg, 1,70 m) bei einem Kind dürfte dagegen nahezu ausgeschlossen sein. Eine derartige Messung wäre also eine multivariate Anomalie, deren generierender Mechanismus mit hoher Wahrscheinlichkeit ein Aufzeichnungsfehler ist.

Kontextuelle Anomalie

Eine weitere Anomalie-Art sind kontextuelle Anomalien. Diese Datenpunkte fallen erst auf, wenn sie in einem größeren Zusammenhang gesehen werden. Als Beispiel sei ein Fall aus der IT-Sicherheit genannt: Der Netzwerkverkehr eines Unternehmens schwankt erheblich zwischen Tag und Nacht. Ein hohes Datenvolumen, das während der Arbeitszeiten üblich, also normal, ist, kann nachts ein Indiz für den unautorisierten Zugriff auf Unternehmensdaten sein. Es liegt eine sicherheitsrelevante kontextuelle Anomalie vor.

Kollektive Anomalie

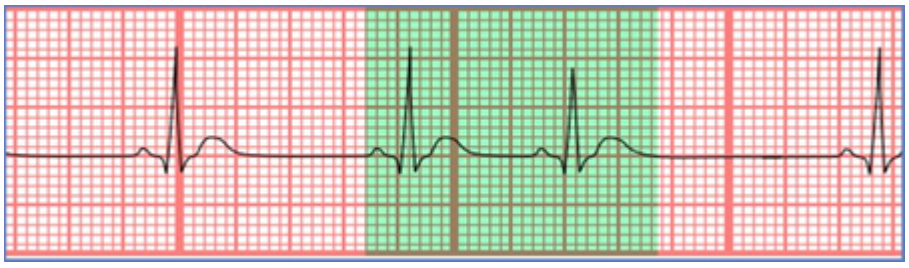


Abb. 2: Extrasystolen im Elektrokardiogramm sind kollektive Anomalien

Die letzte und vielleicht herausforderndste Art einer Anomalie befindet sich im Feld der sogenannten kollektiven Anomalien. Bei diesen sind einzelne Datenpunkte nicht auffällig. Stattdessen ergibt sich erst in der Betrachtung einer Datengruppe eine Auffälligkeit. Das Signal in **Abbildung 2** zeigt ein Elektrokardiogramm, bei dem jeder einzelne Herzschlag normal erscheint. Erst die Unregelmäßigkeit eines zusätzlichen Schlages (Extrasystole) definiert den abnormalen – wie es im medizinischen Kontext bezeichnet wird – Zustand.

