KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Clustering: vom Sortieren bis zum Explorieren: 06_05Evaluation_Clustering

# Evaluation und Interpretation Clustering

Erarbeitet von
Dr. Katarina Boland



## Lernziele

- Du kannst gebräuchliche Metriken zur Messung der Qualität von Clusteringergebnissen nennen
- Du kannst interne und externe Evaluationsmetriken unterscheiden
- Du kannst den Silhouetten-Koeffizient berechnen
- Du kannst das F-Measure für Clustering berechnen



Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

- Du kannst die Best Practices für die Evaluation von Clusteringergebnissen zusammenfassen

# Inhalt

## Einstieg

Du hast gelernt, wie du mithilfe von Clustering Daten partitionierst.

Aber woher weißt du jetzt, wie gut deine Ergebnisse sind?

Vielleicht hast du unterschiedliche Verfahren und Parameter getestet oder mit der Anzahl der Cluster experimentiert. Welches Ergebnis ist denn nun das beste?

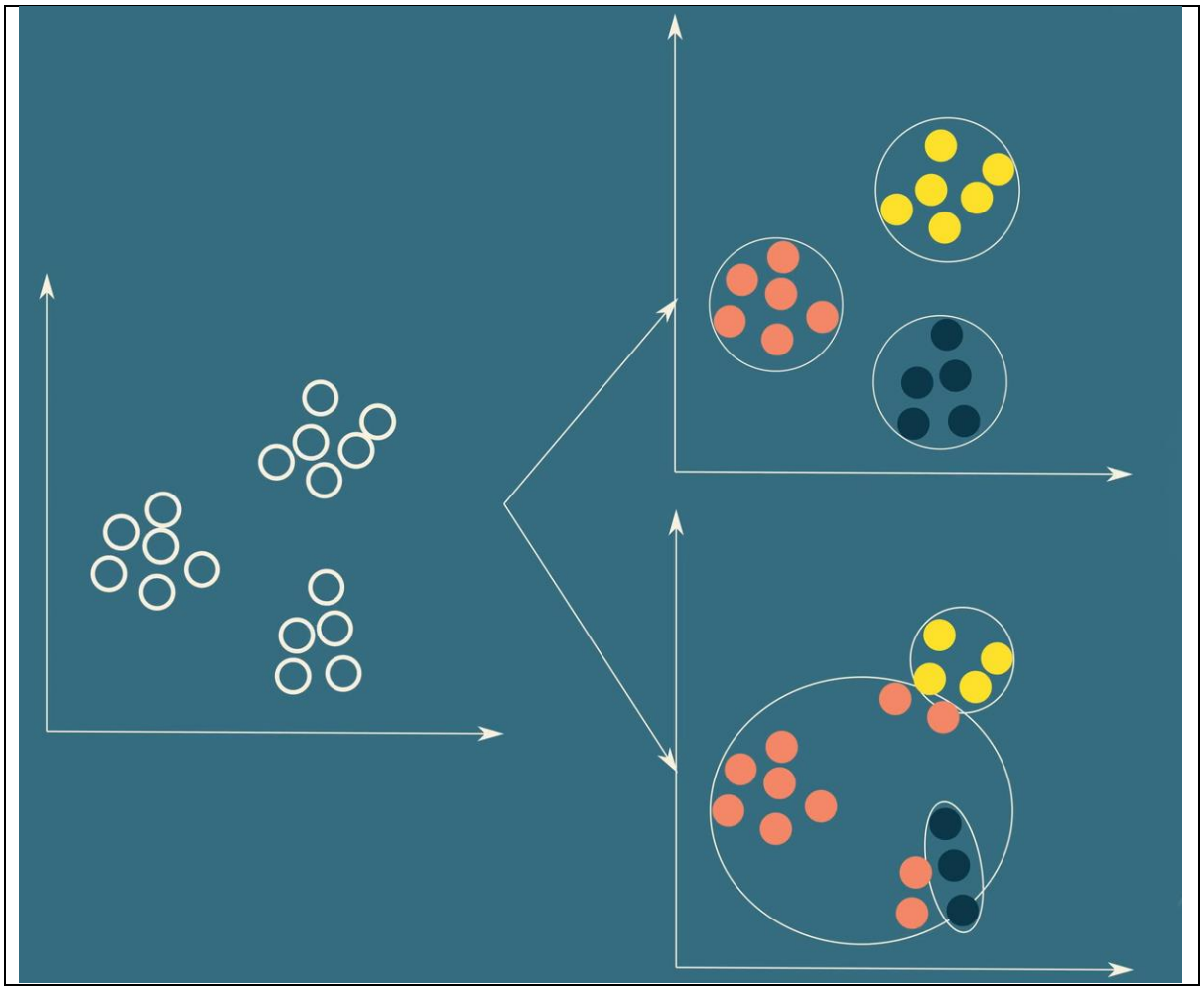
Und ist es ausreichend gut für deinen Anwendungsfall?

