\documentclass[13pt,a4paper,oneside]{article}

\usepackage[ngerman]{babel}

\usepackage[utf8]{inputenc}

\usepackage{amsmath}

\usepackage{amsfonts}

\usepackage{amssymb}

\usepackage{graphicx}

\usepackage[left=2cm,right=2cm,top=2cm,bottom=2cm]{geometry}

\usepackage{fancyhdr}

\usepackage{graphicx}

\usepackage{listings}

\usepackage{float}

\restylefloat{table}

%link

\usepackage{hyperref}

\author{Vantinh}

\title{ICW}

\date{\today}

\pagestyle{fancy}

\fancyhf{}

\lhead{Independent Coursework (ICW)- Master Angewandte Informatik-HTW Berlin-SS2020}

\rfoot{Seite \thepage}

\begin{document}

\begin{center}

{\Huge Dichte-basierte Clusteringverfahren für große Datenmengen}\\

Van Tinh Chu

\textit{HTW Berlin}\\

Berlin, Germany\\

s01567851@htw-berlin.de

\end{center}

\newpage

\tableofcontents

\newpage

**\section{Data Mining-Übersicht}**

Zurzeit leben wir in der Ära 4.0 mit der großen Daten, die immer im kommenden Jahr angestiegen werden würden. Der Prozess der Informationexplosion läuft weiter sehr schnell. Deswegen wurde eine Frage gestellt, was wir diese Daten erkunden werden.

In diesem Bericht präsentiere ich das Begiff "Dichte-basierte Clusteringverfahren für große Datenmengen", mit dem auf drei verschiedene Algorithmen vorgestellt werden.

Diese Algorithmen lassen sich schritt für schritt in nächsten Seiten beschreiben, indem die Funktion und auch die Komplexität von Algorithmen auf den Tisch legen, um Klarheit in der Globalisierungsdebatte zu schaffen, welche Algrithmus wir für die große Datenmengen verwenden sollen.

Bevor wir drei unterschiedliche Algorithmen eingehen, möchte ich die erste Punk des Begriff dichtebasiertes Clustering in Klar machen, dass \textit{Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in a data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. The data points in the separating regions of low point density are typically considered noise/outliers} ". \cite {DFCL} \\

Ziel des Clutering: \\

- Identifizierung des endlichen Datenmenge von Kategorien, Klassen order Gruppen.\\

- die Objekte sollten in einem Cluster gleichste sein.\\

- die andere Objekt in anderem Cluster muss am besten nicht gleich sein. \cite{ZIELCLUSTER}\\

- Vi du: chung ta co du lieu ve du lieu ve cac diem bus bahn hang ngay, bao nhieu luot nguoi su dung , sau do chung ta co the dung phuong phap nay de biet duoc , nhung diem nao la co mat do nguoi su dung lon, va nen dat cac diem bus bahn nhu the nao cho hop ly.

con rat nhieu vi du khac ve thoi tiet, noi chung la ve phan cum du lieu de chung ta co the dua ra cac nhan xet, danh gia phu hop voi tinh hinh.\\

Hien tai co rat nhieu thuat toan cho viec phan cum, nhu quy vi co the nhin thay o abbildung ben duoi, nhung trong noi dung bai bao cao nay, 3 thuat toan dbscann vs hdbscann va optimic se duoc trinh bay.

\begin{figure}[!htb]

\begin{center}

**\includegraphics**[width=\linewidth]{img/arten.png}

\end{center}

\caption{Eine Taxonomie von wichtigen Clusteralgorithmen \cite{Komplexitat2005}}

**\label{fig:boat1}**

\end{figure}

**\section{Dichte-basierte Clusteringverfahren für große Datenmenge}**

Es gibt viele Methode , um die Daten um Cluster zu klassfiziren. Zuerst wird der Algorithmus DBSCAN , HDBSCAN und OPTICS präsentiert. Dann wird alle Informationen zusammengefasst, wechler besser ist

