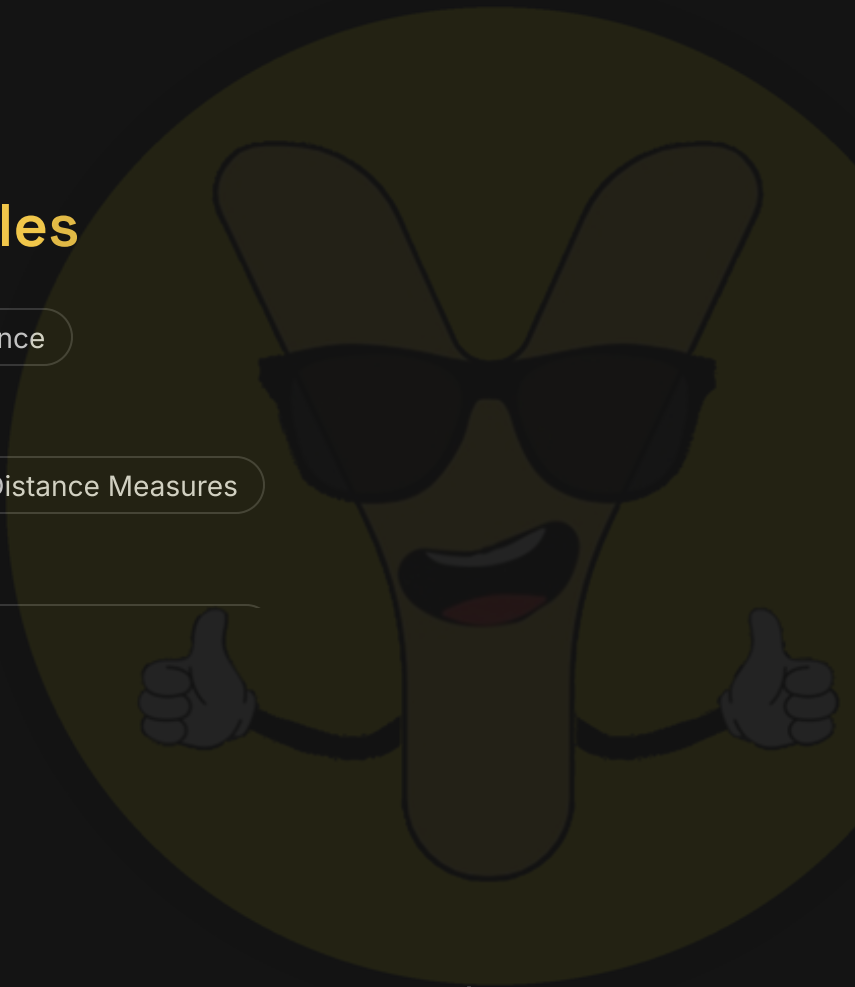
AKTU · 3RD YEAR · SEMESTER 5 · BCS058

Data Warehouse & Data Mining

Unit 4 — Classification, Clustering and Association Rules

- Classification: Definition
- Data Generalization
- Analytical Characterization
- Analysis of Attribute Relevance
- Mining Class Comparisons
- Statistical Measures in Large Databases
- Statistical-Based Algorithms
- Distance-Based Algorithms
- Decision Tree-Based Algorithms
- Clustering: Introduction
- Similarity and Distance Measures
- Hierarchical and Partitional Algorithms
- Hierarchical Clustering: CURE and Chameleon



DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

Classification: Definition and Process

DATA CLASSIFICATION

Data classification is a two-step supervised learning process that builds a predictive model to categorize unseen data instances into predefined discrete class labels based on attribute values.

- Step 1 (Model Construction): Training dataset is analyzed to build classification rules or functions.
- Step 2 (Model Usage): Test dataset evaluates accuracy; if acceptable, model classifies unknown future tuples.
- Supervised learning relies on known target class labels, distinguishing it from unsupervised clustering.

Data Generalization and Analytical Characterization

Generalization replaces low-level primitive data with high-level concepts for characterization.

- Concept Hierarchies map low-level attribute values (e.g., city) to higher-level concepts (e.g., country).
- Data Generalization aggregates target class data to generate summarization profiles.
- Analytical Characterization integrates Attribute Relevance Analysis to filter out non-discrimination features.
- Information Gain and Chi-Square (χ^2) metrics evaluate correlation between attributes and target class.