Clustering basiert auf unüberwachtem Lernen. Die Klassen, Zuordnungen und Ziellabels sind im Normalfall nicht bekannt und es wird daher in vielen Fällen auch keinen Goldstandard geben. Kann man denn die Qualität überhaupt messen, wenn kein Goldstandard zum Vergleich vorliegt? Oder kann man das unüberwachte Clustering doch nur dann sinnvoll einsetzen, wenn man zumindest für eine kleine Stichprobe einen Goldstandard erstellt?

## Evaluationsmaße

Die kurze Antwort ist: doch, man kann die Qualität auch ohne Goldstandard bewerten.

Vielleicht hast du eine Intuition, wenn du die Visualisierungen der beiden Clusterings betrachtest, welche von beiden die bessere ist.



Wir müssen also Qualitätsmaße finden, die diese Intuition abbilden und berechenbar machen.

Hierfür gibt es sogenannte „interne" Qualitätsmaße.

**Quelle [1]**

## Interne Qualitätsmaße

Eine Grundidee ist, dass gute Cluster sich dadurch auszeichnen, dass die enthaltenen Items einander sehr ähnlich, die Cluster also homogen sind, und gleichzeitig sehr unähnlich sind zu den Items in anderen Clustern, die Cluster also eine große Trennschärfe besitzen.

Um diese Aspekte zu messen, gibt es viele verschiedene Metriken. Einige Maße kombinieren beide Aspekte zu einem einzelnen Wert, zu diesen gehören der Calinski-Harabasz Index, der Silhouetten-Koeffizient und DBCV (density-based clustering validation). Wir illustrieren diese Klasse von Maßen am Beispiel des Silhouetten-Koeffizienten.

**Quelle [1]**
**Quelle [2]**
**Quelle [3]**
**Quelle [4]**

### Silhouetten-Koeffizient

Für jeden Datenpunkt wird die durchschnittliche Distanz zu allen anderen Punkten innerhalb desselben Clusters berechnet. Diese bezeichnen wir als A. Sie gibt die Homogenität an. Je kleiner der Wert, desto homogener ist das Cluster.

Zusätzlich wird die durchschnittliche Distanz zu allen Punkten in dem am nächsten liegenden, also dem ähnlichsten Cluster berechnet. Diese Distanz bezeichnen wir als B. Sie kennzeichnet die Trennschärfe. Je größer der Wert, desto trennschärfer ist das Cluster.

Der Datenpunkt bekommt einen Score zugewiesen, der sich berechnet aus der Differenz aus B und A geteilt durch den größeren Wert der beiden = (B-A)/max(B, A).

Die Qualität eines einzelnen Clusters ist der Durchschnitt der Scores aller Datenpunkte innerhalb des Clusters.

Der Gesamtwert für das Clustering ist der Durchschnitt der Scores aller Datenpunkte im gesamten Datensatz.