%------------DBSCANN---------------DBSCANN--------------DBSCANN-------------DBSCANN------%

**\subsection{ Algorithmus DBSCANN}**

Thuat toan DBscan duoc phat trien boi cac nha khoa hoc Martin Ester ,Hans-Peter Kriegel, Jörg Sander und Xiaowei Xu phat trien va su dung chu yeu trong linh vuc Clusteranalyse. Thuat toan dua tren mat do cua cac diem va tao ra cac Cluster khac nhau.\\

**\subsubsection{der Prozess der Algorithmus DBSCANN}**

\begin{figure}[!htb]

\begin{center}

**\includegraphics**[scale=0.1]{img/DBSCAN.png}

\end{center}

\caption{DBScan Algorithmus}

**\label{fig:boat1}**

\end{figure}

Thuat toan nay co 3 tu khoa chinh , Đó là Kernopunkt, Grenzpunkt và Rauschepunkt va 2 parameter cua thuat toan la epsion va minimum punkte\\

- $\varepsilon$ la ban kinh vong tron, voi tam chinh la punkte do\\

- minimum punkte: la so punkte de co the tao thanh 1 cluster\\

sau do cac khai niem kernpunk , grenzpunk, rauschepunk se dua vao 2 parameter nay de quy dinh cac cluster.\\

Vi du chung ta cho $\varepsilon$ la 1, mini punk = 4 nhu hinh ben.\\

Kernopunkt (Core Point) là 1 diem , diem ma thoa man 2 dieu kien ben tren. se duoc coi la kernpunkt. \\

nhu hinh ben thi diem A se la Kernpunk vi no co 4 diem trong 1 vong tron, 3 diem lan can va chinh no nua.\\

- Grenzpunkt (Border Point)-thi lai co it nhat 1 diem la kernpunk la chinh no, nhung no lai ko du so luong punkt nhu dieu kien la 4, chinh vi vay no duoc coi la grenzpunk. \\

Nhu hinh ben thi diem B chinh la grenzpunkt, vi no co maximal la 2 diem, 1 diem ben canh va chinh no\\

- Rauschpunk (Noise Point): thi lai ko dat duoc mat do nhu dieu kien bai toan, vi du trong hinh la diem N . no chi co duy nhat 1 diem trong von ban kinh cho phep , nen duoc coi la diem Noise\\\cite{WIKIDB} \cite{YOUTUBEDB}\\

image\\

Chinh vi vay thuat toan se tim nhung diem kennpunkt va grenzpunkt gan nhau, de tao cluster, theo chieu huong nguoc lai thi cac diem rauschpunk se ko tao duoc cluster. \cite{YOUTUBEDB}\\

**\subsubsection{die Komplexität der Algorithmus DBSCANN}**

Nhu chung ta da biet , thuat toan nay co 2 tham so chinh, chinh vi vay khi chung ta thay doi thong so, chung ta se nhan duoc cac ket qua khac nhau.Sau day toi xin trinh bay ve uu va nhuoc diem cua thuat toan, tu do chung ta se thay duoc do phuc tap cua thuat toan\\

\textbf{- Vorteile:}

co the de dang dieu khien cac tham so, ra duoc ket qua nhu mong muon\\

co the output ra ko chi la hinh cau ma con hinh khac nua \cite{WIKIDB}\\, noi chungla ra nhieu hinh dang khac nhau\\

\textbf{- Nachteile als die Komplexität:} \\ do phuc tap cao, khi gap du lieu nhieu chieu. \cite {YOUTUBE\_DBSCAN\_VN}

- \textbf{die Komplexität} : $"$\textit{ selbst ist von linearer Komplexität. Jedes Objekt wird im Wesentlichen nur einmal besucht. Jedoch ist die Berechnung der $\huge{\epsilon}$ -Nachbarschaft im Regelfall nicht in konstanter Zeit möglich (ohne entsprechende Vorberechnungen). Ohne die Verwendung von vorberechneten Daten oder einer geeigneten Indexstruktur ist der Algorithmus also von quadratischer Komplexität.} $"$ \cite{WIKIDB}\\

- \textbf{ein unbewachstes Problem} Ngoai viec gap van de ve du lieu dua vao , cai ma co van de , leer hoac là bi sai, thi viec ko có du lieu cung la van de cho hoat dong thuat toan , kha nang hoc cua thuat toan \cite {Komplexitat2005}\\

