NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

KNOWLEDGE ★ CHARACTER ★ UNITY

Weather Plotting using Data Mining

On

Density Based Spatial Clustering Application with Noise

Submitted in partial fulfilment of the requirement for the award of Degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

ADITI GORDE	1NT19CS012
K M BHOOMIKA	1NT19CS091
SHARVIN PINTO	1NT19CS174
TEJASWINI MURARI	1NT19CS201

Under the Guidance of

VANI VASUDEVAN

DESIGNATION, Dept. of CS&E, NMIT

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

2020-2021

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

CERTIFICATE

This is to certify that the LA2 Report on **DBScan** is an authentic work carried out **ADITI GORDE(1NT19CS012), K M BHOOMIKA (1NT19CS091), SHARVIN PINTO(1NT19CS174), TEJASWINI MURARI(1NT19CS201)** bonafide students of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of *Bachelor of Engineering* in COMPUTER SCIENCE AND ENGINEERING of Visvesvaraya Technological University, Belagavi during the academic year *2019-2020.* It is certified that all corrections and suggestions indicated during the internal assessment has been incorporated in the report.

Internal Guide   **Signature of the HOD**

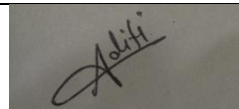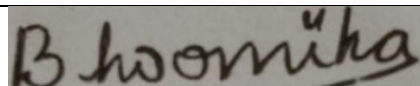Mrs. Vani Vasudevan
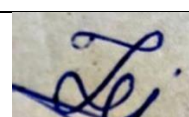Assistant Professor, Dept. CSE,
NMIT Bangalore

Dr. Sarojadevi H
Professor, Head, Dept. CSE,
NMIT Bangalore

DECLARATION

We hereby declare that

(i)   The project work is our original work
(ii) This Project work has not been submitted for the award of any degree or examination at any other university/College/Institute.
(iii) This Project Work does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
(iv) This Project Work does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
      a) their words have been re-written but the general information attributed to them has been referenced;
      b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.
(v)   This Project Work does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

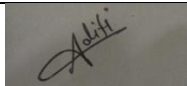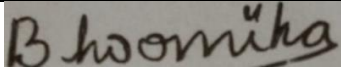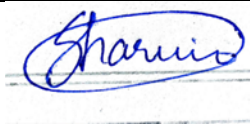| NAME	USN	Signature
ADITI GORDE	1NT19CS012	
K M BHOOMIKA	1NT19CS091	
SHARVIN PINTO	1NT19CS174	
TEJASWINI MURARI	1NT19CS201	

Date: 17-01-2022

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr. Sarojadevi H** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

I hereby like to thank our *Mrs. Vani Vasudevan, Assistant Professor*, Department of Computer Science & Engineering on her periodic inspection, time to time evaluation of the project and help to bring the project to the present form.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the project.

NAME	USN	Signature
ADITI GORDE	1NT19CS012	
K M BHOOMIKA	1NT19CS091	
SHARVIN PINTO	1NT19CS174	
TEJASWINI MURARI	1NT19CS201	

Date: 17-01-2022

<u>ABSTRACT</u>

Clustering algorithms are attractive for the task of class identification in spatial databases. However, the application to large spatial databases rises the following requirements for clustering algorithms: minimal requirements of domain knowledge to determine the input parameters, discovery of clusters with arbitrary shape and good efficiency on large databases. The well-known clustering algorithms offer no solution to the combination of these requirements. In this paper, we present the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it. We performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and real data of the SEQUOIA 2000 benchmark. The results of our experiments demonstrate that

(1) DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS, and that

(2) DBSCAN outperforms CLARANS by factor of more than 100 in terms of efficiency.

1. INTRODUCTION

1.1 MOTIVATION

Clustering algorithms are attractive for the task of class identification. However, the application to large spatial databases rises the following requirements for clustering algorithms:

(1) Minimal requirements of domain knowledge to determine the input parameters, because appropriate values are often not known in advance when dealing with large databases.

(2) Discovery of clusters with arbitrary shape, because the shape of clusters in spatial databases may be spherical, drawn-out, linear, elongated etc.