**Quelle [2]**
**Quelle [3]**

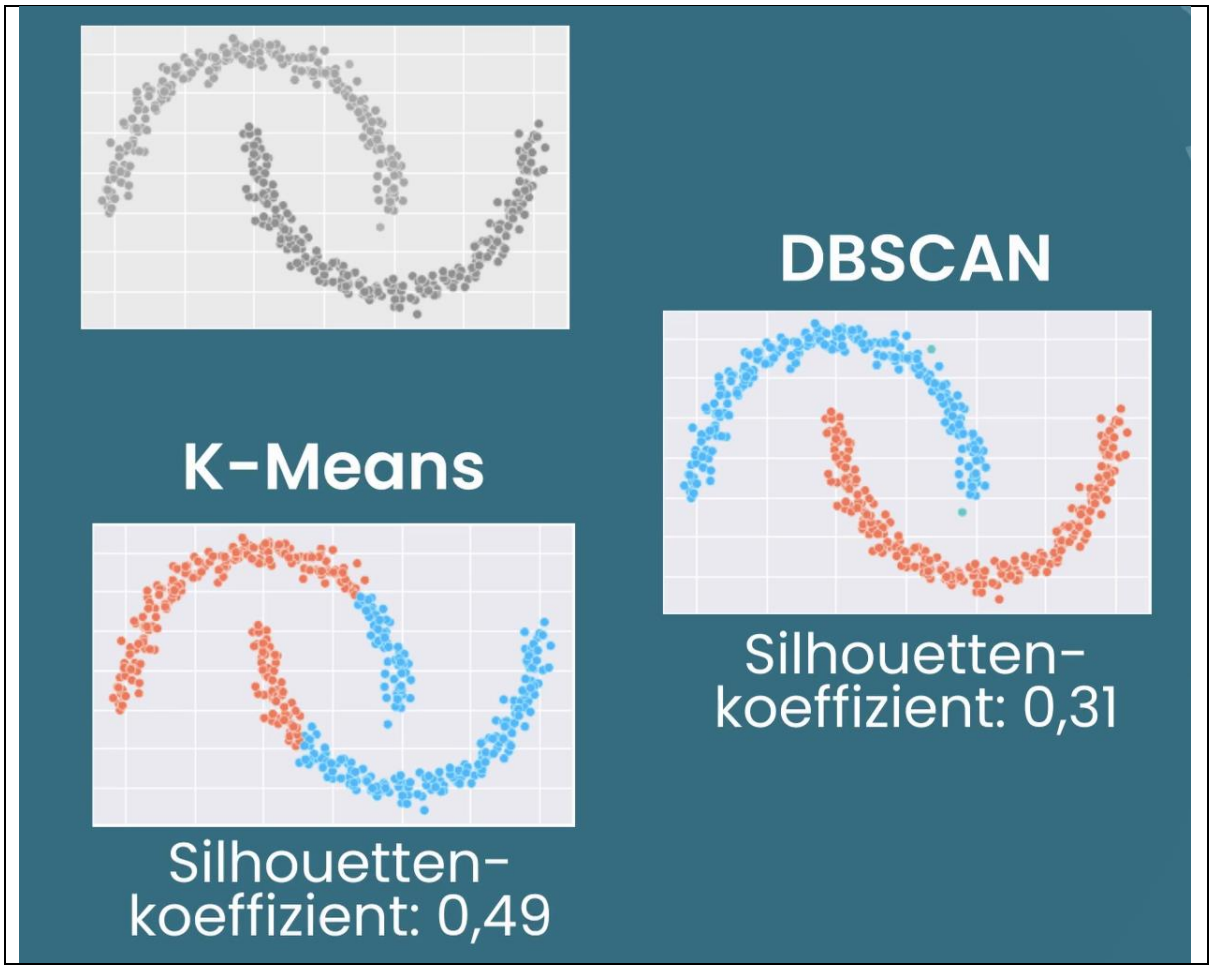
Der Silhouetten-Koeffizient kann Werte zwischen -1 und +1 annehmen. Je näher der Wert sich +1 annähert, desto homogener und trennschärfer sind die Cluster. Negative Werte deuten auf Instanzen hin, die sich näher an Instanzen in anderen als dem eigenen Cluster befinden und somit auf geringe Homogenität und Trennschärfe.

**Quelle [2]**

Wie der Name auch schon andeutet, ist der Silhouetten-Koeffizient geeignet zur Validierung von Clustern, die eine konvexe und im Idealfall kugelförmige Form haben.

Ist dies nicht der Fall, ist dieses Maß weniger gut geeignet. Density-basierte Verfahren wie HDBScan haben den Vorteil, dass sie Cluster mit beliebigen Formen bilden können. Für ihre Evaluation sollte der Einsatz von anderen Validierungsmethoden, wie beispielsweise DBCV, erwogen werden.

**Quelle [4]**
**Quelle [5]**
**Quelle [6]**



Beispiele für Werte des Silhouetten-Koeffizienten. Quellen: [5],[6]

**Quelle [5]**
**Quelle [6]**

### Externe Qualitätsmaße

Ist ein Goldstandard verfügbar, so können „externe" Qualitätsmaße eingesetzt werden, bei denen Vergleiche zu den wahren Clusterpartitionierungen und -zugehörigkeiten gezogen werden können. Hierzu gehören Precision, Recall und F-Measure, Rand Index, Normalized Mutual Information, der Fowlkes-Mallows Index und das V-Measure.

