

BIG DATA

CLASSIFICATION AND CLUSTERING

GVHD: PGS.TS. Thoại Nam

HVTH:

1. Lê Công – 1670214
2. Lê Tự Đức – 1670217
3. Võ Hoàng An – 1670211



CONTENT

- Classification
 - Basic Concepts
 - Classification Methods
 - Case Study
- Clustering
 - Basic Concepts
 - Clustering Methods
 - Case Study
- Model Evaluation
 - Classification
 - Clustering

CONTENT

- **Classification**
 - Basic Concepts
 - Classification Methods
 - Case Study
- Clustering
 - Basic Concepts
 - Clustering Methods
 - Case Study
- Model Evaluation
 - Classification
 - Clustering

CLASSIFICATION

Basic Concepts

Classification



Dogs



Cats

Given a training set of labeled objects, learn a decision rule

CLASSIFICATION

Basic Concepts

How does classification work?

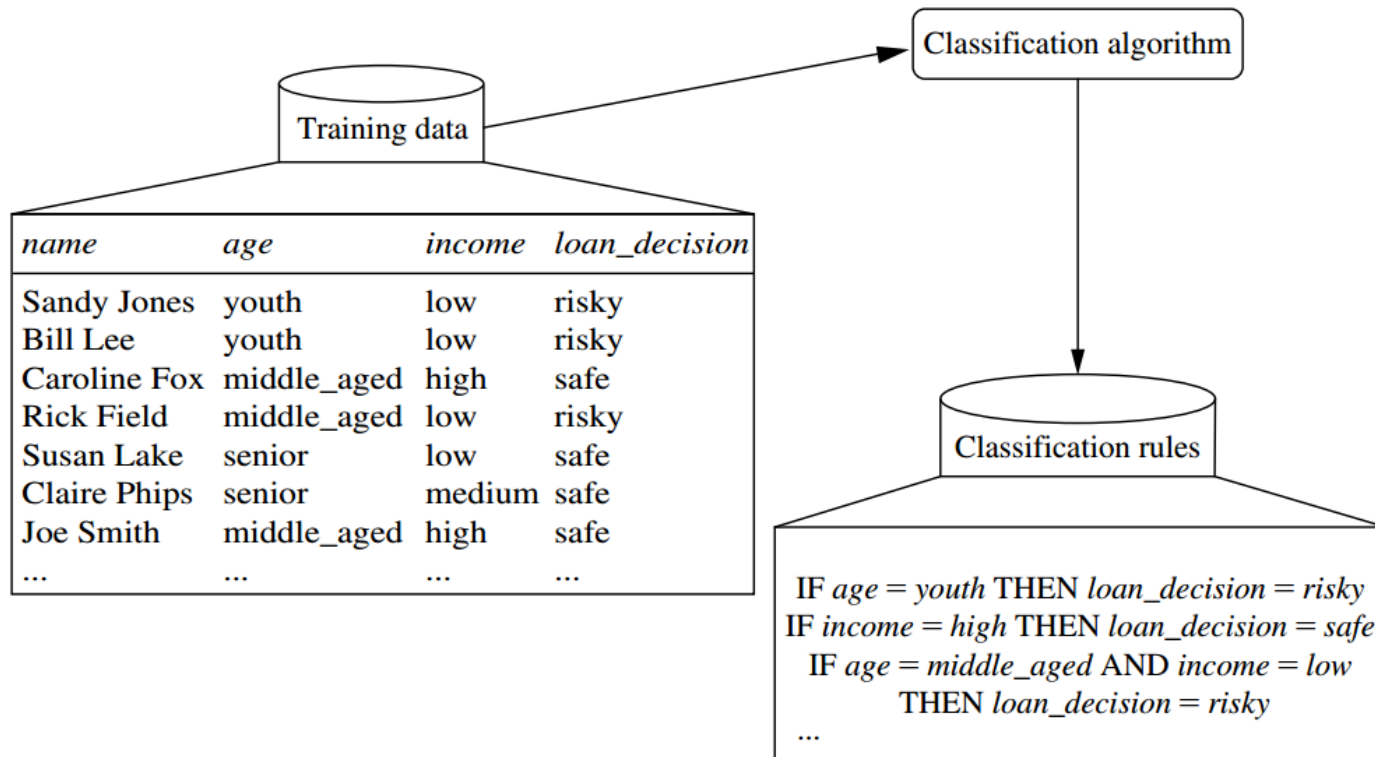
- Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data).

CLASSIFICATION

Basic Concepts

How does classification work?

- In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase)

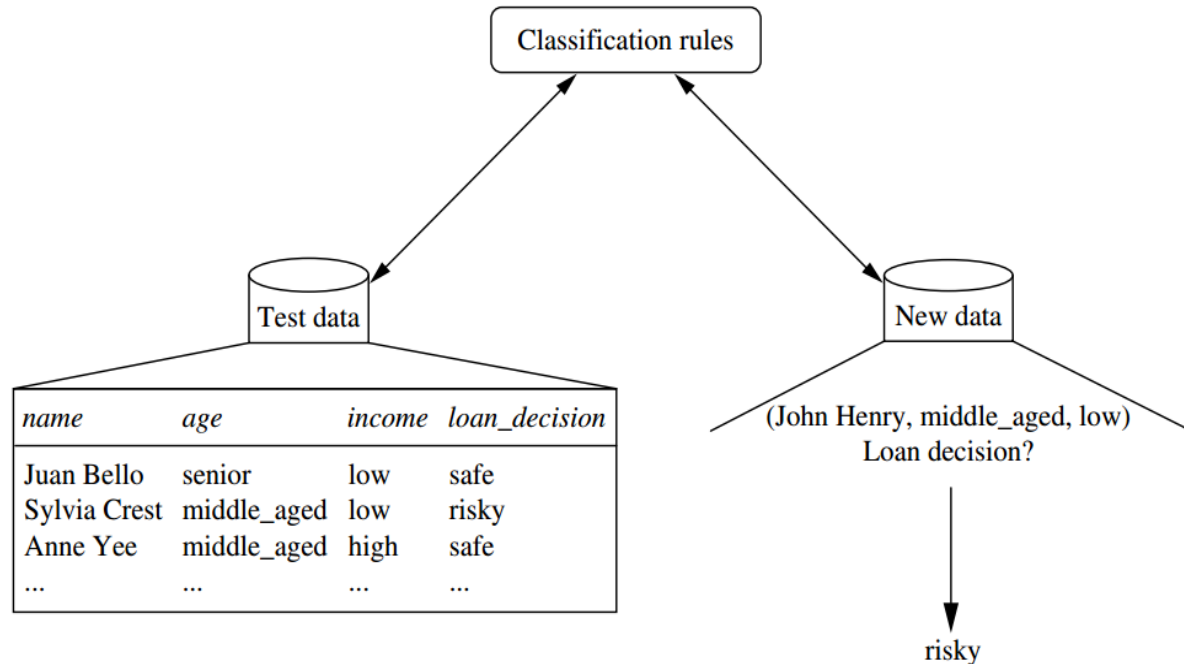


CLASSIFICATION

Basic Concepts

How does classification work?

- In the second step, the predictive accuracy of the classifier is estimated. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known.