(3) Good efficiency on large databases, i.e., on databases of significantly more than just a few thousand objects.

DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it. It discovers clusters of arbitrary shape. Finally, DBSCAN is efficient even for large spatial database

1.2 PROBLEM DOMAIN

Weather forecasting is one of the most prominent topics that has been challenging scientists and engineers due to an expected increase of weather events and climate changes around the world.

Weather forecasting is the application of science and technology to predict the state of the atmosphere for a given location. Human beings have attempted to predict the weather since ancient times and started developing scientific forecast models since the nineteenth century. Nowadays, forecasters use more advanced methods and technologies to gather weather data, along with the world's most powerful computers. With the ability to launch satellites and supercomputers and harvest data from IoT devices, the new arrivals are advancing in information-gathering capabilities. Also, the use of data analytics techniques, as well as using machine learning, artificial intelligence and cloud-based warning systems; for example a system that indicates an airline company when to reschedule flights to avoid thunderstorms or a farmer when to irrigate a particular row of crops. These models can be programmed to predict how the atmosphere and the weather will change. Despite these advances, weather forecasts are still often incorrect. Weather is extremely difficult to predict because it is a complex and chaotic system.Weather forecasting contributes to the social and economic welfare in many sections of the society. Weather forecast provides vital information to a wide range of fields: agriculture, aviation, energy, commerce, marine, advisories, insurance companies, etc. It can also significantly influence decision and policymaking, global food security, construction planning, productivity and environmental risk management (Wiston, 2018).

Weather forecasting is often used to predict and warn about natural disasters that are caused by abrupt change in climatic conditions. Catalyzed by climate change, extreme weather is an increasing liability to the economy, with approximately 10 weather and climate disasters costing

1.3 AIM & OBJECTIVES

Weather plotting is a crucial science that helps in determining any future changes in the climatic conditions. The probablity of snow and hail reacting the surface can be determined with the use of latitudes. It also aids in identifying the thermal energy from the sun that is exposed to a region. The use of weather forecast plays an important role in the farming and agriculture. It also assists in the process of cultural operations, food grain transportation and implementation of livestock protection initiatives.

Distribution of Mean Temperature
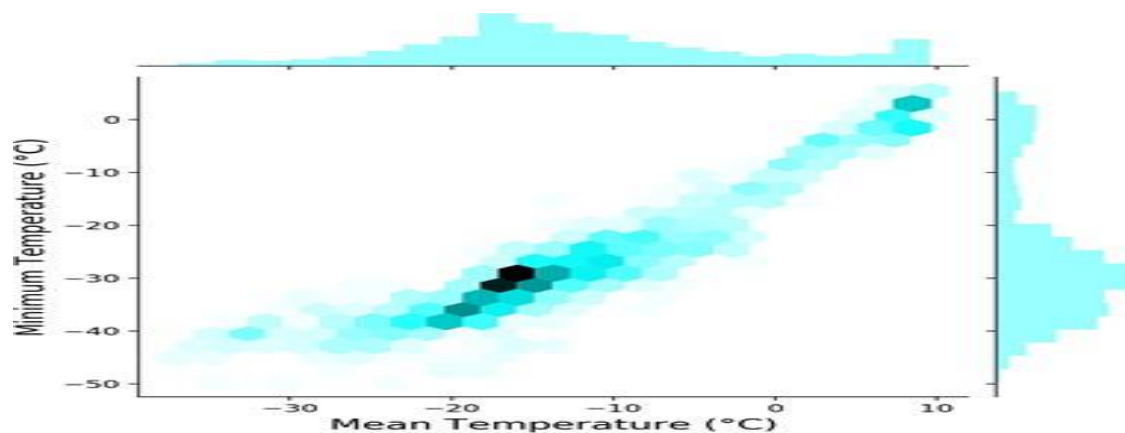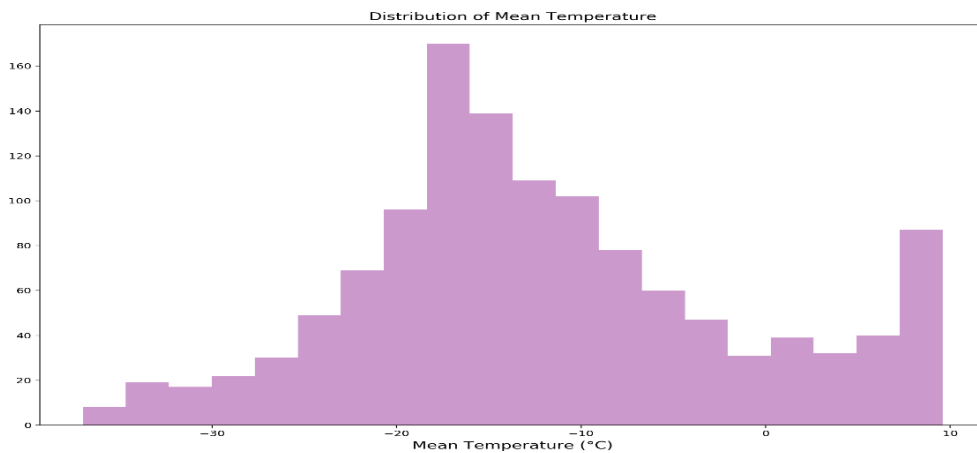
2. DATA SOURCE AND DATA QUALITY