**Quelle [2]**
**Quelle [7]**
**Quelle [8]**
**Quelle [9]**
**Quelle [10]**

Stellvertretend für diese Klasse schauen wir uns hier das F-Measure an.

## F-Measure

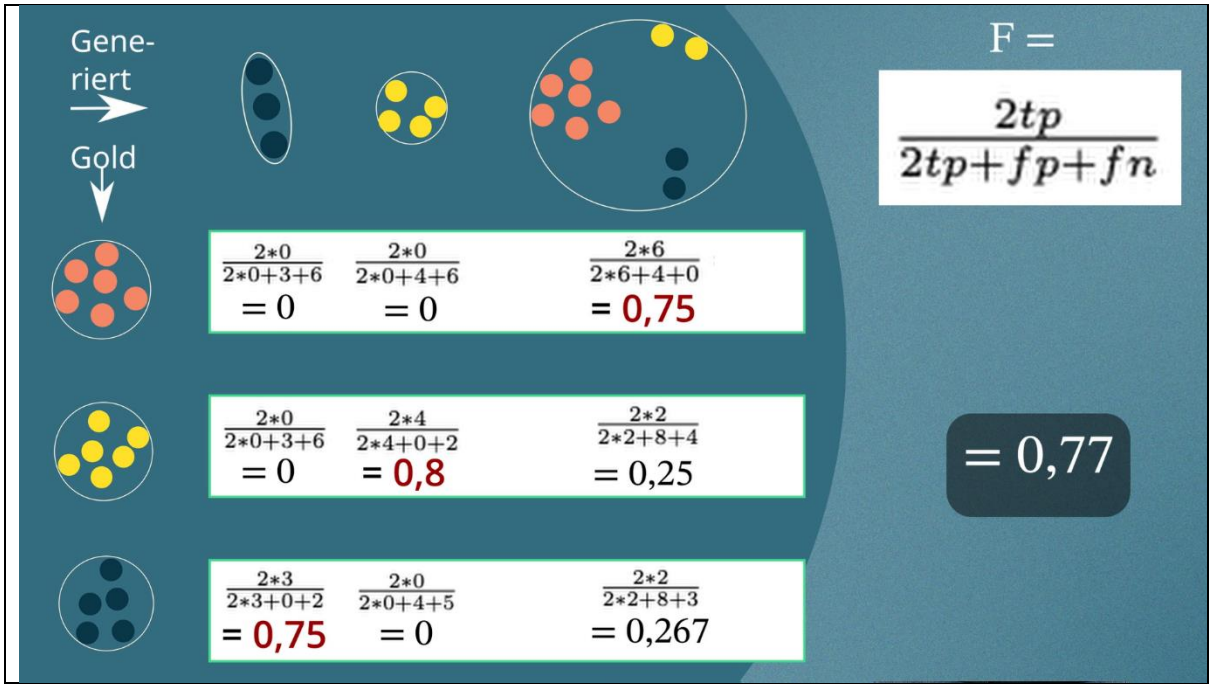
Das F-Measure ist eine Kombination aus Precision und Recall, also Korrektheit und Vollständigkeit von Klassifikationen. Wie wird dies aber nun auf Clustering übertragen?

Im Wesentlichen wird so getan, als wäre jedes generierte Cluster die Antwort auf eine Suchanfrage und die Cluster im Goldstandard die Menge an korrekten Ergebnissen. Da man nicht weiß, welches generierte Cluster welchem Cluster im Goldstandard entspricht, wird jedes generierte Cluster mit jedem Cluster im Goldstandard verglichen und die Paarung mit dem höchsten Score für die Berechnung des Gesamtscores verwendet.

Genauer gesagt wird für alle Elemente in den Goldclustern Precision und Recall berechnet in Bezug auf jedes der generierten Cluster.

Aus beiden Werten wird dann das F-Measure berechnet. Das F-Measure des Gesamtclusterings entspricht der gewichteten Summe aller dieser F-Measures für die einzelnen Goldcluster.

**Quelle [9]**




Auch hier liegt der Wert des F-Measures zwischen 0 und 1, wobei höhere Werte für eine bessere Qualität des Clusterings stehen.