CLASSIFICATION

Classificaiton Methods

- Decision Tree method
- Bayes method
- Rule-Based method

CLASSIFICATION

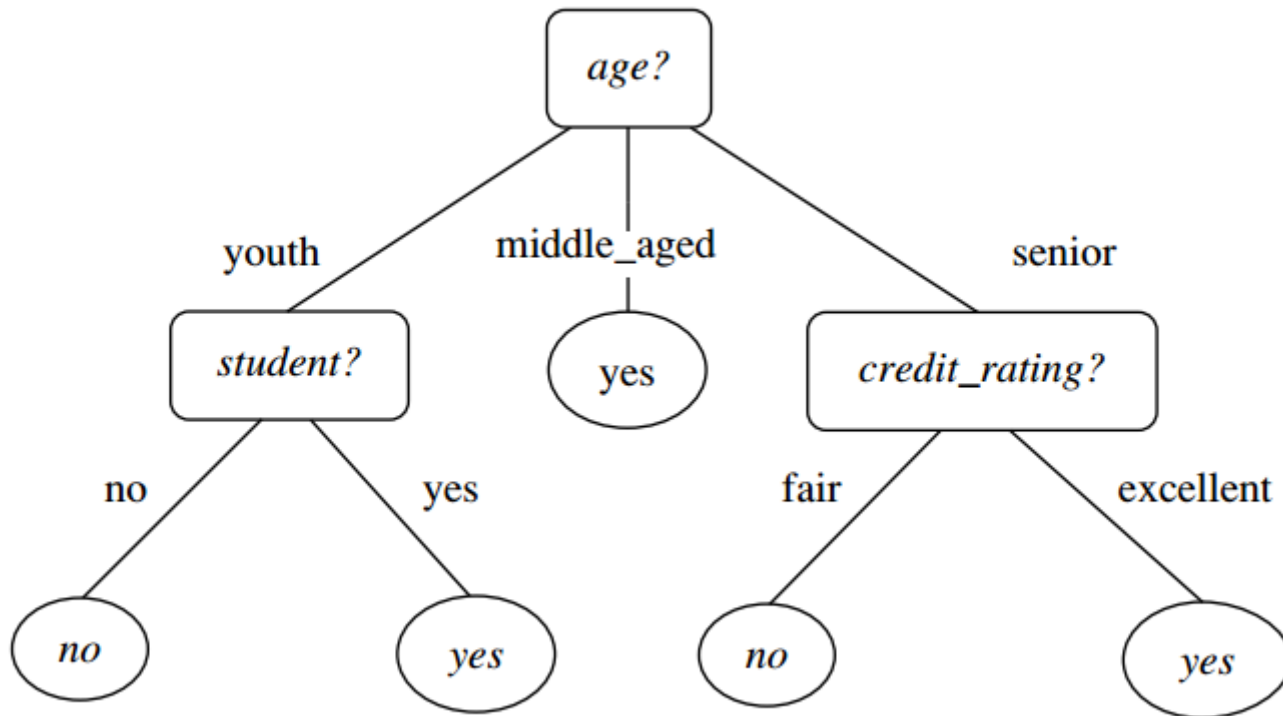
Decision Tree method

- A decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node

CLASSIFICATION

Decision Tree method

- Buy computer or not?



CLASSIFICATION

Bayes method

- Bayesian classifiers are statistical classifiers. They can predict class membership probabilities such as the probability that a given tuple belongs to a particular class
- Bayesian classification is based on Bayes' theorem, described next. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers.
- Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

CLASSIFICATION

Rule-Based method

- A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form: IF condition THEN conclusion
- An example is rule R1
 - R1: IF age = youth AND student = yes THEN buys computer = yes.
- In comparison with a decision tree, the IF-THEN rules may be easier for humans to understand, particularly if the decision tree is very large.

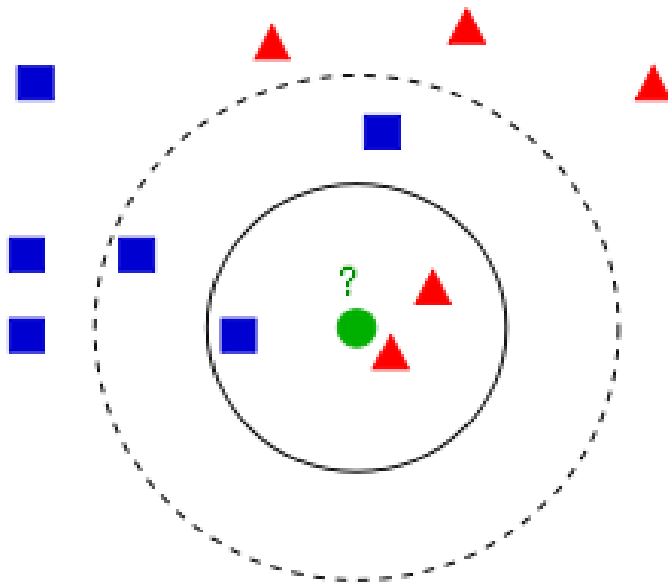
CLASSIFICATION

Case Study

- The k-Nearest Neighbor classifier is one of the most well known methods in data mining because of its effectiveness and simplicity
- In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).
- If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.

CLASSIFICATION

Case Study



- Example of k-NN classification. The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles.
- If $k = 3$ (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle.
- If $k = 5$ (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

CLASSIFICATION

Case Study

Pros

- Simple to implement
- Flexible to feature / distance choices
- Naturally handles multi-class cases
- Can do well in practice with enough representative data

Cons

- Runtime: The complexity of the traditional k-NN algorithm is $O((n \cdot D))$, where n is the number of instances and D the number of features.
- Memory consumption: For a rapid computation of the distances, the k-NN model may normally require to store the training data in memory. When T_R is too big, it could easily exceed the available RAM memory.

CLASSIFICATION

Case Study

How to deal with cons in big data?

