CS-F415: Data Mining

18 PK-Means MR-DBSCAN

Dr. Kamlesh Tiwari
Associate Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

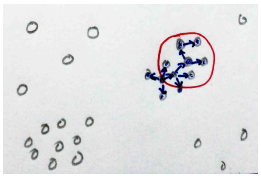
Apr 08, 2024 M/W/F 4:00pm 6101 @ BITS-Pilani [Jan-May 2024]

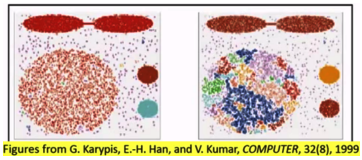
http://ktiwari.in/dm

DBSCAN

DBSCAN [1] (Density-Based Spatial Clustering of Applications with Noise) is a spatial clustering algorithm of KDD96

- Parameters (Eps/MinPts) and points (core/border/noise)
- Uses DFS



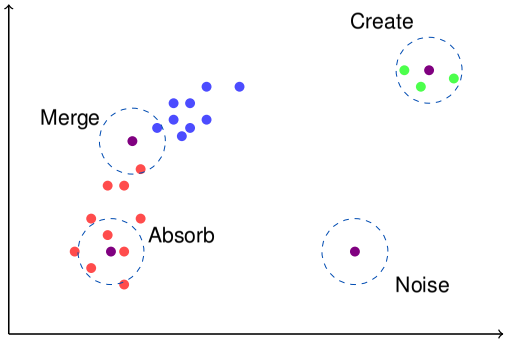


Figures from G. Karypis, E.-H. Han, and V. Kumar, *COMPUTER*, 32(8), 1999

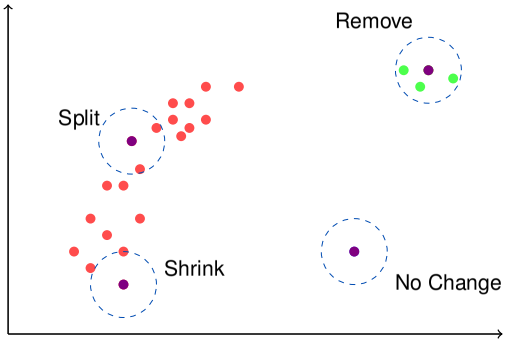
- Advantage: clusters of arbitrary shape
- Disadvantage: Sensitive to parameters

[1] A density-based algorithm for discovering clusters in large spatial databases with noise, *M Ester, HP Kriegel, J Sander, X Xu*, KDD, 96(34), pp 226–231, 1996

Incremental DBSCAN (Addition)



Incremental DBSCAN (Deletion)



Incremental DBSCAN

- Insertion and deletion are treated separately
- Based on change in density in affected region, clusters are updated.
- Update cost is proportional to number of points in affected region that is high
- You may be doing redundant operations

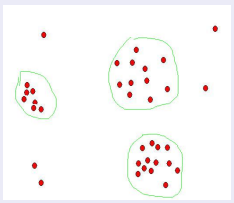
Differ update for some time

Assume periodic arrival of updates.

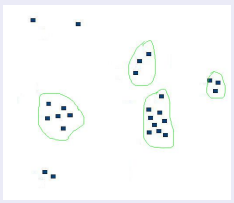
- Cluster new data
- Merge it with previous clusters (it is easy to see the density change)

Incremental DBSCAN

Initial Database

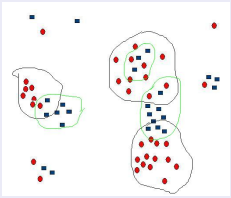


New Date



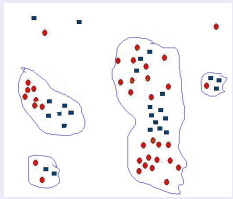
- Region based merging is applied

Incremental DBSCAN



Overlapping of clusters made from original and the new data points looks as

- Point p is in set of intersection I' if $\exists p' \in D$ such that p and p' are neighbor
- It is necessary an sufficient to process all $p \in I'$
- Efficiently compute I'. How?