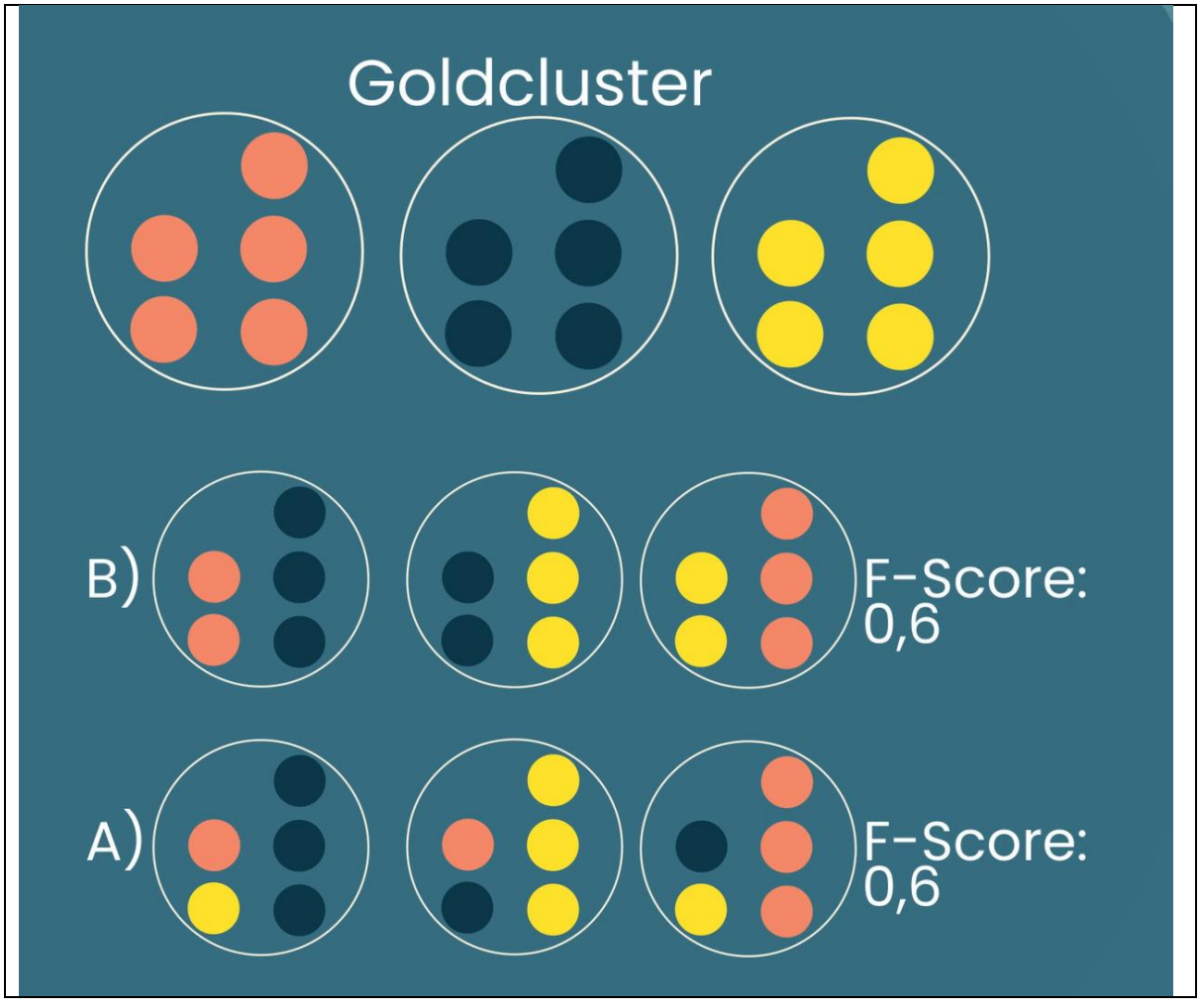
$$2 * \frac{precision*recall}{precision+recall} = \frac{2tp}{2tp+fp+fn}$$

Bei diesem Maß werden alle Instanzen gleichbehandelt. Sprich, wenn eine sehr unähnliche Instanz fälschlicherweise in ein Cluster einsortiert und die Varianz innerhalb des Clusters stark erhöht wird, zählt dieser Fehler nicht stärker als wenn eine relativ ähnliche Instanz, die geradeso außerhalb des Clusters liegt, fälschlicherweise in dieses einsortiert wird.

