Vergleich von häufigsten Clustering-Algorithmen auf Kaggle

Florian Merlau Friedrich-Alexander Universität
Erlangen, Deutschland
Email: florian.merlau@fau.de

***Zusammenfassung*—This document describes the most common article elements and how to use the IEEEtran class with LaTeX to produce files that are suitable for submission as part of the Big Data Seminar.**

I. Einführung

CLUSTERING-Algorithmen sind eine der zentralen Techniken in der Data Science und ermöglichen die Gruppierung von Datenpunkten basierend auf Ähnlichkeiten innerhalb der Daten. Diese Methode wird in einer Vielzahl von Anwendungen wie Mustererkennung, Datensegmentierung und der Entwicklung von Vorhersagemodellen eingesetzt. Die Auswahl des geeigneten Clustering-Algorithmus hängt dabei von mehreren Faktoren ab, darunter die Datenstruktur, der spezifische Anwendungsfall und die Skalierbarkeit des Algorithmus.

Auf Plattformen wie Kaggle, einer beliebten Umgebung für Data-Science-Wettbewerbe, wird eine Vielzahl von Clustering-Algorithmen eingesetzt, um reale Analyseprobleme zu lösen. Diese Wettbewerbe bieten eine wertvolle Gelegenheit, die aktuellen Präferenzen und Trends in der Anwendung von Clustering-Methoden zu untersuchen. Gleichzeitig stellt sich die Frage, ob die auf Kaggle bevorzugten Algorithmen auch in industriellen Kontexten eine ähnlich hohe Relevanz besitzen.

Ziel dieser Arbeit ist es, die am häufigsten verwendeten Clustering-Algorithmen in den letzten 20 Kaggle-Wettbewerben systematisch zu analysieren und zu dokumentieren. Dabei werden sowohl die spezifischen Anwendungsbereiche als auch die Unterschiede zwischen den Algorithmen untersucht. Ein besonderer Fokus liegt auf der Identifikation der Faktoren, die die Wahl eines Algorithmus beeinflussen, und der Evaluation, welche dieser Ansätze auch auerhalb von Kaggle bevorzugt werden.

Durch diese Analyse wird ein umfassender Überblick über die aktuellen Trends bei der Anwendung von Clustering-Algorithmen gewonnen, was sowohl für die Forschung als auch für industrielle Anwendungen von Bedeutung ist.

Dieser Beitrag entstand im Rahmen des “Big Data Seminar”s, das im Wintersemester 2024/2025 vom Lehrstuhl für Informatik 6 (Datenmanagement) der Friedrich-Alexander Universität Erlangen-Nürnberg durchgeführt wurde.

II. Grundlagen

A. Clustering

Clustering ist eine Methode des unüberwachten maschinellen Lernens, die darauf abzielt, Datenpunkte basierend auf ihren inhärenten Ähnlichkeiten in Gruppen, sogenannte Cluster, zu unterteilen. Dieses Verfahren ermöglicht es, in Datensätzen verborgene Strukturen zu identifizieren, ohne dass vorherige Labels oder Kategorisierungen erforderlich sind. Ein exemplarisches Clustering-Ergebnis ist in Abbildung 1 dargestellt. [1]

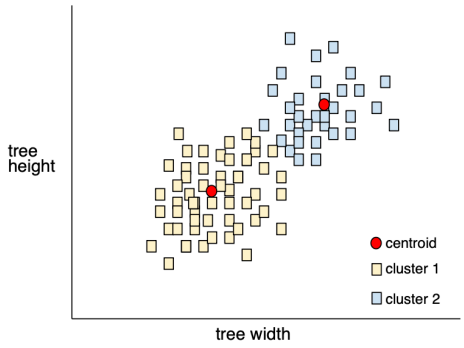

Abbildung 1. Beispiel für ein Clustering-Ergebnis, basierend auf der Darstellung von Google Developers [1]

B. Liste Clustering Algos

Es existiert eine Vielzahl von Clustering-Algorithmen, die je nach Datenstruktur und Anwendungsfall unterschiedliche Ansätze verfolgen. Zu den prominentesten zählen:

- **K-Means-Algorithmus**: Ein zentroidbasierter Ansatz, der die Daten in Cluster unterteilt, wobei jeder Cluster durch den Mittelwert seiner Datenpunkte repräsentiert wird. Der Algorithmus minimiert die Varianz innerhalb der Cluster und ist besonders effektiv bei konvexen Datenstrukturen [2].
- **Hierarchisches Clustering**: Dieser Ansatz erstellt eine hierarchische Struktur von Clustern, entweder agglomerativ (von einzelnen Punkten zu gröeren Clustern) oder divisiv (von einem groen Cluster zu kleineren). Dies ermöglicht die Analyse der Daten auf verschiedenen Granularitaetsebenen [3].
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Ein dichtebasierter Algorithmus, der

Cluster als Bereiche höherer Dichte definiert und in der Lage ist, Cluster beliebiger Form zu identifizieren sowie Ausreier effektiv zu erkennen [3].
- **Gaussian Mixture Models (GMM)**: Ein modellbasierter Ansatz, der die Daten als Mischung mehrerer normalverteilter Komponenten modelliert. GMM kann Cluster mit unterschiedlichen Formen und Gröen erfassen und bietet probabilistische Zugehörigkeiten der Datenpunkte zu den Clustern [4].
- **Mean-Shift-Clustering**: Ein nichtparametrischer Algorithmus, der versucht, die Dichtemodi in den Daten zu finden, indem er Datenpunkte iterativ in Richtung höherer Dichte verschiebt. Dies ermöglicht die Identifizierung von Clustern ohne vorherige Angabe der Anzahl der Cluster [3].

Die Wahl des geeigneten Clustering-Algorithmus hängt von den spezifischen Eigenschaften des Datensatzes sowie den Zielen der Analyse ab. Faktoren wie die Form und Dichte der Cluster, das Vorhandensein von Rauschen und Ausreissern sowie die Skalierbarkeit des Algorithmus spielen hierbei eine entscheidende Rolle.

III. Mehtodik

In diesem Abschnitt wird die geplante Methodik zur Analyse von Wettbewerben auf Kaggle im Hinblick auf Clustering-Algorithmen beschrieben. Die Methodik umfasst die folgenden Schritte:

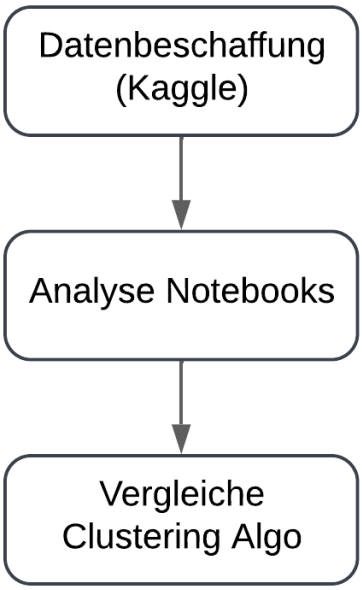

Abbildung 2. Ablaufdiagramm, eigene Darstellung

1) **Datenbeschaffung:** Extrahieren der letzten 20 Wettbewerbe auf Kaggle, wobei der Fokus auf solchen Wettbewerben liegt, die Clustering-Algorithmen beinhalten.
2) **Notebook-Analyse:** Analyse der zugehörigen Kaggle-Notebooks durch die Suche nach relevanten Schlagwörtern im Zusammenhang mit Clustering-Algorithmen. Dies umfasst Algorithmenamen (z. B. k-means, DBSCAN, hierarchisches Clustering) sowie verwandte Begriffe (z. B. Clustering-Bewertungsmetriken).
3) **Vergleich der Algorithmen:** Vergleich der identifizierten Clustering-Algorithmen im Hinblick auf ihre Unterschiede, Häufigkeit der Verwendung und Implementierungsdetails.

Zur Effizienzsteigerung und Automatisierung wird ein Python-Skript eingesetzt, das folgende Aufgaben übernimmt:

- Automatischer Download der Notebooks aus den identifizierten Kaggle-Wettbewerben.
- Parsen und Durchsuchen der Inhalte nach Clustering-relevanten Schlüsselwörtern.
- Aggregation und Organisation der Ergebnisse für die weitere Analyse.