2.1 Dataset Used

Data set used is Canada weather data for the year 2014 to cluster weather station. Data set consists of 1341 rows and 25 columns.

After dropping the rows containing NaN values in the above mentioned columns, we are left with 1255 samples. We received a loss in data around ~2.3%.

Latitude and Longitude are converted to x/y map projection coordinates using the command below

xs, ys =my_map(np.asarray(weather_df.Long), np.asarray(weather_df.Lat))

So there are 25 columns and now we need to learn more about the Columns, The names that are difficult to guess

- Stn_Name === Station Name
- Prov === Province
- Tm === Mean Temperature (°C)
- Tn === Lowest Monthly Minimum Temperature
- Tx === Highest Monthly Maximum Temperature
- DwTm === Days Without Valid Mean Temperature
- DwTx === Days Without Valid Maximum Temperature
- DwTn === Days Without Valid Minimum Temperature
- D === Mean Temperature Difference from Normal
- S === Snowfall (cm)
- DwS === Days Without Snowfall
- S%N === Percent of Normal Snowfall
- P === Total Precipitation (mm)
- DwP === Days Without Valid Precipitation
- P%N === Percent of Normal Precipitation
- Pd === No. of days with precipitation 1mm or More
- BS === Bright Sunshine days
- DwBS === Days Without valid Bright Sunshine
- BS% === Percent of Normal Bright Sunshine
- HDD === Degree Days Below 18°C
- CDD === Degree Days Above 18°C
- Stn_No === Station Number; Climate Station Identifier (1st 3 Digits==Indicate drainage basin, Last 4 Digits Sorting Alphabetically)

2.2 Data Pre-processing

Appropriate parameters "Eps" and "MinPts" of each cluster and at least one point from the respective cluster. Then, we could retrieve all points that are density-reachable from the given point using the correct parameters.

Let d be the distance of a point p to its k-th nearest neighbour, then the d-neighbourhood of p contains exactly k+l points for almost all points p. The d-neighbourhood of p contains more than k+l points only if several points have exactly the same distance d from p which is quite unlikely. Furthermore, changing k for a point in a cluster does not result in large changes of d. This only happens if the k-th nearest neighbours of p for k= 1,2, 3 are located approximately on a straight line which is in general not true for a point in a cluster.

DBSCAN needs two parameters, Eps and MinPts. However, our experiments indicate that the k-dist graphs for k > 4 do not significantly differ from the 4-dist graph and, furthermore, they need considerably more computation. Therefore, we eliminate the parameter MinPts by setting it to 4 for all databases (for 2-dimensional data). We propose the following interactive approach for determining the parameter Eps of DBSCAN :