A Question

Example: Assume you have a database that contains coordinates of points in a 2D plane. Now I give you one more point P and ask you the following question.

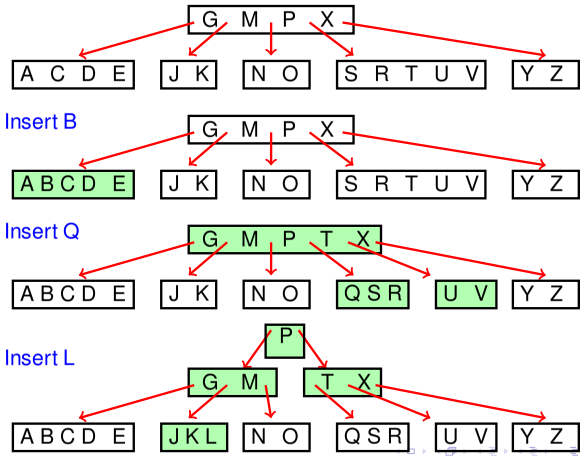
Give me the point from database which is less than 3cm away from P.

- What approach you would follow?
- Evaluate distance from all points in database and sort
- Evaluate distance from all points in database and take minimum
- Something else?

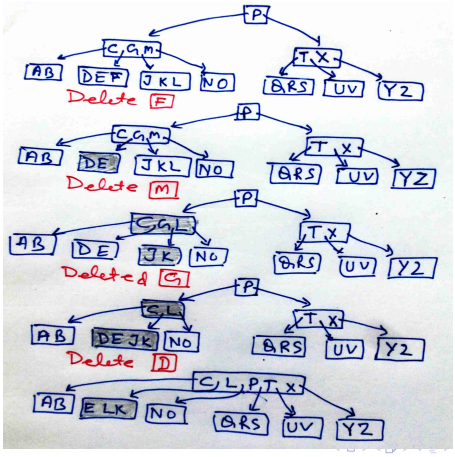
Recall B-tree

- k is a parameter specifying any non-root node can have $k-1$ to $2k-1$ number of keys
- B-Tree is a rooted tree. Every node x has $n(x)$ number of keys with $key_1[x] \leq key_2[x] \leq ... \leq key_{n[x]}[x]$ and $leaf[x] = TRUE$ if it is a leaf node
- Internal node also contains $n(x)+1$ pointers $c_1[x], c_2[x], ..., c_{n[x]+1}[x]$
- For chosen $k > 1$, three exists at least $k-1$ and at most $2k-1$ keys at every node except root
- Height of tree $h \leq \log_t \frac{n+1}{2}$

B-tree (insertion)



B-tree (deletion)



R-tree an efficient date structure

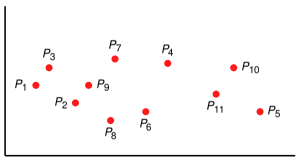
- R-tree[2] provides an approach to index multidimensional spatial data
- Locating a nearest object to a current location is easy by using R-tree
- Or finding all objects in vicinity
- Uses a minimum bounding rectangle (MBR)
- It is a smallest rectangle that always contains the specified object
- Each node in the index contains its children
- Leaves of the tree points the actual objects
- Tree is height-balanced so height is $O(\log n)$
- R-Tree node usually corresponds to database points
- Similar to B-Tree

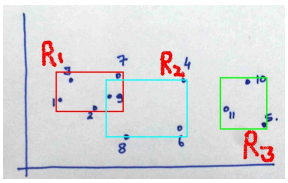
[2] R-trees: A dynamic index structure for spatial searching, Guttman, Antonin, 14(2),:ACM-1984

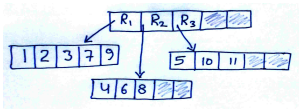
R-tree at Work

Consider arrival of

- P_1, P_2, P_3, P_4, P_5
- P_6 (split and region formation)
- P_7 (R1 expands)
- P_8 (R2 expands)
- P_9, P_{10}
- P_{11} (Split in R2)