Dieses Vorgehen bietet einen systematischen Rahmen, um Trends zu identifizieren und den Einsatz von Clustering-Algorithmen in realen Datensätzen und Wettbewerben zu bewerten.

Den Source Code dazu ist unter folgendem Link zu finden: https://github.com/XplorodoX/BigData

IV. Analyse der Clustering-Algorithmen

In den folgenden Abbildungen wird die Häufigkeit der eingesetzten Clustering-Algorithmen in verschiedenen Wettbewerben dargestellt.
Abbildung 3 illustriert die Häufigkeit der Algorithmen innerhalb einzelner Wettbewerbe. Es zeigt sich, dass bestimmte Algorithmen, wie beispielsweise der *RandomForestClassifier*, deutlich häufiger verwendet werden als andere.

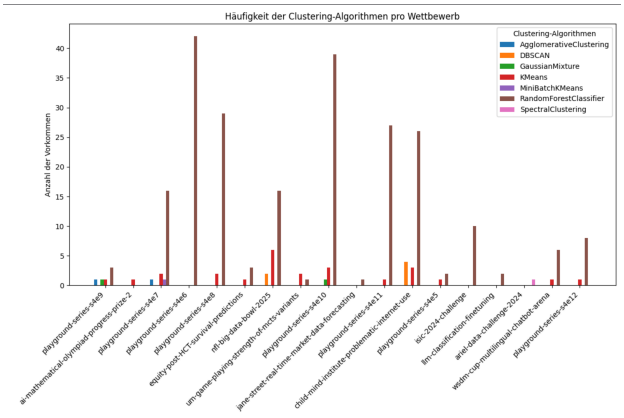

Abbildung 3. Häufigkeit der Clustering-Algorithmen pro Wettbewerb, eigene Darstellung.

Abbildung 4 stellt hingegen die Gesamthäufigkeit der Algorithmen über alle Wettbewerbe hinweg dar und gibt somit eine umfassendere Perspektive auf die generellen Präferenzen.

Der hufigste Algo war hier der RandoForestClassifier der mit ber 200 Notebooks am hufigsten vertreten war. Anschlieend war der KMeans mit knapp 40 Notebooks am zweirt

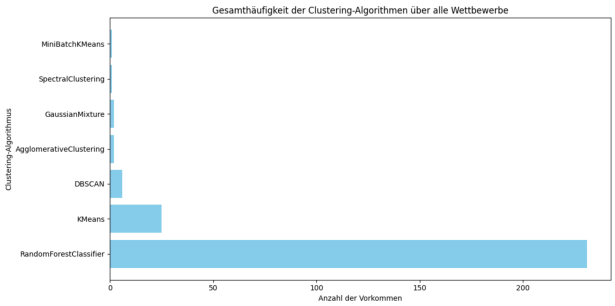

Abbildung 4. Gesamthäufigkeit der Clustering-Algorithmen über alle Wettbewerbe, eigene Darstellung.

hufigsten vertreten. DBSCAN gefolgt mit AgglomerativeClustering und zu guter Letst der Gaussian Mistrure Clustering Algo. Aus der Analyse ergibt sich, dass die fünf am häufigsten verwendeten Algorithmen die folgenden sind:

1) RandomForestClassifier
2) KMeans
3) DBSCAN
4) GaussianMixture
5) AgglomerativeClustering

Diese Algorithmen werden im folgenden Kapitel miteinander verglichen, um ihre Vor- und Nachteile sowie Einsatzmöglichkeiten näher zu beleuchten.

V. Vergleich der Algos

RandomForest:

VI. Fazit

Please Standby...

Literatur

[1] Google Developers, Clustering-Algorithmen — Machine Learning, verfügbar unter: https://developers.google.com/machine-learning/clustering/clustering-algorithms?hl=de
[2] FreeCodeCamp, 8 Clustering Algorithms in Machine Learning that All Data Scientists Should Know, verfügbar unter: https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/
[3] GeeksforGeeks, Clustering in Machine Learning, verfügbar unter: https://www.geeksforgeeks.org/clustering-in-machine-learning/
[4] D. Xu und Y. Tian, A Comprehensive Survey of Clustering Algorithms, Annals of Data Science, 2015.