- The system computes and displays the 4-dist graph for the database.
- If the user can estimate the percentage of noise, this percentage is entered and the system derives a proposal for the threshold point from it.
- The user either accepts the proposed threshold or selects another point as the threshold point. The 4-dist value of the threshold point is used as the Eps value for DBSC
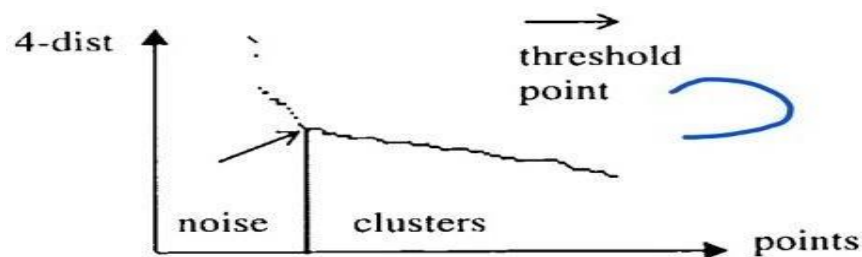
3. METHODS AND MODELS

3.2 DATA MINING ALGORITHMS

In the following, we present a basic version of DBSCAN omitting details of data types and generation of additional information about clusters:

DBSCAN (SetOfPoints, Eps, MinPts)

// SetOfPoints is UNCLASSIFIED

 ClusterId := nextId(NOISE);

FOR i FROM 1 TO SetOfPoints.size DO

Point := SetOfPoints.get(i);

 IF Point.CiId = UNCLASSIFIED THEN

      IF ExpandCluster(SetOfPoints, Point, ClusterId, Eps, MinPts) THEN

   ClusterId := nextId(ClusterId)

  END IF

 END IF

END FOR END; // DBSCAN

    ExpandCluster (SetOfPoints, Point, CiId, Eps,

    MinPts) : Boolean;

    seeds : =SetOfPoints. regionQuery (Point, Eps)

    IF seeds.siz<MinPts THEN// no core point

    SetOfPoint. changeCl Id (Point, NOISE)

    RETURN False;

    ELSE // all points in seeds are density-

    // reachable from Point

```
SetOfpoints. changeCiIds ( seeds, C1 Id)
seeds .delete (Point);
WHILE seeds <> Empty DO
currentP := seeds.first()
result := setofPoints.regionQuery(currentP, Eps)
IF result.size >= MinPts THEN
FOR i FROM 1 TO result.size DO
 resultP := result.get(i)
 IF resultP. CiId
 IN (UNCLASSIFIED, NOISE} THEN
 IF resultP.CiId = UNCLASSIFIED THEN
seeds, append (resultP)
END IF;
SetOfPoints. changeCiId ( resultP, CiId)
END IF; // UNCLASSIFIED or NOISE
END FOR ;
END IF; // result.size >= MinPts
seeds, delete (currentP)
END WHILE; // seeds <> Empty
RETURN True ;
 END IF
 END; // ExpandCluster
```
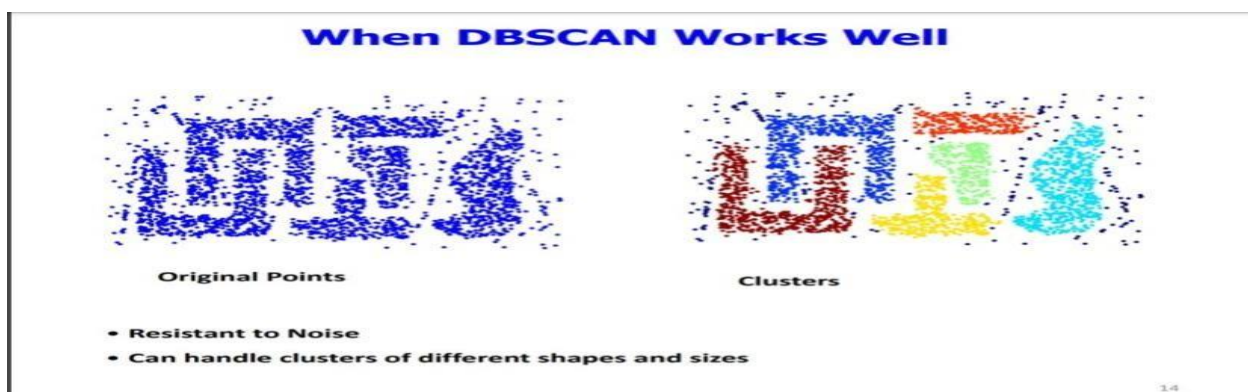
3.3 DATA MINING MODELS