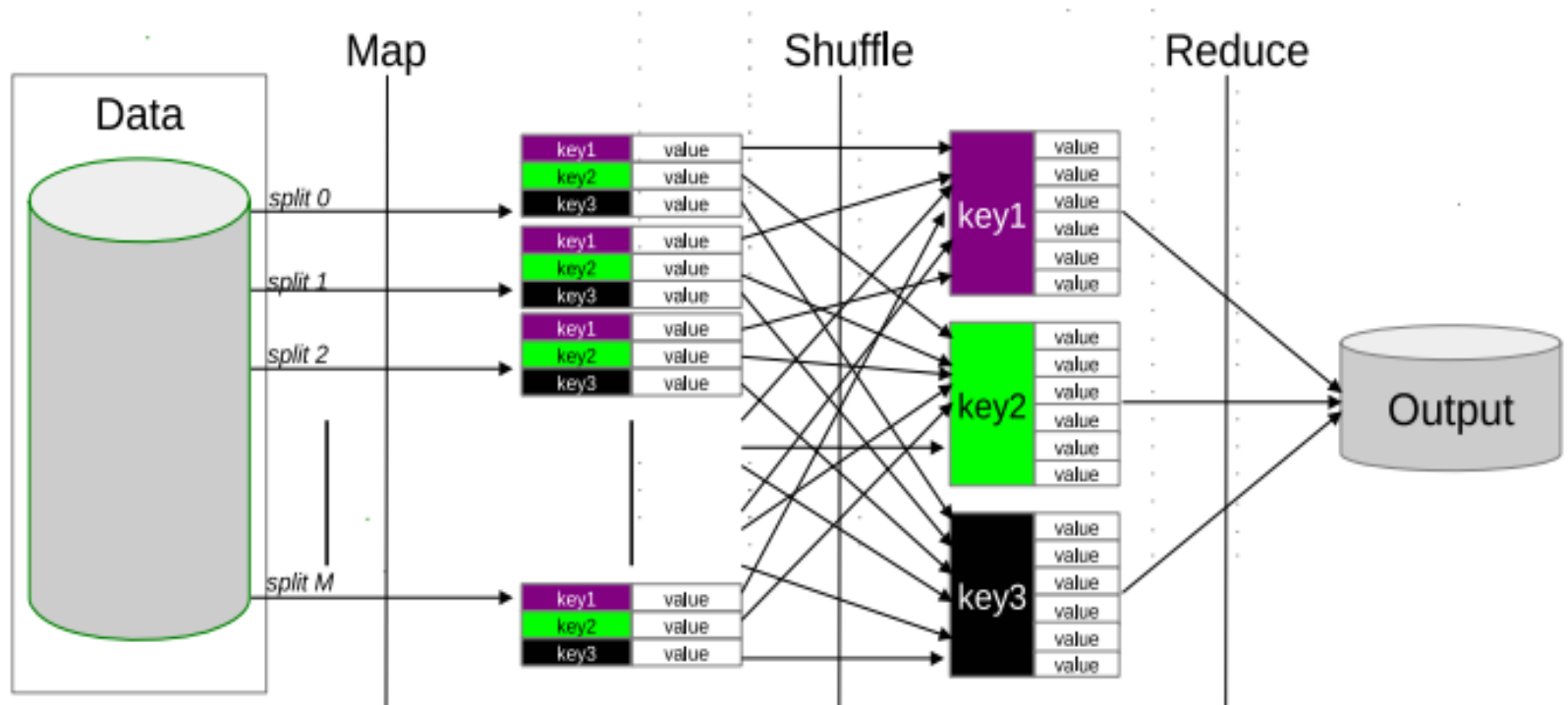
CLASSIFICATION

Case Study

MapReduce-based
k-Nearest Neighbor Approach

CLASSIFICATION Case Study

- Remind about MapReduce



CLASSIFICATION

Case Study

- <Key, value> Format

TABLE I. ENCODING THE RESULTING k NEAREST NEIGHBORS (CLASSES AND DISTANCES) FOR A CERTAIN MAPPER Map_j

| | Neighbor 1 | Neighbor 2 | ... | Neighbor k |
|-----------------------|---|---|-----|---|
| $\mathbf{x}_{test,1}$ | $\langle Class(neigh_1), Dist(neigh_1) \rangle_1$ | $\langle Class(neigh_2), Dist(neigh_2) \rangle_1$ | ... | $\langle Class(neigh_k), Dist(neigh_k) \rangle_1$ |
| $\mathbf{x}_{test,2}$ | $\langle Class(neigh_1), Dist(neigh_1) \rangle_2$ | $\langle Class(neigh_2), Dist(neigh_2) \rangle_2$ | ... | $\langle Class(neigh_k), Dist(neigh_k) \rangle_2$ |
| ... | | | | |
| $\mathbf{x}_{test,t}$ | $\langle Class(neigh_1), Dist(neigh_1) \rangle_t$ | $\langle Class(neigh_2), Dist(neigh_2) \rangle_t$ | ... | $\langle Class(neigh_k), Dist(neigh_k) \rangle_t$ |

CLASSIFICATION

Case Study

- Map phase

Algorithm 1 Map function

Require: $TS; k$

- 1: Constitute TR_j with the instances of split j .
 - 2: **for** $i = 0$ **to** $i < size(TS)$ **do**
 - 3: Compute k-NN ($\mathbf{x}_{test,i}, TR_j, k$)
 - 4: **for** $n = 0$ **to** $n < k$ **do**
 - 5: $CD_j(i, n) = \langle Class(neigh_n), Dist(neigh_n) \rangle_i$
 - 6: **end for**
 - 7: **end for**
 - 8: $key = idMapper$
 - 9: $EMIT(\langle key, CD_j \rangle)$
-

CLASSIFICATION

Case Study

- Reduce phase

Algorithm 2 Reduce operation

Require: $size(TS)$, k , CD_j

Require: Setup procedure has been launched.

```
1: for  $i = 0$  to  $i < size(TS)$  do
2:    $cont = 0$ 
3:   for  $n = 0$  to  $k$  do
4:     if  $CD_j(i, cont).Dist < CD_{reducer}(i, n).Dist$  then
5:        $CD_{reducer}(i, n) = CD_j(i, cont)$ 
6:        $cont++$ 
7:     end if
8:   end for
9: end for
```

CLASSIFICATION

Case Study

- Reduce cleanup process

Algorithm 3 Reduce cleanup process

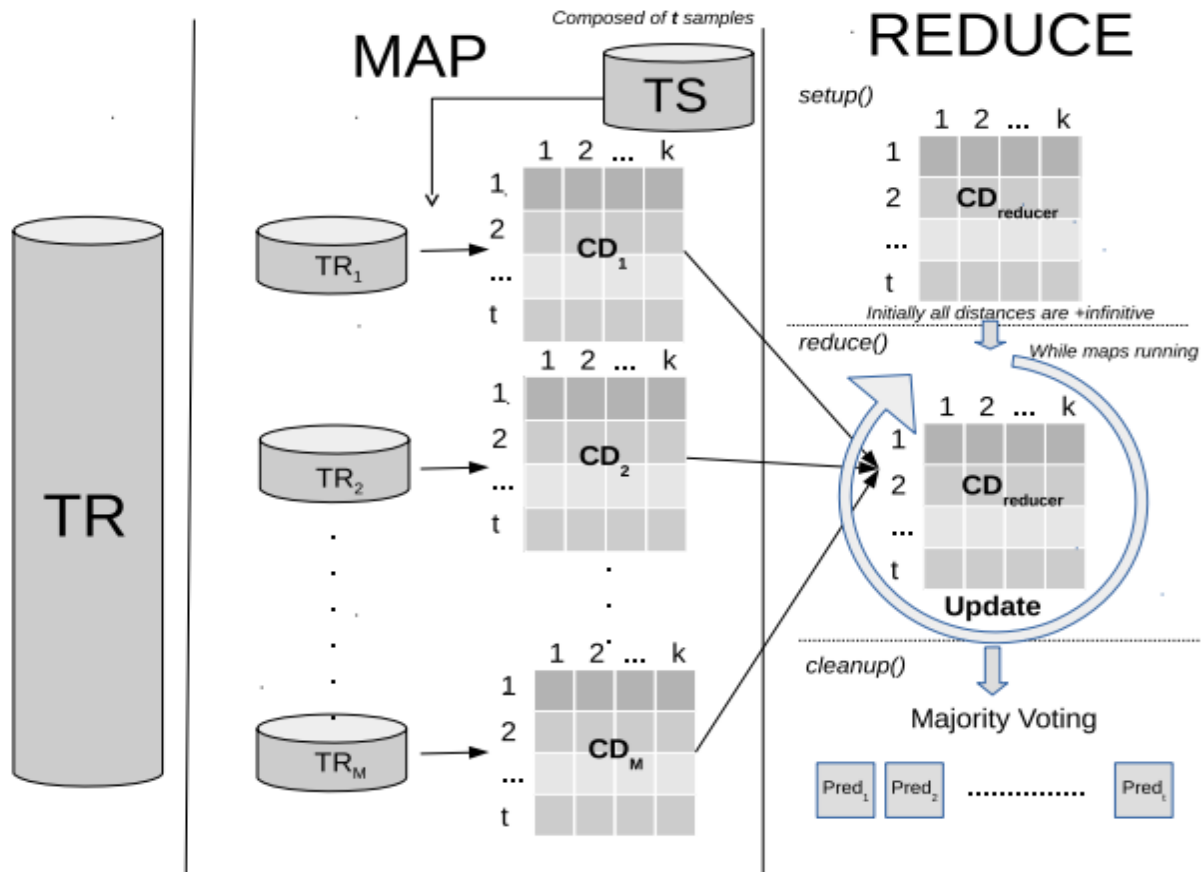
Require: $size(TS)$, k

Require: Reduce operation has finished.

```
1: for  $i = 0$  to  $i < size(TS)$  do  
2:    $PredClass_i = MajorityVoting(Classes(CD_{reduce}))$   
3:    $key = i$   
4:    $EMIT(< key, PredClass_i >)$   
5: end for
```