Zeitreihen

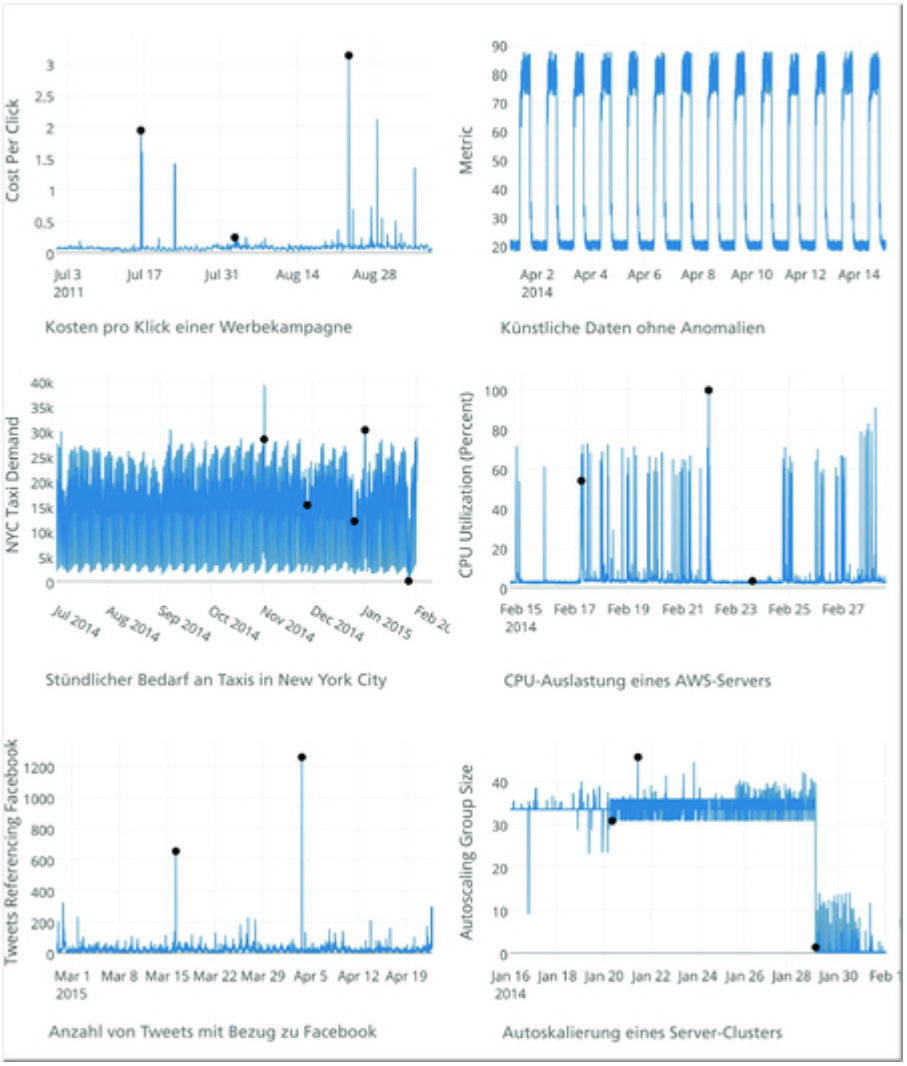


Abb. 3: Beispiele für Zeitreihen – Anomalien sind mit einem Symbol markiert (Quelle: [Sub17], modifiziert, Creative Commons Lizenz CC BY 4.0)

Eine besondere Stellung nehmen Zeitreihen ein, bei denen neben univariaten oftmals auch kontextuelle oder kollektive Anomalien vorliegen **(siehe Abbildung 3)**. Zum einen ermöglicht ein Fortschreiben der Zeitreihe in die Zukunft im Rahmen von Predictive Analytics das Erkennen zukünftiger Anomalien, bevor diese auftreten. Hierzu werden statistische Verfahren wie ARIMA oder Machine-Learning-Algorithmen wie neuronale Netze (LSTM) verwendet. Zum anderen können die vorhergesagten Werte und ihre Konfidenzintervalle als Definition von Normalität verwendet werden, um Anomalien beim Auftreten zu erkennen [Gup14].

Algorithmen für die Anomalie-Erkennung

Zur algorithmischen Identifizierung von Ausreißern stehen verschiedene Zugänge zur Verfügung, deren Auswahl neben den spezifischen Eigenschaften der Fragestellung auch von generellen Dateneigenschaften abhängig gemacht werden muss. Zu unterscheiden sind drei Situationen:

1. Neben normal gekennzeichneten Daten liegen auch bereits bekannte anomale Daten vor (überwachtes Lernen).

2. Es liegen nur als normal gekennzeichnete Daten vor, es sind aber keine Anomalien markiert (semi-überwachtes Lernen).

3. Es liegen nur ungekennzeichnete Daten vor (unüberwachtes Lernen).

Eine zweite Klassifizierung erfolgt anhand des algorithmischen Zugangs:

Bei **probabilistischen Verfahren** wird ein statistisches Modell an die Daten angepasst. Die Wahrscheinlichkeit, einen Datenpunkt aus diesem Modell zu generieren, wird bewertet, um Ausreißer zu erkennen. Die statistischen Modelle müssen hierbei geeignet gewählt und gegebenenfalls parametrisiert werden.

Abstands- und Dichte-Verfahren wie der k-NN-Algorithmus hingegen betrachten jeden Datenpunkt im Kontext seiner Umgebung bzw. der Ähnlichkeit zu anderen Datenpunkten. Liegt für eine Instanz eine hinreichend große Menge ähnlicher Daten vor, so bewertet das Verfahren den Datenpunkt als normal.

Nach einem ähnlichen Prinzip funktionieren **Clustering-Verfahren**, bei denen Machine-Learning-Algorithmen wie k-means verwendet werden, um die Daten in Gruppen einzuteilen. Instanzen, die von allen Gruppen weit entfernt sind, werden als Ausreißer identifiziert.

Klassenbasierte Verfahren setzen einen zumindest teilweise klassifizierten Trainingsdatensatz voraus (überwachtes oder semi-überwachtes Lernen). Ein Machine Learning Classifier wird mit den Trainingsdaten trainiert, um die Zugehörigkeit eines Datenpunktes zu einer Klasse vorherzusagen. Weit verbreitet sind One-Class Support Vector Machines (SVM), die eine Grenze zwischen Normalität und Anomalie ermitteln und somit auch als Domänenverfahren bezeichnet werden.

Bei den **Rekonstruktionsverfahren** und den **spektralen Verfahren** werden die Daten in eine niedrigere Dimensionalität überführt und so komprimiert. Instanzen, die in diesem Kompressionsprozess schlecht abgebildet werden können, gelten als Anomalie. Zu diesen Verfahren gehören die Hauptkomponenten-Analyse (PCA) und auch Replicator Neural Networks. Eine vergleichbare Methode ergibt sich in **informationstheoretischen Verfahren**, in denen Kenngrößen wie Entropie und Kolmogorow-Komplexität bewertet werden.