A Density Based Notion of Clusters

The main reason why we recognize the clusters is that within each cluster we have a typical density of points which is considerably higher than outside of the cluster. Furthermore, the density within the areas of noise is lower than the density in any of the clusters

The key idea is that for each point of a cluster the neighborhood of a given radius has to contain at least a minimum number of points, i.e. the density in the neighborhood has to exceed some threshold. The shape of a neighborhood is determined by the choice of a distance function for two points p and q, denoted by dist(p,q). For instance, when using the Manhattan distance in 2D space, the shape of the neighborhood is rectangular. Note, that our approach works with any distance function so that an appropriate function can be chosen for some given application. For the purpose of proper visualization, all examples will be in 2D space using the Euclidean distance.

Definition 2: (directly density-reachable) A point p is directly density-reachable from a point q wrt. Eps, MinPts if 1) p @ NEps(q) 2) INEps(q)l > MinPts (core point condition). Obviously, directly density-reachable is symmetric for pairs of core points. In general, however, it is not symmetric if one core point and one border point are involved.

When DBSCAN Works Well

Original Points   Clusters

• **Resistant to Noise**
• **Can handle clusters of different shapes and sizes**

14

4. Model Evaluation

Basemap does not do any plotting on its own, but provides the facilities to transform coordinates to one of 25 different map projections". One of the important properties of Basemap is — calling a Basemap class instance with the arguments latitude/longitude (in degrees, as in our data-frame), to get x/y map projection coordinates.

These map projection coordinates will be used as features to cluster the data points spatially along with the temperatures

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.
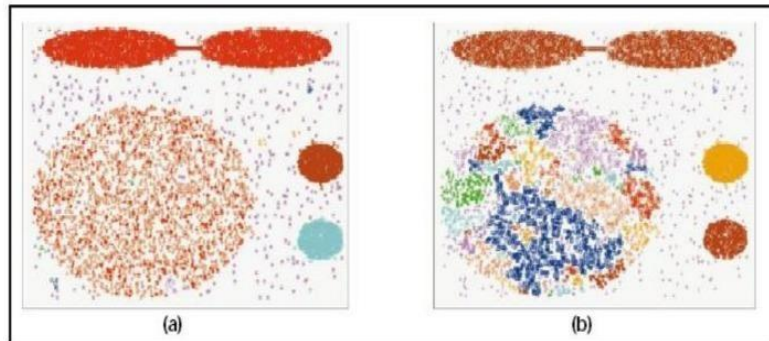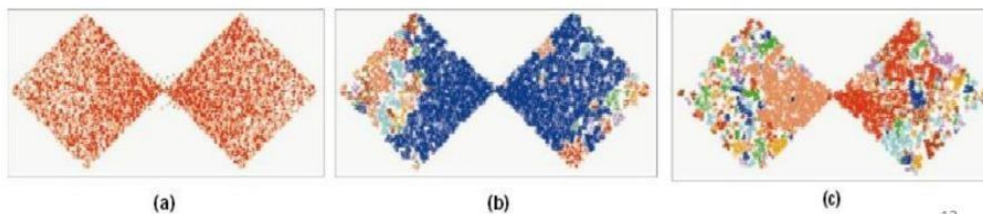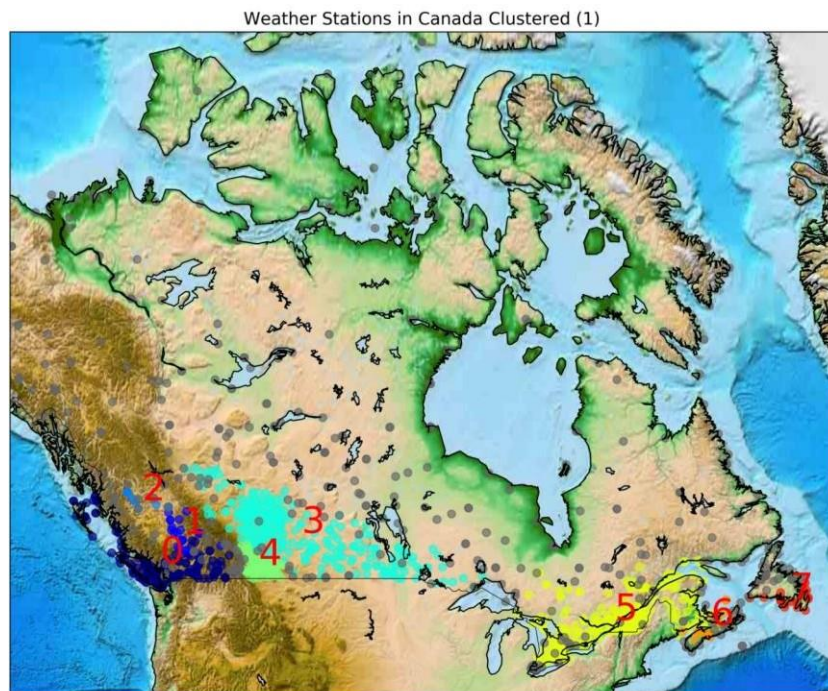
Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.