CLASSIFICATION Case Study

- Flowchart



CLASSIFICATION Case Study

- Summary

TABLE II. SEQUENTIAL K-NN PERFORMANCE

| Number of Neighbors | AccTest | Runtime(s) |
|---------------------|---------|-------------|
| 1 | 0.5019 | 105475.0060 |
| 3 | 0.4959 | 105507.8470 |
| 5 | 0.5280 | 105677.1990 |
| 7 | 0.5386 | 107735.0380 |

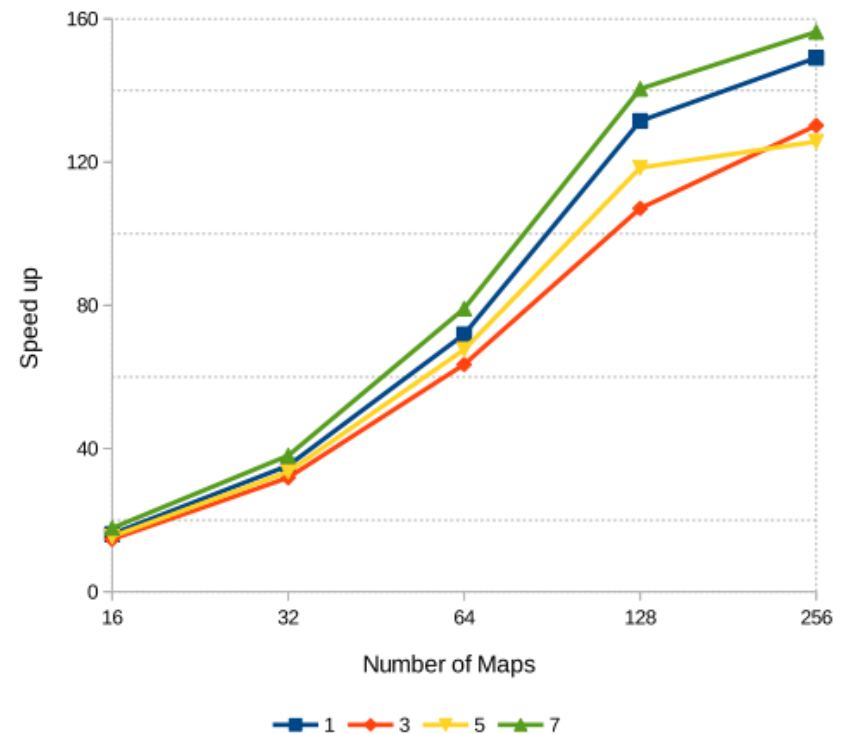


Fig. 3. Speedup

CONTENT

- Classification
 - Basic Concepts
 - Classification Methods
 - Case Study
- Clustering
 - Basic Concepts
 - Clustering Methods
 - Case Study
- Model Evaluation
 - Classification
 - Clustering

CLUSTERING

Basic Concepts

Clustering



Given a collection of (unlabeled) objects, find meaningful groups

CLUSTERING

Basic Concepts

What is a cluster?

"A group of the same or similar elements gathered or occurring closely together"



Galaxy clusters



Birdhouse clusters



Cluster munition



Cluster computing



Cluster lights



Hongkeng Tulou cluster

CLUSTERING

Methods & Case Study

❖ Big Data Challenges

- Heterogeneous data
- Autonomous
- Complexity
- Evolving

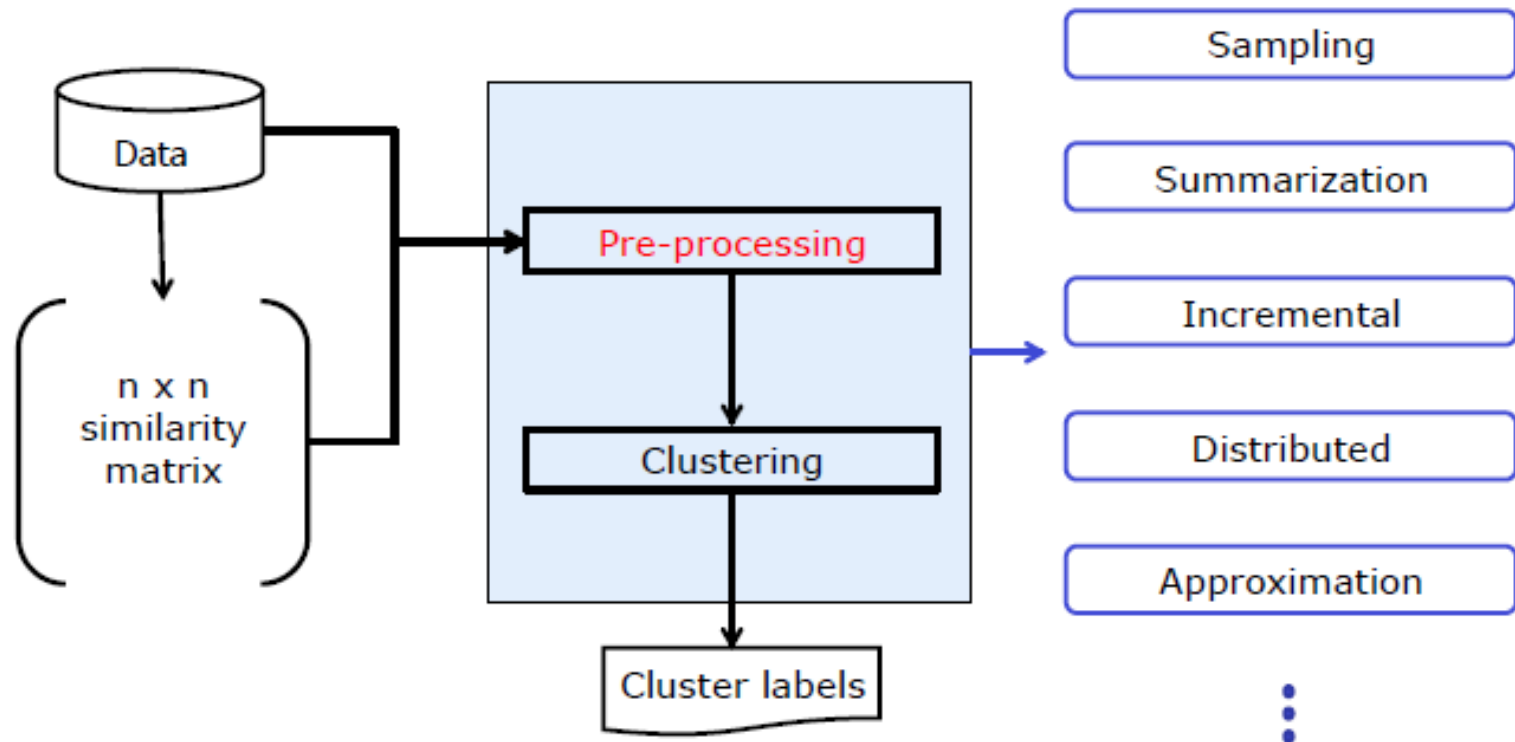
❖ The problem that need to be solved

- Manage large volume of data
- Keep an acceptable resource needs
 - ✓ Time
 - ✓ Memory
 - ✓ Implementation cost