Dieses Maß bewertet vor allem die Vollständigkeit von Clustern und weniger stark die Homogenität.

Spielt die Homogenität eine wichtige Rolle, können also andere Qualitätsmaße die bessere Wahl sein, wie beispielsweise das V-Measure.

**Quelle [10]**



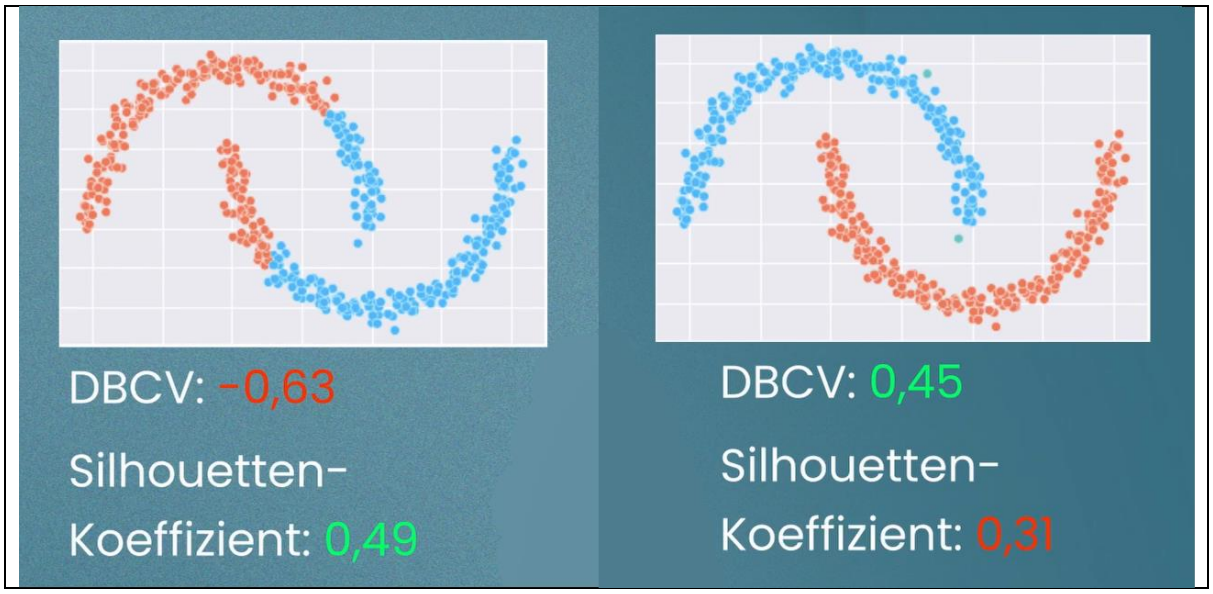

## Abschluss

Es mangelt nicht an verschiedenen Maßen zur Evaluation von Clustering. Der Grund dafür ist: Die Evaluation von Clustering ist alles andere als einfach. Was die „beste" Anzahl an Clustern ist und wo die Trennlinien zwischen den Clustern verlaufen sollen, ist oft subjektiv.

Hier kann es selbst unter Fachexpert*innen unterschiedliche Meinungen geben, gerade für Datenpunkte, die sich am Rand eines Clusters befinden.

Zudem ist es anwendungsabhängig, welche Kriterien die relevantesten sind.