\textbf{die zweite Komplexität} : no dung vong foreach 1 lan de duyet qua tat ca cac diem va se tra lai cac diem khi ket thuc vong lap . khong su dung index trong qua trinh xu ly, dan den do , vi du cac diem co khoang cach nho hon, thi no van dung vong lap nhu vay, dan den thoi gian cua thuat toan ko duoc cai thien nhieu.\cite {WIKIVN\_DBSCAN}

**\subsubsection{DBSCAN-Algorithmus Pseudocode}**

\begin{table}[!h]

\begin{lstlisting}[language=python]

regionQuery(P, eps)

return all points within P`s eps-neighborhood (including P)

\end{lstlisting}

\begin{tabular}[\linewidth]{ c c }

\begin{lstlisting}[language=Python]

DBSCAN(D, eps, MinPts)

C = 0

for each unvisited point P in dataset D

mark P as visited

N = D.regionQuery(P, eps)

if sizeof(N) < MinPts

mark P as NOISE

else

C = next cluster

expandCluster(P, N, C, eps, MinPts)

\end{lstlisting}

&

\begin{lstlisting}[language=Python]

expandCluster(P, N, C, eps, MinPts)

add P to cluster C

for each point P` in N

if P is not visited

mark P` as visited

N` = D.regionQuery(P`, eps)

if sizeof(N`) >= MinPts

N = N joined with N`

if P` is not yet member of any cluster

add P` to cluster C

unmark P` as NOISE if necessary

\end{lstlisting}

\\

\end{tabular}

\caption{Test environment details}

\end{table}

\begin{center}

\begin{flushleft}

D ist Datenmengen\\

eps: radius $\varepsilon$\\

minPts: minum Punkte \cite{WIKIDB}\\

\end{flushleft}

\end{center}

%-------------------HDBCANN-------------HDBCANN----------------HDBCANN---------------------%

**\subsection{ Algorithmus HDBCANN}**

**\subsubsection{der Prozess der Algorithmus HDBCANN}**

1 thuat toan tiep theo duoc gioi thieu do la dhdb scann, thuat toan nay duoc cac nha khoa hoc tiep tuc phat trien va ke thua tu thuat toan dbscann.

HDBSCAN is a clustering algorithm developed by Campello, Moulavi, and Sander.

thuat toan nay giup chung ta phan cum du lieu 1 cach khac la tot .

thuat toan nay cung cap 1 so parameter de chung ta co the custom duoc ket qua tra ve, sau day chung ta se nghien cuu ve no.\\

+ min cluster size: do lon cua cluster, so luong cua cac punkte trong 1 cluster. vi du: set la 15 thi so cum la 10, nhung khi set la 30 thi so cum se bi giam di con 5.\\

+ min samples : cai nay de than trong khi tao cluster, khi tang he so nay len thi cac rauschpunkte se it di\\

+ cluste selecting epsilon : trong truong hop chung ta can chon cluster nho hon, tap trung chu yeu vao object do thi ta co the thay doi he so nay,

cai he so nay thi phu thuoc vao khoang cach cua cac points,

\cite{HDBSCAN\_DOC}\\

+ alpha : cai nay giup chon cluster than trong hon (conservative = konservativ) , chinh xac hon

mac dinh dc set la 1.0 \\

Leaf clustering : cum la, ho tro them tham so ( cluster selection method).

mac dinh method la "eom", \\

+ allow single cluster :Trong truong hop nhan duoc nhieu cluster nho ,nhung co the cac cum nho do co gi dac biet, hoac co cau truc lon hon, chung ta co the su dung paramter nay va mac dinh cua thuat toan la khong tra ve single. \\

+ Outlier Detection : \\ tim kiem nhung diem khong phu hop, no giong nhu kieu noise point. THong qua thuat toan chung ta khong chi tim nhung diem ngoai duong bien, ma no co the phat hien nhung cluster ben trong , cai cung co nhung diem outlier.

- Predicting clusters for new points : No co 1 function approximate-predict() de co the thuc hien them points vao cac cluster , cai ma da duoc tao truoc do. Sau khi thuc hien call api nay thi cac diem se duoc xac dinh no la noise point order in a cluster.