CLUSTERING

Methods & Case Study

Clustering Big Data





















CLUSTERING

Methods & Case Study

Similarity Matrix

Polynomial kernel: $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + 1)^4$

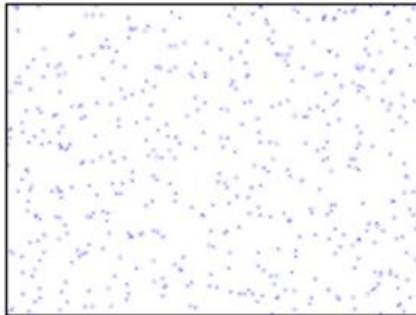
| |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|--|---|---|---|---|
|  | 16 | 15 | 14 | 4 | 6 | 6 | 4 | 3 | 1 |
|  | 15 | 16 | 14 | 4 | 5 | 5 | 6 | 4 | 3 |
|  | 14 | 14 | 16 | 9 | 9 | 9 | 8 | 7 | 4 |
|  | 4 | 4 | 9 | 16 | 15 | 15 | 9 | 10 | 6 |
|  | 6 | 5 | 9 | 15 | 16 | 16 | 7 | 8 | 4 |
|  | 6 | 5 | 9 | 15 | 16 | 16 | 7 | 8 | 4 |
|  | 4 | 6 | 8 | 9 | 7 | 7 | 16 | 16 | 14 |
|  | 3 | 4 | 7 | 10 | 8 | 8 | 16 | 16 | 14 |
|  | 1 | 3 | 4 | 6 | 4 | 4 | 14 | 14 | 16 |

n x n similarity matrix

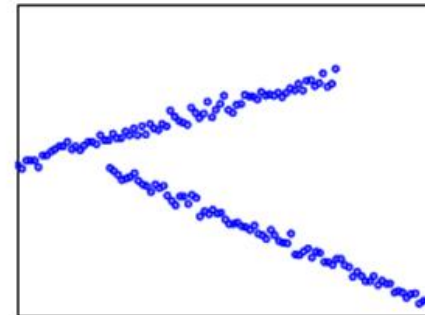
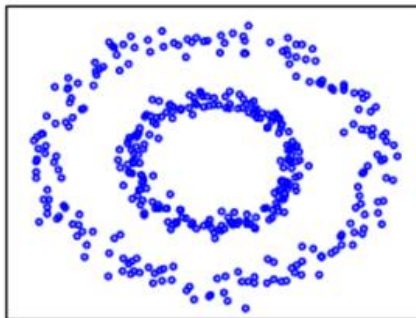
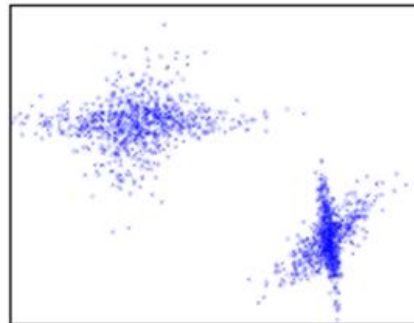
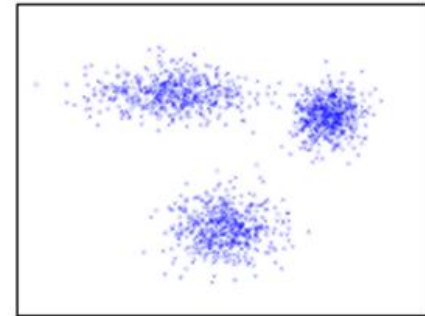
CLUSTERING