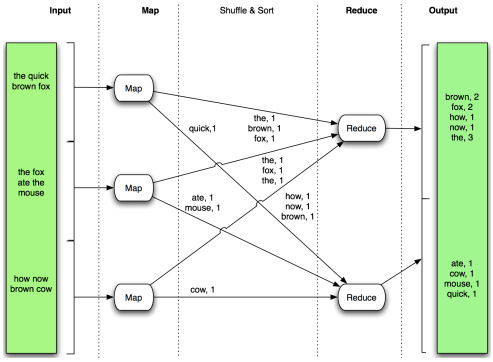


Big Data: MapReduce/Hadoop

- **Grid** (heterogeneous) or **cluster** (homogeneous) computing
- Distributed computing **have to deal with** synchronization, deadlocks, data dependency, mutual exclusion, replication, reliability, platform scalability and provisioning
- Ready-made solution is **MapReduce**/**Hadoop**[3] or **spark** that provides a high level of abstraction for data parallel tasks
- Hadoop/PIG combo is very effective
- Need is there for a scalable distributed computing framework that provides both abstraction and performance (by exploiting all kinds of parallelisms that exist in an algorithm)

[3] MapReduce: Simplified Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemawat, OSDI, 2004

Big Data



map is a function that executes on each partition.

PK-Means

- **K-Means:** steps involved are
  1. Select Seed
  2. Assignment //(most compute intensive)
  3. Compute Centroid
- **PK-Means:** [4]
  - Input dataset is stored on GFS/HDFS. A sequence file of $< key, value >$ pairs. (Key: offset, Value: string of whole record)
  - Dataset is split and globally, and broadcast to all mappers

  Map
  - Distance calculations are parallel executed
  - Each mapper has array of centers
  - Computes closest center for each sample
  - Intermediate values/output: $< key', value' >$ (Key': index of the closed centre, value': sample)

[4] Du, Zhihua and Wang, Yiwei and Ji, Zhen, "PK-means: A new algorithm for gene clustering" in Computational Biology and Chemistry, pages=243–247, vol 32(4), Elsevier 2008

PK-Means

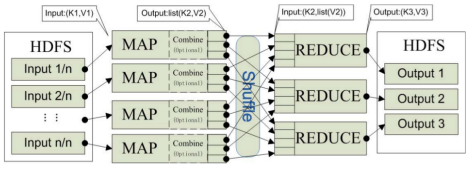
- **PK-Means:**
  Combiner
  - Combines intermediate data of each map task and stores locally
  - Partial sum the values assigned to the same cluster
    - Record number of samples in each cluster and
    - Sum of values at each dimension
  - Key': key' & Value': string of num and sums

  Reducer
  - Input: output of combiner
  - Compute all the samples assigned to a center
  - Calculate new centers

Scalability is high

MR-DBSCAN

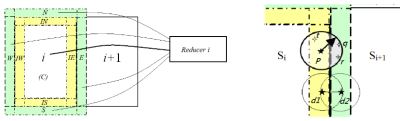
- MR-DBSCAN [5] involves following steps
  1. Preprocessing
  2. Local DBSCAN
  3. Find Merging Mapping
- Uses quadtree, a spacial data structure
- Extended regions (ϵ-extended) is taken in partition



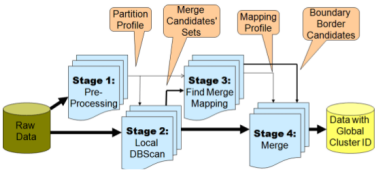
[5] MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm using MapReduce He, Yaobin et.al. Parallel and Distributed Systems (ICPADS) 473–480, IEEE 2011

MR-DBSCAN

- Extended regions (ϵ-extended) is taken in partition



- Cross connection files are processed during reduce
- This makes the algorithm data parallel



Thank You!

Thank you very much for your attention!

Queries ?