Precipitaion Cluster

Weather Stations in Canada Clustered (1)

Temp Difference Cluster

5. CONCLUSION

Clustering algorithms are attractive for the task of class identification in spatial databases. However, the well-known algorithms suffer from severe drawbacks when applied to large spatial databases. In this paper, we presented the clustering algorithm DBSCAN which relies on a density-based notion of clusters. It requires only one input parameter and supports the user in determining an appropriate value for it. We performed a performance evaluation on synthetic data and on real data of the SEQUOIA 2000 benchmark. The results of these experiments demonstrate that DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS. Furthermore, the experiments have shown that DBSCAN outperforms CLARANS bya factor of at least 100 in terms of efficiency. Future research will have to consider the following issues. First, we have only considered point objects. Spatial databases, however, may also contain extended objects such as polygons. We must develop a definition of the density in an Eps-neighbourhood in polygon databases for generalizing DBSCAN. Second, applications of DBSCAN to high dimensional feature spaces should be investigated. In particular, the shape of the k-dist graph in such applications has to be explored.

6. REFLECTION & PORTFOLIO

DBSCAN discovers all clusters (according to definition5) and detects the noise points (according to definition from all sample databases. CLARANS, however, splits clusters if they are relatively large or if they are close to some other cluster. Furthermore, CLARANS has no explicit notion of noise. Instead, all points are assigned to their closest medoid .

To test the efficiency of DBSCAN and CLARANS, we use the SEQUOIA 2000 benchmark data. The SEQUOIA 2000 benchmark database (Stone braker et al. 1993) uses real data sets that are representative of Earth Science tasks. There are four types of data in the database: raster data, point data, polygon data and directed graph data. The point data set contains 62,584 Californian names of landmarks, extracted from the US Geological Survey's Geographic Names Information System, together with their location. The point data set occupies about 2.1 M bytes. Since the run time of CLARANS on the whole data set is very high, we have extracted a series of subsets of the SEQUIOA 2000 point data set containing from 2% to 20% representatives of the whole set.

The results of our experiments show that the run time of DBSCAN is slightly higher than linear in the number of points. The run time of CLARANS, however, is close to quadratic in the number of points. The results show that DBSCAN outperforms CLARANS by a factor of between 250 and 1900 which grows with increasing size of the database.

number of points	1252	2503	3910	5213	6256
DBSCAN	3.1	6.7	11.3	16.0	17.8
CLAR-ANS	758	3026	6845	11745	18029
number of points	7820	8937	10426	12512	
DBSCAN	24.5	28.2	32.7	41.7	
CLAR-ANS	29826	39265	60540	80638	

7. REFERENCES

1. https://github.com/suvoooo/Machine_Learning/tree/master/DBSCAN_Complete/images
2. Clustering in Large Spatial Databases, Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, Montreal, Canada, 1995, AAAI Press, 1995.
3. Gueting R.H. 1994. An Introduction to Spatial Database Systems.The VLDB Journal 3(4): 357°399