Methods & Case Study

Challenges in Data Clustering

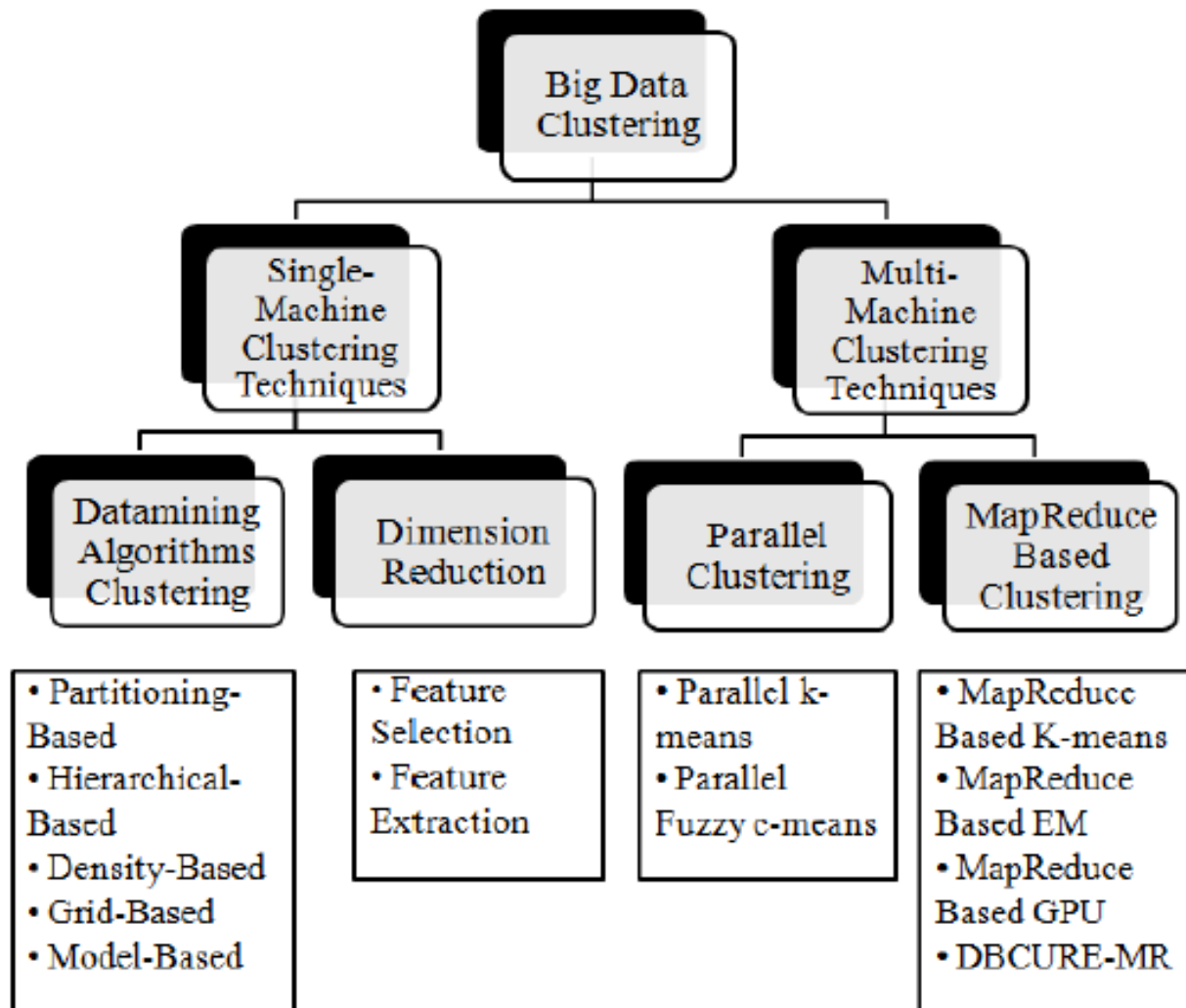


- Measure of similarity
- No. of clusters
- Cluster validity



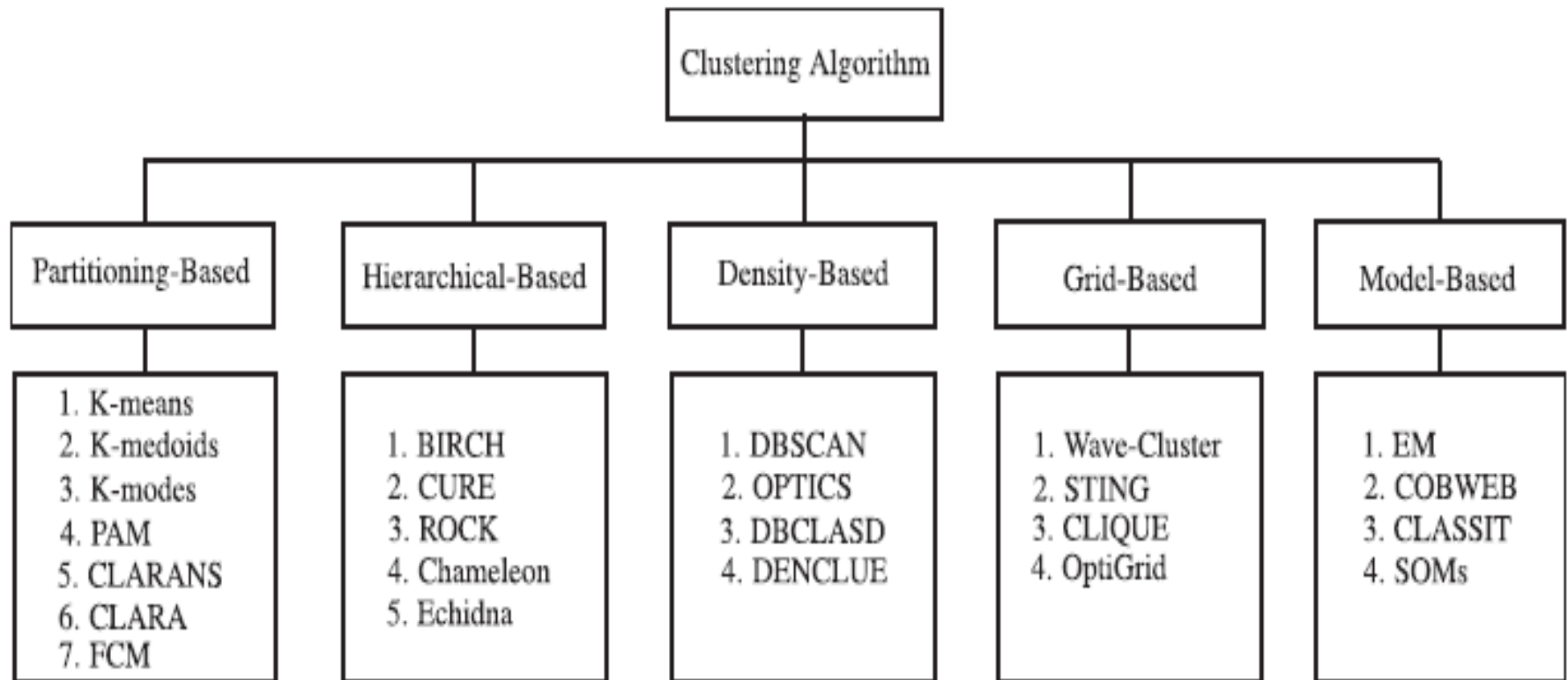
CLUSTERING

Methods & Case Study



CLUSTERING

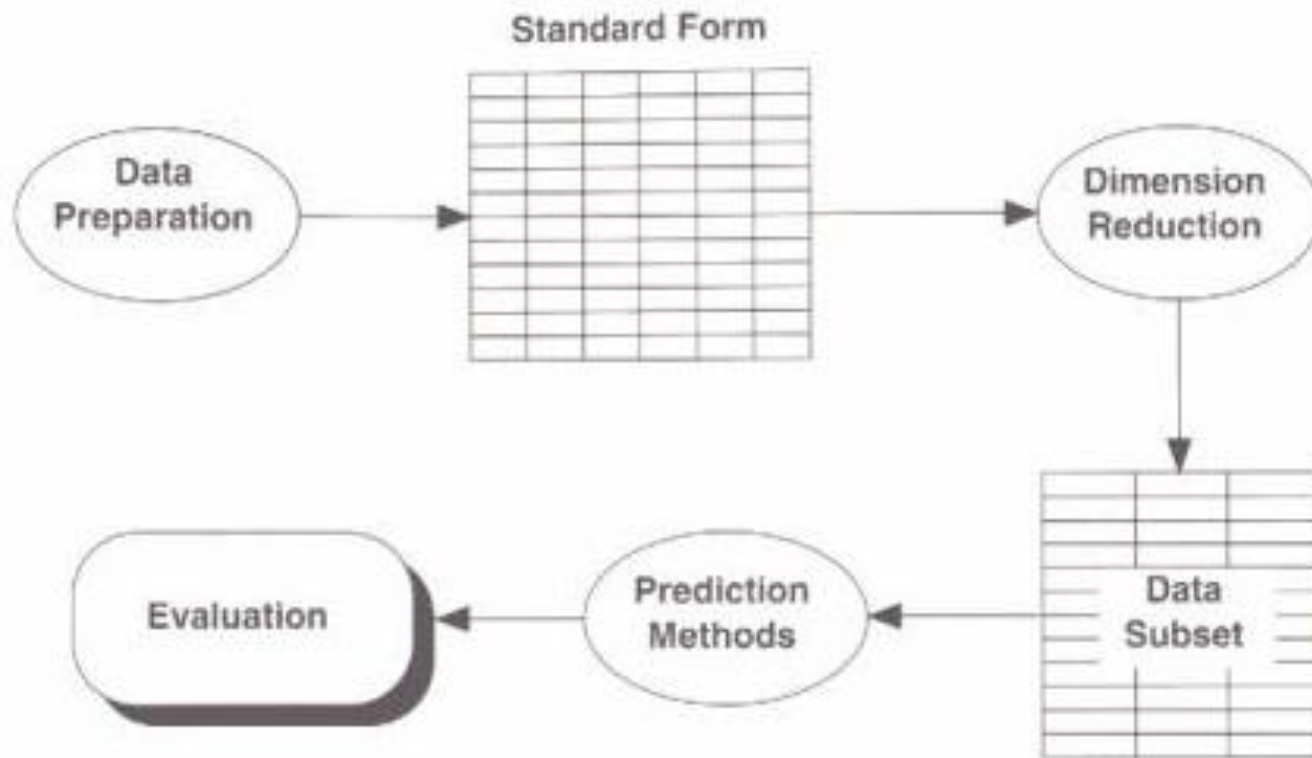
Methods & Case Study



CLUSTERING

Methods & Case Study

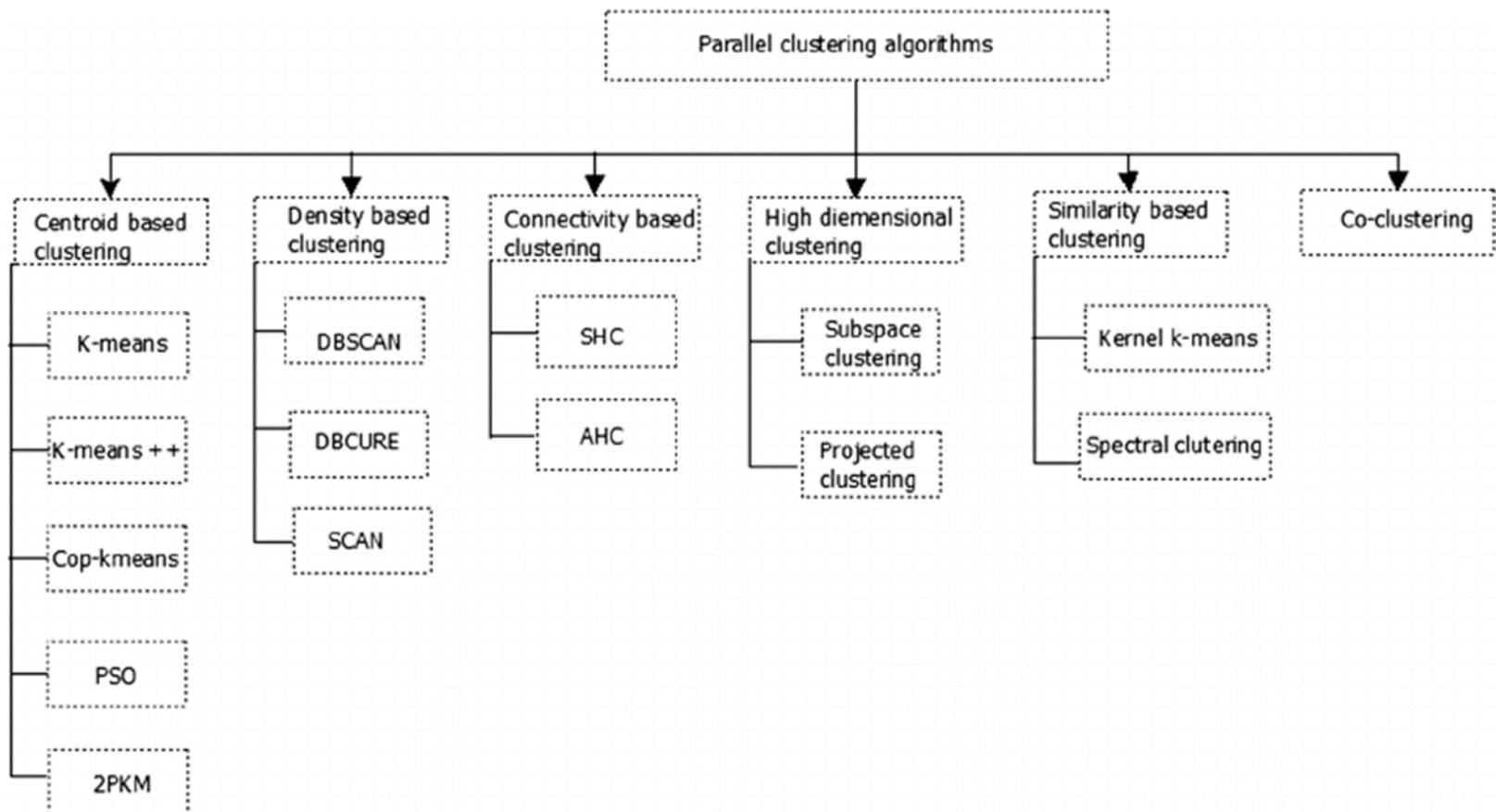
- Dimension Reduction



CLUSTERING

Methods & Case Study

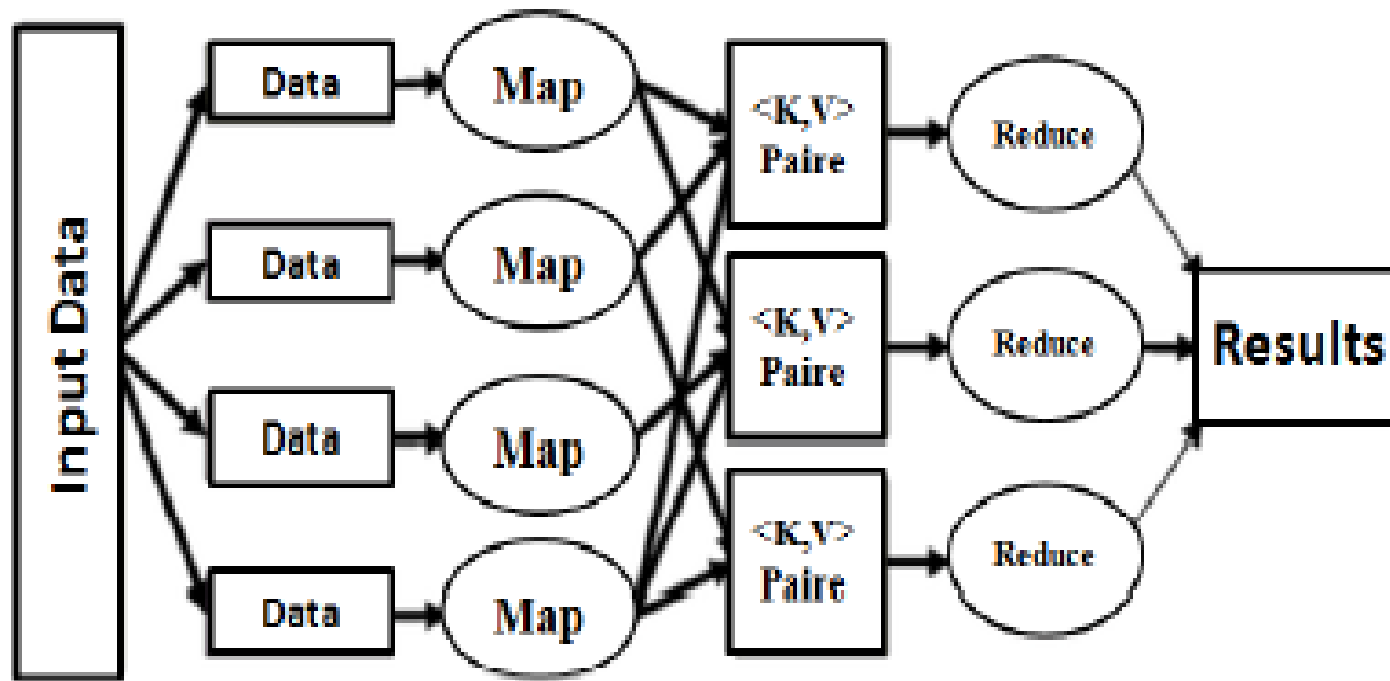
- Parallel Clustering



CLUSTERING

Methods & Case Study

- MapReduce Based Clustering



CLUSTERING

Methods & Case Study

| Clustering Techniques | Advantages | Limitations |
|---|---|---|
| <i>Datamining clustering algorithms</i> | -Simple implementation | -Don't have the capacity to deal with huge amount of data |
| <i>Dimension reduction</i> | -Reduce the dataset -Optimize treatment cost -Very fast and scales algorithm | -Don't offer an efficient solution for high dimensional datasets -Should be performed before applying the classification algorithm |
| <i>Parallel classification</i> | -Minimize the time of execution -More scalable | -Implementing the algorithms can't easily be done |
| <i>MapReduce framework</i> | -Offer impressive scalability -Generate instance responses -Inherently Parallel | -Need more resources -Implementing each query as a MR program is difficult -No primitives for common operations(selection/extraction) |