Mining Class Comparisons and Statistical Measures

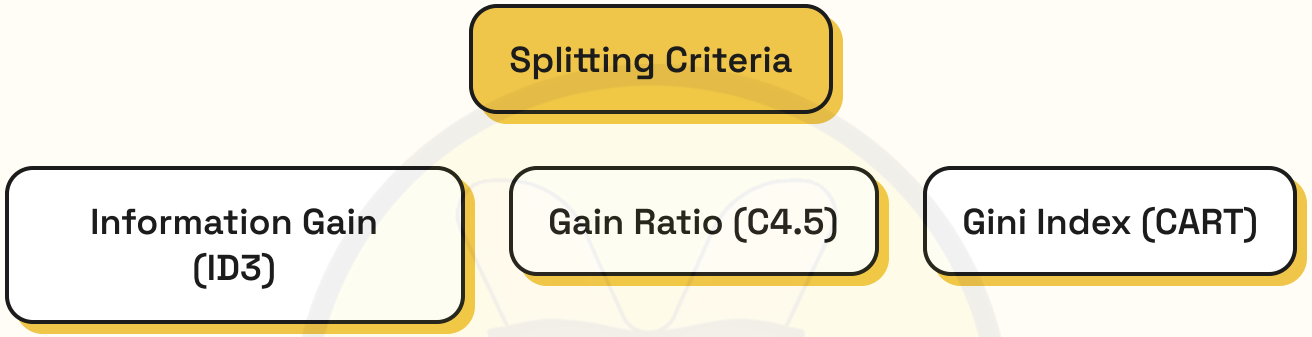
Feature	Class Characterization	Class Discrimination
Objective	Summarizes common properties of target class	Highlights distinguishing features between classes
Input Data	Target class instances only	Target class compared against benchmark classes
Output Result	Generalized profile of target data	Comparative discrimination rule set
Statistical Measures	Mean, count, variance, standard deviation	Relative frequency, discrimination index
Data Mining Context	OLAP data cube summarization	Multi-dimensional class comparisons

Statistical and Distance-Based Classification

Classification algorithms apply probability theory or spatial metric spaces for class assignment.

- Naive Bayes Classifier applies Bayes Theorem assuming class-conditional attribute independence.
- Bayes Theorem: $P(C|X) = \frac{P(X|C)P(C)}{P(X)}$ computes posterior probability of class C given tuple X.
- Maximum A Posteriori (MAP) rule selects class C_k that maximizes $P(C_k|X)$.
- K-Nearest Neighbor (K-NN) is a non-parametric distance-based classifier using spatial proximity.
- Distance metrics like Euclidean distance $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ determine nearest neighbors.

Decision Tree Induction Metrics



Attribute selection measures used to construct decision nodes in tree algorithms.

Decision Tree: Information Gain Example

Calculate Information Gain for dataset S (9 positive, 5 negative) split by attribute A into S1 (6 positive, 2 negative) and S2 (3 positive, 3 negative).

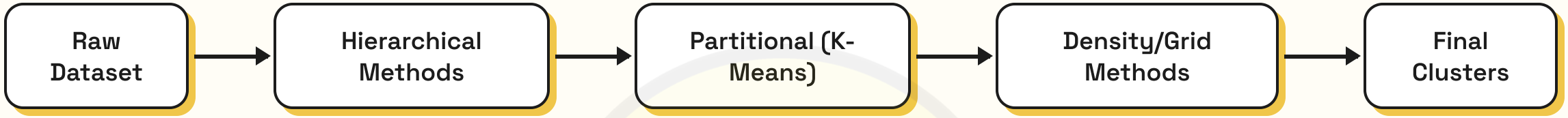
1. Calculate parent Entropy: $Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} \approx 0.940$
2. Calculate Entropy(S1): $Entropy(S_1) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} \approx 0.811$
3. Calculate Entropy(S2): $Entropy(S_2) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.000$
4. Calculate Weighted Entropy: $Entropy_A(S) = \frac{8}{14}(0.811) + \frac{6}{14}(1.000) \approx 0.892$
5. Compute Gain: $Gain(S, A) = Entropy(S) - Entropy_A(S) = 0.940 - 0.892 = 0.048$

Information Gain of Attribute A is 0.048 bits.

Clustering: Distance and Similarity Measures

Distance Metric	Mathematical Formula	Application Context
Euclidean Distance	$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$	Continuous numeric feature spaces
Manhattan Distance	$d(x, y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$	Grid-based or city-block dimensions
Minkowski Distance	$d(x, y) = (\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p)^{1/p}$	Generalized metric space parameter p
Cosine Similarity	$sim(x, y) = \frac{x \cdot y}{\lVert x \rVert \lVert y \rVert}$	High-dimensional sparse text vectors
Jaccard Coefficient	$sim(A, B) = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	Asymmetric binary transactional sets

Taxonomy of Clustering Paradigms



Workflow of major clustering methodologies used in data mining.

Hierarchical Clustering: CURE and Chameleon

Advanced hierarchical algorithms handle complex non-spherical clusters and noisy datasets.

