Dichte-basierte Clusteringverfahren für große Datenmengen Teil B

Van Tinh Chu *HTW Berlin*
Berlin, Germany
s01567851@htw-berlin.de

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstrakt Aus Teil A von Independent Course haben wir schon drei Algorithmen(DBSCAN, HDBSCAN, OPTICS) probiert, um den besten Algorithmus finden zu können. Im Rahmen des Projekts erweitern wir weiterhin im drei Monat, um nach dem Ersatz der oben Algorithmus zu suchen, nämlich ein paar Algorithmen in diesem Bericht erforscht werden. Dabei werden das Verfahren und die Komplexität von Algorithmen auch geforscht, damit ein guter Überblick für die Forschung ausgeworfen wird. Dann wird ein Datensatz in 100 bis 200 Dimension mit 2 bis 5 Millionen Datenmenge geschaffen und untersucht, damit ein Algorithmus mit diesen Datenmenge innerhalb 5 Tage laufen kann.

Keywords: Clustering, density-based clustering, hierarchical clustering, dichbasierte Clustering,High dimension multiview data OptimizationSpark

1 Einleitung

Clustering spielt eine wichtige Rolle in dem Bereich Machine Learning. Mit Hilfe der machenden Cluserting werden Datenmenge in der getrennt Klasse sortiert. Aus diese Analyse wird die wichtige Information ausgezogen. Im IT Bereich gibt es mehre Algorithmen , um die Datenmenge zu klassifizieren. Trotzdem sollte ein suchender Algorithmus gefunden werden, damit das Problem der großen Datenmenge innerhalb 5 Tage gelöst werden, damit die Cluster im dieser Datenmenge entdeckt werden kann.

2 eine Umfrage der Clustering für große Datenmenge

Im Internet gab es schon ein paar Berichte über die Verarbeitung der Clusteringsprozess. Es ist jedoch unklare Ergebnisse für die riesig Datenmenge in der höhe Dimension. Deswegen sollte es klar sein,welcher Algorithmus besser verwendet werden sollte. Durch den Bericht [1] können wir erkennen, dass die Ergebnisse der Algorithmen in einer Tabelle eingefügt werden.

Clustering - Algorithment - Kategorien Damit die Verwirrung von Clustering-Kategorien vermieden werden kann, muss eine Überblick der Cluster ausgegeben werden, weil es nun viele Algorithmen für Clustering gibt. Zudem werden das Verfahren und Verarbeitung von jedem Algorithmus unter zusammengefasst.
Abbildung 1

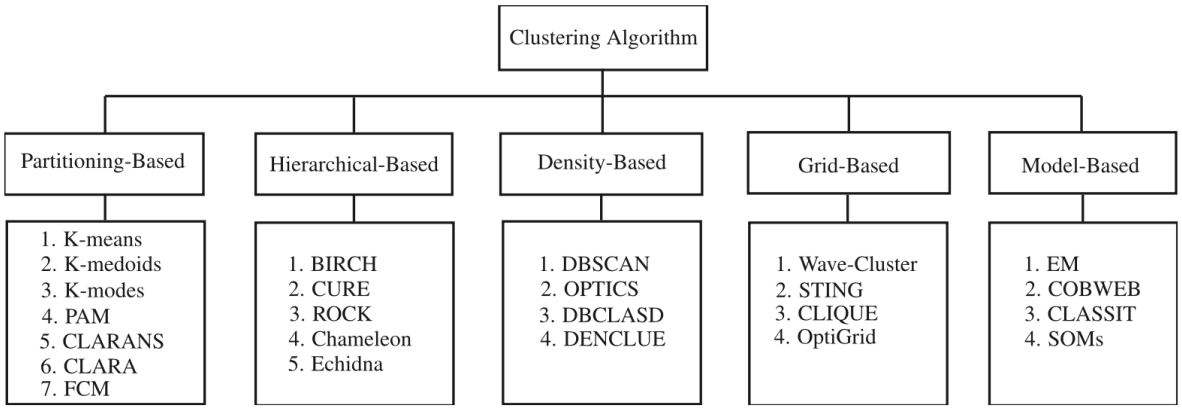

Abbildung 1: Eine Taxonomie von wichtigen Clus- teralgorithmen

- **Partitioning-based:**
  Hierarchical-ased:
- **Density-based:**
- **Grid-based:**
- **Model-based:**

[1]

3 5 häufigste Algorithmen für Clustering in Machine Learning

Es gab zurzeit verschiedene Algorithmen für den Process des Clustering [1]

4 K-Means Clustering

O(n) komplexitat k means ist noch langer als minibatch kmeans https://scikitlearn.org/stable/modules/generated/

5 Minibatch kmean

fragen auf stack over flow https://stackoverflow.com/questions/47270604/python-kmeans-clustering-for-large-datasets

he so la : $kmeans = MiniBatchKMeans(n_clusters = 100, random_state = 0, batch_size = 300, max_iter = 100).fit(X)3trrieu = 79s1trieu = 22s4trieu = 116.7942s(5000000, 100)5trieuElapsedtimetoclusterinMInibatchkmeans : 190.0804s$

6 ALl Sklearn

all clustering aus clutstering https://github.com/scikit-learn/scikit-learn/tree/0fb307bf39bbdacd6ed713c00724f8f871d60370/s

7 FUZZY-CMEANS (FCM) und EM

$https://github.com/scikit-fuzzy/scikit-fuzzy$

8 OptiGrid

Beispiel: Optigrid https://github.com/mexxexx/Optigrid/blob/master/optigrid.py

9 B. BIRCH

not good

10 C. DENCLUE

not gut ,1000 du lieu ma da bi die, va cai thang kia no tu viet chu ko phai cua bon sklearn

11 EXPECTATION-MAXIMIZATION (EM)

12 Zusammenfassung

[1] https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Literatur

[1] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras. A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 2(3):267–279, 2014.