Einsatz der Algorithmen

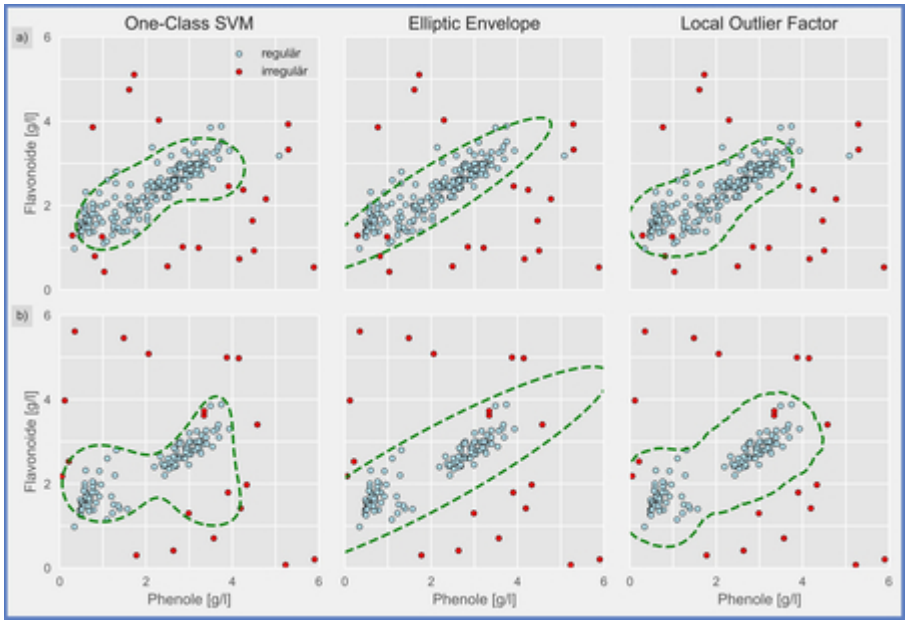


Abb. 4: Vergleich von Algorithmen zur Anomalie-Erkennung: klassenbasiert (One-Class SVM), probabilistisch (Elliptic Envelope) und dichtebasiert (Local Outlier Factor)

Die Auswahl eines geeigneten Zugangs zur Anomalie-Erkennung ist von vielen Faktoren abhängig. An dieser Stelle sollen daher einige illustrative Beispiele die Möglichkeiten der Algorithmen darstellen. **Abbildung 4** zeigt dazu die Phenolgehalte von Weinen aus drei verschiedenen Rebsorten [DKT17]. Auf der y-Achse sieht man den Gehalt an Flavonoiden, die vor allem für die Färbung

des Weins verantwortlich sind. Die x-Achse zeigt den Gesamtgehalt aller Phenole an. Diese haben einen weitreichenden Einfluss auf den Geschmack des Weins. Zusätzlich zu den gemessenen Datenpunkten sind zufällige Verunreinigungen dargestellt, die von den Algorithmen als Anomalie erkannt werden sollen.

Je ein Algorithmus aus dem Bereich der klassenbasierten, probabilistischen Verfahren und Dichte-Verfahren wurde verwendet, um einen Bereich der Normalität zu finden – **in Abbildung 4** durch die gestrichelte Linie markiert. Die Qualität der Algorithmen ist insgesamt vergleichbar hoch.

Für die Anomalie-Erkennung in Abbildung 4 standen hingegen nur die Daten zweier Rebsorten als Definition der Normalität zur Verfügung, die in zwei Cluster zerfallen. Bei dieser schwierigeren Datenlage zeichnet sich vor allem das Dichte-Verfahren als geeignet ab. Es ist zu beachten, dass in der Praxis meist höherdimensionale Daten vorliegen, die den Algorithmen eine klarere Trennung zwischen Normalität und Anomalie ermöglichen.

Fazit

Die in modernen Unternehmen entstehende Datenmenge und die damit verbundene Datengranularität machen den Einsatz von Machine-Learning-Algorithmen zunehmend attraktiv. Verbunden mit einer automatisierten Echtzeitanalyse bieten diese Ansätze einen spürbaren Mehrwert gegenüber klassischen BI-Tools und helfen, den Nutzen der Daten zugänglich zu machen. Die Auswahl und der Einsatz der richtigen Werkzeuge müssen aber auch bei automatisierten Verfahren sorgfältig erfolgen. Es gilt, den richtigen Grad der Granularität zu finden, bei dem Anomalien sicher erkannt und Fehlalarme gleichzeitig minimiert werden. Nur unter diesen Voraussetzungen kann eine automatisierte Lösung im Unternehmen Akzeptanz finden.

Literatur

[CBK09] Chandola, V. / Banerjee, A. / Kumar, V.: Anomaly detection: A survey. In: ACM Computing Surveys 41, 2009

[DKT17] Dua, D. / Karra Taniskidou, E.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, 2017, archive.ics.uci.edu/ml

[Gup14] Gupta, M. et al.: Outlier Detection for Temporal Data: A Survey. In: IEEE Transactions on Knowledge and Data Engineering 26, 2014

[Jol02] Jolliffe, I. T.: Principal Component Analysis. New York, Berlin, Heidelberg: Springer 2002

[Pim14]Pimentel, M. A. F. et al.: A review of novelty detection. In: Signal Processing 99, 2014, S. 215–249

[Sub17] Subutai, A. et al.: Unsupervised real-time anomaly detection for streaming data. In: Neurocomputing 262, 2017, S. 134–147

Dr. Torben Ott

ist Senior Consultant Data Science bei der Consist Software Solutions GmbH.

E-Mail: ott(at)consist.de

Bildnachweise:

Consist Software Solutions GmbH