- soft Clustering: la 1 tinh nang quan trong cua hdbscann . Vi du chung ta co 1 tap cluster tren du lieu 2 chieu. sau do chung ta co the generate soft clustering voi muc dich la xem dc nhieu thong tin hon cua cac diem noise points.\\

Ngoai ra chung ta cung co the get tat ca cac thong tin cua tat ca cac diem , cac diem ma chung ta da dua vao ban dau la dataset, dung function all points membership vectors(). Hon the nua, chung ta co the get soft clustering cua nhung diem, nhung diem chua nhin thay moi, chung ta co the su dung function membership vector() .

- wie wird hdbscan funktionieren ? khonng giong nhudb scann, hdbscan co the hoat dong ma ko can dua vao het so epsilon va so minimum points in cluster . cluster selection epsilon mac dinh la 0 .

**\subsubsection{wie eingentlich funktioniert HDBCANN}**

Es gibt ingesamt 5 Schritte , um die Daten zu Cluster zu kassfizieren.

**\paragraph**{Transform} Core cua thuat toan chinh la phuong phap phan cum lien ket don

cac diem noise se duoc xu ly de dua chung lai gan nhau hon.THuat toan nay cung dua vao 1 yeu to goi la core distance de xac dinh cum, theo cong thuc ben duoi day\\

$d\_{\mathrm{mreach-}k}(a,b) = \max \{\mathrm{core}\_k(a), \mathrm{core}\_k(b), d(a,b) \}$\\

Nhung diem ma khoang cach gan nhu nhau thi se dc giu nguyen, con cac diem ma khoang cach xa se bi day ra xa , cac diem bi day ra xa la cac diem noise point

trong hinh duoi cac kreis thi chung ta co the tinh duoc "khoang cach tiep can lan nhau", khoang cach nay la khoang cach giua 2 cluster, vi du blue cluster und green cluster

**\paragraph**{Build the minimum spanning tree of the distance weighted graph}

KHai niem co ban de lam la graph: o buoc nay chung ta co the dua vao weighted cua data , ma moi diem co a weighted graph and an edge points. Trong so(a weighted ) la khoang cach giua 2 diem .

sau khi chung ta thuc hien buoc nay thi chung ta co so lieu ve kha nang tiep can lan nhau giua cac cluster. Bay gio chung ta xem xet gia tri nguong, bat dau la 1 so lon, sau do se dc ha xuong dan dan, cac weight trong knoten se bi thay doi, khi do cac knoten bi thay doi, se tao ra 2 component , 1 loaij ket noi voi nhau 1 loai khong ket noi voi nhau theo cac nguong nhat dinh.

**\paragraph**{Construct a cluster hierarchy of connected components}

sau khi co cay do thi chung ta se tiep tuc tao ra cum phan cap du lieu , sap xep cac canh theo du lieu khoang cach ( tang ) va lap lai qua trih nay de merge cac cluster for each edge.

hdbscann co ho tro de tao ra dendiagramm is\\ $clusterer.single\_linkage\_tree\_.plot(cmap='viridis', colorbar=True)$

sau do chung ta se co 1 dendagramm nhu hinh ben duoi , \\

image o day\\

thong qua hinh tren chung ta nhan duoc diem dung lien ket don manh me,

DBSCAN cung co tao qua trinh nhu nay, ho dung 1 bien parameter de chon clusster, dieu nay co ve ko duoc khach quan so voi HDBscann, vi hbdscan co the tao dc cluster tot hon thong qua \textbf{lien ket don manh me (robust single linkage)}

**\paragraph**{Condense the cluster hierarchy based on minimum cluster size.}

trong phan nay thi chung ta se tiep tuc tim hieu cach ma hdscan tao ra cluster \\

truoc do chung ta co 1 param trong hdbscann do la \textbf{minimum cluster size} thuat toan se so sanh trong cluster do , rang no co so points it hon tham so chung ta dua vao hoac la nhieu hon, de tu do dua ra quyet dinh . Neu it so points so voi quy dinh thi cac diem do se bi day rangoai khoi cluster , nhung diem nay cung dc danh dau lai. Thuat toan se luot qua 1 luot Sau khi toan bo he thong phan cap va cuoi cung la tao 1 bieu do ngan hon.

sau do dung 1 funtion tiep theo de tao bieu do ngan gon hon voi it knoten hon. \\Tuy nhien lachieu rog cua cac line la bi thu hep honso voi buoc tren, vi cac diem ko phu hop da bi loai bo ra ngoai.