CLUSTERING

Methods & Case Study

❖ Big Data Challenges

- Heterogeneous data
- Autonomous
- Complexity
- Evolving

❖ The problem that need to be solved

- Manage large volume of data
- Keep an acceptable resource needs
 - ✓ Time
 - ✓ Memory
 - ✓ Implementation cost

CONTENT

- Classification
 - Basic Concepts
 - Classification Methods
 - Case Study
- Clustering
 - Basic Concepts
 - Clustering Methods
 - Case Study
- **Model Evaluation**
 - Classification
 - Clustering

Model Evaluation

Classification

- **positive tuples** and **negative tuples**
- Given two classes, for example, the positive tuples may be `buys_computer = yes` while the negative tuples are `buys_computer = no`

(age, student, credit_rating, buys_computer)

- (senior, no, excellent, **yes**) ← positive tuple
- (senior, no, fair, **no**) ← negative tuple

Model Evaluation

Classification

There are four additional terms we need to know

- True positives (TP): These refer to the positive tuples that were correctly labeled by the classifier. Let TP be the number of true positives.
- True negatives (TN): These are the negative tuples that were correctly labeled by the classifier. Let TN be the number of true negatives

Model Evaluation

Classification

There are four additional terms we need to know

- False positives (FP): These are the **negative tuples** that were incorrectly labeled as positive (e.g., tuples of class buys computer = no for which the classifier predicted buys computer = yes). Let FP be the number of false positives.
- False negatives (FN): These are the **positive tuples** that were mislabeled as negative (e.g., tuples of class buys computer = yes for which the classifier predicted buys computer = no). Let FN be the number of false negatives.

Model Evaluation

Classification

There are four additional terms we need to know

| | | Predicted class | | |
|--------------|------------|-----------------|-----------|--------------|
| | | <i>yes</i> | <i>no</i> | Total |
| Actual class | <i>yes</i> | <i>TP</i> | <i>FN</i> | <i>P</i> |
| | <i>no</i> | <i>FP</i> | <i>TN</i> | <i>N</i> |
| Total | | <i>P'</i> | <i>N'</i> | <i>P + N</i> |

Model Evaluation

Classification

- What is accuracy and error rate in this case?

| <i>Classes</i> | <i>buys_computer = yes</i> | <i>buys_computer = no</i> | <i>Total</i> |
|----------------------------|----------------------------|---------------------------|--------------|
| <i>buys_computer = yes</i> | 6954 | 46 | 7000 |
| <i>buys_computer = no</i> | 412 | 2588 | 3000 |
| Total | 7366 | 2634 | 10,000 |

Model Evaluation

Classification

- The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier.
- $\text{accuracy} = (\text{TP} + \text{TN}) / (\text{P} + \text{N}) = (6954 + 2588) / 10000 = 95.42\%$

| <i>Classes</i> | <i>buys_computer = yes</i> | <i>buys_computer = no</i> | <i>Total</i> |
|----------------------------|----------------------------|---------------------------|--------------|
| <i>buys_computer = yes</i> | 6954 | 46 | 7000 |
| <i>buys_computer = no</i> | 412 | 2588 | 3000 |
| Total | 7366 | 2634 | 10,000 |

Model Evaluation

Classification

- error rate or misclassification rate of a classifier, M , which is simply $1 - \text{accuracy}(M)$, where $\text{accuracy}(M)$ is the accuracy of M .
- error rate = $(FP+FN)/(P+N) = (412+46)/10000 = 4.58\%$

| <i>Classes</i> | <i>buys_computer = yes</i> | <i>buys_computer = no</i> | <i>Total</i> |
|----------------------------|----------------------------|---------------------------|--------------|
| <i>buys_computer = yes</i> | 6954 | 46 | 7000 |
| <i>buys_computer = no</i> | 412 | 2588 | 3000 |
| Total | 7366 | 2634 | 10,000 |

Model Evaluation

Classification

- We now consider the **class imbalance problem**

| <i>Classes</i> | <i>yes</i> | <i>no</i> | <i>Total</i> |
|----------------|------------|-------------|---------------|
| <i>yes</i> | 90 | 210 | 300 |
| <i>no</i> | 140 | 9560 | 9700 |
| Total | 230 | 9770 | 10,000 |

Model Evaluation

Classification

- sensitivity = $TP/P = 90/300 = 30\%$
- specificity = $TN/N = 9560/9700 = 98.56\%$

| <i>Classes</i> | <i>yes</i> | <i>no</i> | <i>Total</i> |
|----------------|------------|-------------|--------------|
| <i>yes</i> | 90 | 210 | 300 |
| <i>no</i> | 140 | 9560 | 9700 |
| <i>Total</i> | 230 | 9770 | 10,000 |

Model Evaluation

Classification

- sensitivity = $TP/P = 6954/7000 = 99.34\%$
- specificity = $TN/N = 2588/3000 = 86.27\%$

| <i>Classes</i> | <i>buys_computer = yes</i> | <i>buys_computer = no</i> | <i>Total</i> |
|----------------------------|----------------------------|---------------------------|--------------|
| <i>buys_computer = yes</i> | 6954 | 46 | 7000 |
| <i>buys_computer = no</i> | 412 | 2588 | 3000 |
| Total | 7366 | 2634 | 10,000 |

Model Evaluation

Clustering

- The silhouette coefficient is such a measure
- For a data set, D , of n objects, suppose D is partitioned into k clusters, C_1, C_2, \dots, C_k
- For each object $o \in D$, we calculate $a(o)$ as the average distance between o and all other objects in the cluster to which o belongs.

$$a(o) = \frac{\sum_{o' \in C_i, o \neq o'} \text{dist}(o, o')}{|C_i| - 1}$$

Model Evaluation

Clustering

- For each object $o \in D$, we calculate $a(o)$ as the average distance between o and all other objects in the cluster to which o belongs.

$$a(o) = \frac{\sum_{o' \in C_i, o \neq o'} \text{dist}(o, o')}{|C_i| - 1}$$

- The value of $a(o)$ reflects the compactness of the cluster to which o belongs.
- The smaller the value, the more compact the cluster

Model Evaluation

Clustering

- $b(o)$ is the minimum average distance from o to all clusters to which o does not belong

$$b(o) = \min_{C_j: 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} \text{dist}(o, o')}{|C_j|} \right\}$$

- The value of $b(o)$ captures the degree to which o is separated from other clusters.
- The larger $b(o)$ is, the more separated o is from other clusters

Model Evaluation

Clustering

- The silhouette coefficient of o is then defined as

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

- The value of the silhouette coefficient is between -1 and 1
- When the silhouette coefficient value of o approaches 1, the cluster containing o is compact and o is far away from other clusters

Model Evaluation

Clustering

- To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster.
- To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

Reference

- Data Mining: Concepts and Techniques, 3rd Edition. Jiawei Han, Micheline Kamber, Jian Pei.
- J. Maillo, I. Triguero, F. Herrera, “A mapreduce-based k-nearest neighbor approach for big data classification”, in 9th International Conference on Big Data Science and Engineering (IEEE BigDataSE-15) (2015), pp. 167–172
- Zerhari, Btissam, Ayoub Ait Lahcen, and Salma Mouline. "Big data clustering: Algorithms and challenges." Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15). 2015.
- Fahad, Adil, et al. "A survey of clustering algorithms for big data: Taxonomy and empirical analysis." IEEE transactions on emerging topics in computing 2.3 (2014): 267-279.
- Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.
- https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

THANK YOU FOR
YOUR ATTENTION !

Q & A