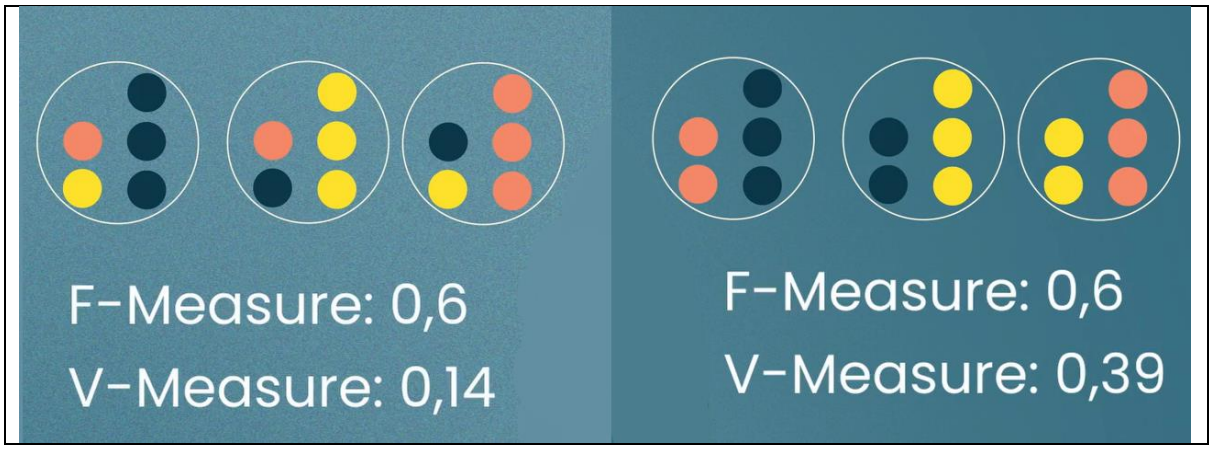
Die Evaluationsmaße bilden genau das ab: Verschiedene Maße gewichten unterschiedliche Kriterien unterschiedlich stark. Die Wahl der Evaluationsmetriken ist auch vom eingesetzten Clusteringverfahren abhängig.




Beispiele für Werte von Silhouetten-Koeffizienten vs. DBCV. Höhere Werte stehen für besseres Clustering. Quellen: [5],[6]

Somit sind sich auch verschiedene Evaluationsmaße nicht immer einig, welches Clusteringergebnis das beste ist.

Es ist daher gute Praxis, eine Kombination verschiedener Evaluationsmetriken zu verwenden um ein umfassenderes Bild von den generierten Ergebnissen zu bekommen.




Beispiele für Werte von F-Measure vs. V-Measure. Höhere Werte stehen für besseres Clustering. Adaptiert aus Quelle [10]

Zu Beginn steht oft die Visualisierung von Clustern und ein manuelles Prüfen der Plausibilität.

Die Auswahl des Evaluationsmaßes sollte sich dann nach dem Anwendungsfall, dem Clusteringverfahren und nach der Verfügbarkeit von Vergleichsdaten richten. Im Idealfall verwendest du mehrere Metriken, um die Qualität des Clusterings hinsichtlich verschiedener Aspekte bewerten zu können. Sofern du Clustering innerhalb einer Pipeline anwendest, also das Clustering nur ein Schritt in deiner Anwendung ist, kannst du auch eine extrinsische Evaluation in Erwägung ziehen:

Anstatt die Qualität des Clusterings direkt zu messen, kannst du messen, wie gut deine Gesamtanwendung unter Zuhilfenahme des Clusterings funktioniert. So würdest du also nicht die Qualität, sondern den Nutzen deines Clusterings messen, und auf den kommt es oft am Ende an.

## Quellen


Quelle [1] Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. 2010 IEEE International Conference on Data Mining, 911–916. https://doi.org/10.1109/ICDM.2010.35

Quelle [2] Han, J., Kamber, M., & Pei, J. (2012). 10 - Cluster Analysis: Basic Concepts and Methods. In J. Han, M. Kamber, & J. Pei (Hrsg.), Data Mining (Third Edition) (S. 443–495). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00010-1

Quelle [3] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Quelle [4] Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). Density-Based Clustering Validation. In Proceedings of the 2014 SIAM International Conference on Data Mining (SDM) (S. 839–847). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611973440.96

Quelle [5] Chawla, A. (2023, April 20). Intrinsic Measures for Clustering Evaluation. Abgerufen 27. August 2024, von https://blog.dailydoseofds.com/p/intrinsic-measures-for-clustering

Quelle [6] Daily-Dose-of-Data-Science/Machine Learning/KMeans-vs-DBSCAN.ipynb at main · ChawlaAvi/Daily-Dose-of-Data-Science · GitHub. (o. J.). Abgerufen 27. August 2024, von https://github.com/ChawlaAvi/Daily-Dose-of-Data-Science/blob/main/Machine%20Learning/KMeans-vs-DBSCAN.ipynb

Quelle [7] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. Journal of Intelligent Information Systems, 17(2), 107–145. https://doi.org/10.1023/A:1012801612483

Quelle [8] Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc.

Quelle [9] Zaki, M. J., & Jr, W. M. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms (1. Aufl.). Cambridge University Press.

Quelle [10] Rosenberg, A., & Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In J. Eisner (Hrsg.), Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (S. 410–420). Association for Computational Linguistics. https://aclanthology.org/D07-1043

## Disclaimer