Imgage hier

\textbf{$clusterer\\_condensed\\_tree.plot()$}\\

**\paragraph**{Extract the stable clusters from the condensed tree.}

Dau tien la ta se tao ra 1 bien de so sanh voi khoang cach (distance), de xem xet su ben bi , on dinh cua cac cluster , cong thuc la : $\lambda = \frac{1}{distance} $. Dua vao cac cluster chung ta co khai niem duoc dinh nghia $ \lambda\_{birth} $ und $ \lambda^{death} $

$ \lambda\_{birth} $ la gia tri cluster khi bi split ra\\

$ \lambda^{death} $ la khi tro thanh 1 cluster riegg \\

$ \lambda\_{p} $ la gia tri cua moi diem (each point), la gia tri khi roi khoi cluster \\

thuat toan se tinh toan do on dinh theo cong thuc :\\

\textbf{stability} = $ \sum\_{p \in \mathrm{cluster}} (\lambda\_p - \lambda\_{\mathrm{birth}})$ \\

Sau do thuat toan se so sanh : \\

do on dinh cua child cluster > stability of cluster => set stability of child to cluster

if child cluster < stability of cluster => set clusster to to selected cluster und unselect all its descendants.\\ khi dat duoc wurzelknoten thi tra ve selected cluster va tao thah cluster cuoi cung. \newpage

**\subsubsection{die Komplexität der Algorithmus HDBCANN}**

+ HDB Scann la hieu qua hon dbscann , chung ta co the thay nhu hinh sau.

ve thoi gian thuc hie cua thuat toan thi nhu hinh chung ta co the thay. \cite {HDBSCAN\_SCIPY}\\

+ toi nghi la hdb scann la hieu qua hon vi nhieu chuyen gia noi vay, con do kho de thuc hien thi no phuc tap hon, nhung hieu qua thi tot hon.

\begin{figure}[!htb]

\begin{center}

**\includegraphics**[scale=0.1]{img/compare.png}

\end{center}

\caption{HDBSCANN effectiver als DBSCANN }

**\label{fig:boat1}**

\end{figure}

\begin{figure}[!htb]

\begin{center}

**\includegraphics**[scale=0.1]{img/numberofdata.png}

\end{center}

\caption{Number of datapoints }

**\label{fig:boat1}**

\end{figure}

%-------------------OPTICS----------------OPTICS-------------OPTICS-----------------------------%

\newpage

**\subsection{ Algorithmus OPTICS }**

\cite {YOUTUBE\_OPTICS}, \cite {WIKIOPTICS} für ganze Section

**\subsubsection{ Prozess der Algorithmus OPTICS}**

Optiics ist ordering poinst to indentify clustering structure\\

Der Algorithmus ist von Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel und Jörg Sander entwickelt worden. y tuong cua thuat toan cung bat dau tu dbscann , nhung duoc ke thua va phat huy tot hon . NO co the nhan biet duoc cac diem co mat do khac nhau, loai bo parameter $ \epsilon$.\\

Du lieu duoc sap xep tuyen tinh theo 1 chieu, theo tieu chi la Erreichbarkeitsdistanz cua cac diem , cac cluster dc ve nhu la 1 thung lung (taler), do do co the xac dinh duoc cluster ve sau .

dau tien cac cluster ban dau se duoc sap xep ordered theo kha nang tiep can (reachability).

\begin{figure}[!htb]

\begin{center}

**\includegraphics**[scale=0.2]{img/optics/visualiesierung.png}

\end{center}

\caption{Visualiesierung }

**\label{fig:boat1}**

\end{figure}

\begin{figure}[!htb]

\begin{center}

**\includegraphics**[scale=0.4]{img/optics/psecode.png}

\end{center}

\caption{ Pseudocode }

**\label{fig:boat1}**

\end{figure}

**\subsubsection{die Komplexität der Algorithmus OPTICS}**

Die Komplexität ist : Optics hat keine $\varepsilon$ . Deswegen co the coi $\varepsilon$ = $\infty$ .