- CURE (Clustering Using Representatives) represents each cluster by multiple well-scattered points.
- Selected points in CURE are shrunk towards cluster centroid by a fraction factor α.
- Shrinking representative points mitigates sensitivity to outliers and handles arbitrary shapes.
- Chameleon uses dynamic modeling on a k-nearest neighbor (k-NN) graph.
- Chameleon merges clusters based on relative interconnectivity and relative closeness.

Density-Based Methods: DBSCAN and OPTICS

Density-based algorithms form clusters as high-density regions separated by low-density regions.

- DBSCAN parameters: ϵ (radius neighborhood) and $MinPts$ (minimum points density).
- Core Point: Contains at least $MinPts$ within its ϵ-neighborhood.
- Border Point: Lies within ϵ-neighborhood of a Core Point but has fewer than $MinPts$.
- Noise Point: Point that is neither a Core Point nor a Border Point.
- OPTICS generates an augmented cluster ordering using Core Distance and Reachability Distance.

Grid-Based Methods: STING vs CLIQUE

Parameter / Feature	STING	CLIQUE
Full Name	Statistical Information Grid	Clustering in Quest
Primary Paradigm	Multi-resolution grid cell hierarchy	Grid-based combined with subspace clustering
Statistical Storage	Stores count, mean, min, max, distribution per cell	Identifies dense grid units across dimensional projections
Time Complexity	$O(K)$ where K is number of grid cells	Linear scalability with total data count $O(n)$
Apriori Utilization	Does not apply Apriori property	Uses Apriori principle to prune non-dense subspaces

Model-Based Clustering: Statistical Approach

Model-based methods optimize the fit between given data and underlying mathematical models.

- Statistical approaches assume data is generated by a mixture of probability distributions.
- Gaussian Mixture Models (GMM) represent clusters as multivariate normal distributions $N(\mu_k, \Sigma_k)$.
- Expectation-Maximization (EM) algorithm iteratively estimates distribution parameters.
- E-step (Expectation): Computes soft membership probability of data points belonging to each component.
- M-step (Maximization): Re-estimates mean μ_k and covariance Σ_k to maximize likelihood.

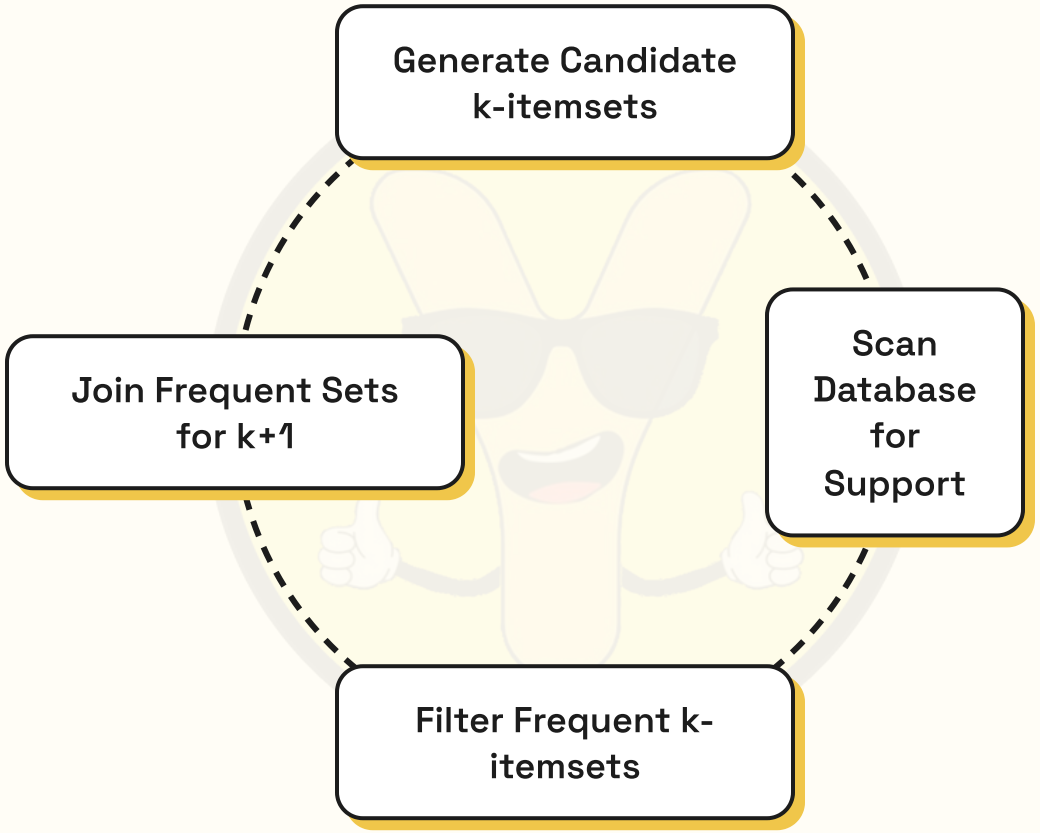
Association Rule Mining Principles

ASSOCIATION RULES

Association Rule Mining identifies strong relationships, frequent patterns, and casual structures among itemsets in transactional databases, expressed as implication rules $X \Rightarrow Y$.