So ist die Komplexität von Optics $ O(n^{2})$.Da Optics indexstructur benutzt, ist sie als $ O(n\cdot \log n)$ abgestiegen worden.\\

" Die „Erreichbarkeitsdistanz“ eines Punktes \textbf{p} von einem zweiten Punkt \textbf{o} ist definiert als $ \max(kerndistanz(o),dist(o,p))$ , also als das Maximum des echten Abstandes und der Kerndistanz des verweisenden Punktes. " \\

sau do optics se luu tru du lieu vao daten bank, cac diem bat ki chua dc xu ly truoc, sau do la nachbarn , sau do la cac diem co khoang cach tu thap nhat den tot nhat (Erreichbarkeitsdistanz). Qua trinh xu ly se giup Erreichbarkeitsdistanz cua cac diem chua duoc xu ly se tot hon. \\

\newpage

%------------thebibliography----------------thebibliography-------------thebibliography-------------%

\begin{thebibliography}{}

\bibitem{DFCL}

Sander J. (2011) Density-Based Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA, url: \url{https://doi.org/10.1007/978-0-387-30164-8\_211}

\bibitem{ZIELCLUSTER}

Skript zur Vorlesung,Knowledge Discovery in Databases,WS2007,Kapitel 5: Clustering, Slide 5, Ludwig Maximilians Universität München,Institut für Informatik, Lehr- und Forschungseinheit für Datenbanksysteme\\

url: \url{https://www.dbs.ifi.lmu.de/Lehre/KDD/WS0708/skript/kdd-6-outlier.pdf}

\bibitem{WIKIDB} Wikipedia DBSCAN,

\url{https://de.wikipedia.org/wiki/DBSCAN}

\bibitem{YOUTUBEDB}

Krish Naik, Published 05 Juni 2019, Accessed at 29 April 2020,\\ url: \url{https://www.youtube.com/watch?v=C3r7tGRe2eI}

\bibitem{Komplexitat2005}

Analyse dichtebasierter Clusteralgorithmen am Beispiel von DBSCAN und MajorClus , Michael Busch

Studienarbeit ,WS 2004/2005, Universitat Paderborn ,22. Marz 2005, url : \url{https://webis.de/downloads/theses/papers/busch\_2005.pdf}

\bibitem{HDBSCAN\_DOC}

Copyright 2016, Leland McInnes, John Healy, Steve Astels Revision 109797c7.\\

url: \url{https://hdbscan.readthedocs.io/en/latest/parameter\_selection.html#selecting-min-cluster-size}

\bibitem{HDBSCAN\_SCIPY}

High Quality, High Performance Clustering with HDBSCAN | SciPy 2016 | Leland McInnes, date 15 July 2016,

url: \url{https://www.youtube.com/watch?v=AgPQ76RIi6A}

\bibitem{WIKIVN\_DBSCAN}

SALEX KORSAN Founder and CEO of w3ki , URL: \url{https://www.vi.w3ki.com/data\_mining/dbscan.html}

\bibitem {YOUTUBE\_DBSCAN\_VN}

DBscan Algorithmus auf Vietnamesich, url: \url{https://www.youtube.com/watch?v=XKtsja3Kl34}

Code HDBSCAN: https://github.com/lmcinnes/hdbscan

%--------------optics--------------------%

\bibitem{YOUTUBE\_OPTICS}

OPTICS Algorithms | Clustering Algorithms | Clustering Analysis - ExcelR, ExcelR Solutions-Raising Excellence, 02.Juli.2018

\url{https://www.youtube.com/watch?v=zAbnJ7kERXk}

\bibitem{WIKIOPTICS}

Wikipedia ,05.May.2020,\url{https://de.wikipedia.org/wiki/OPTICS#cite\_note-1}

%--------end optics--------------------%

\end{thebibliography}

\end{document}