- Support: Probability that a transaction contains both itemsets $X \cup Y$, i.e., $P(X \cup Y)$.
- Confidence: Conditional probability that a transaction containing X also contains Y, i.e., $P(Y|X)$.
- Large (Frequent) Itemsets: Itemsets whose support is greater than or equal to minimum support threshold (min_sup).

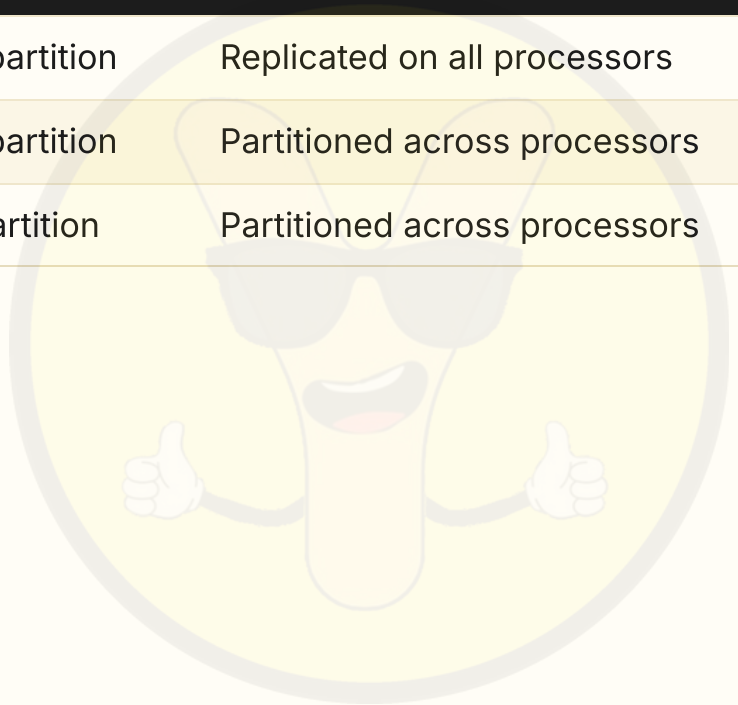
Apriori Algorithm Execution Cycle



Iterative candidate generation and database pruning cycle in Apriori.

Parallel and Distributed Association Mining

Algorithm Strategy	Data Partitioning	Candidate Storage	Communication Overhead
Count Distribution	Horizontal data partition	Replicated on all processors	Low (synchronizes counts per pass)
Data Distribution	Horizontal data partition	Partitioned across processors	High (broadcasts transaction tuples)
Candidate Distribution	Selective data partition	Partitioned across processors	Moderate (point-to-point exchanges)



Neural Network Approach to Classification and Association

Artificial Neural Networks construct complex non-linear decision boundaries for predictive mining.

- Multi-Layer Perceptron (MLP) consists of Input, Hidden, and Output layers connected feed-forward.
- Backpropagation algorithm uses Gradient Descent to adjust synaptic weights and minimize squared error.
- Activation functions like Sigmoid $\sigma(z) = \frac{1}{1+e^{-z}}$ introduce non-linear mapping capabilities.
- Associative Neural Networks (e.g., Hopfield Networks) store patterns for auto-associative itemset retrieval.
- Neural networks offer high fault tolerance and robustness against noisy data tuples.

Exam Corner: Unit 4 Review Questions

- **Q1** Explain Naive Bayes classification with mathematical formulation and discuss the class conditional independence assumption.
- **Q2** Construct a Decision Tree using Information Gain and Entropy for a given sample dataset.
- **Q3** Differentiate between Hierarchical and Partitional clustering algorithms with suitable examples.
- **Q4** Explain DBSCAN density-based clustering algorithm. Define Core, Border, and Noise points with diagrams.
- **Q5** Illustrate the Apriori algorithm step-by-step for finding frequent itemsets with minimum support threshold.
- **Q6** Compare Count Distribution, Data Distribution, and Candidate Distribution parallel algorithms for association mining.
- **Q7** Explain the Backpropagation learning algorithm in Artificial Neural Networks for classification problems.

Unit 4 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 5 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

