Clustering – An Overview

Clustering is the process of grouping a collection of objects into classes of similar objects. It is a crucial tool in data analysis, involving methodologies for automatic classification based on similarity.

Key Points

- **Clustering vs. Supervised Classification**: Clustering is unsupervised classification, unlike supervised classification which involves pre-labeled data.
- **Typical Pattern Clustering Steps**:
  - Pattern representation (including feature extraction and/or selection)
  - Definition of a pattern proximity measure appropriate to the data domain

- Clustering
- Data abstraction
- Assessment of output

Cluster Analysis

- **Exploratory Discovery Process**: Used to discover structures in data without providing explanations.
- **Major Aspects**:
  - **Clustering**: Partitioning objects into groups based on criteria.
  - **Cluster Validation**: Evaluating the quality of clustering results to find the best clustering scheme for a specific application.

Process and Utility

- Clustering divides data points into clusters, making data more manageable and revealing the internal structure of statistical information.
- It improves data readiness for artificial intelligence techniques and can be used as a pre-processing step for other algorithms.

Simple Explanation

- A cluster is a group of related objects where the distance between members is less than the distance between members of different clusters.
- Clustering is represented as a multidimensional space segment with a high density of related objects.

Applications of Cluster Analysis in Data Mining

- **Various Fields**: Data analysis, market research, pattern identification, image processing.
- **Internet**: Assigning documents, credit card fraud detection, insight into data distribution.
- **Biology**: Plant and animal taxonomy, gene classification, population structure analysis.
- **Earth Observation**: Identifying similar land regions, grouping houses by type, value, and location.
- **Search Engines**: Presenting similar objects together, ignoring dissimilar objects, and fetching related objects.
- **Academics**: Associative analysis of documents for plagiarism, copyright infringement, patent analysis.
- **Bioinformatics**: Detecting cancerous cells in medical imagery.
- **OTT Platforms**: Implementing movie recommendations.
- **News Summarization**: Grouping articles by related topics.
- **Sports Training**: Recommending training regimens for athletes based on goals and body metrics.
- **Marketing and Sales**: Identifying demand-supply gaps from past metrics.
- **Job Search Portals**: Organizing job postings for easier job-seeker targeting.
- **Resume Segmentation**: Grouping resumes by skill sets, experience, strengths, expertise.
- **Traffic Analysis**: Detecting patterns and suggesting best routes from GPS data.
- **Satellite Imagery**: Segmenting for agricultural suitability.
- **Customer Persona Analysis**: Building user profiles from recency, frequency, and monetary metrics for customer loyalty.
- **Document Clustering**: Preventing the spread of fake news on social media.
- **Website Traffic Analysis**: Segmenting network traffic to prioritize requests and detect malicious activities.
- **Customer Segmentation in Eateries**: Targeting campaigns effectively to increase engagement.

Clustering Methods

Major Goals for Successful Grouping

1. **Similarity**: Between data points.
2. **Distinction**: Between similar data points and different data points.

Challenges in Clustering

- **Scalability**: Handling large datasets.
- **Data Attributes**: Dealing with categorical and continuous data.
- **Multidimensional Data**: Managing data with multiple dimensions.
- **Cluster Shape**: Ensuring clusters are inclusive and not limited to geometric shapes.
- **Noise**: Handling unwanted features in data.
- **Interpretation**: Making clustering outputs understandable and fitting business criteria.

Types of Clustering Methods

1. **Partitioning Method**
2. **Hierarchical Method**
3. **Density-based Method**
4. **Grid-Based Method**
5. **Model-Based Method**
6. **Constraint-based Method**

Partitioning Method

- Breaks data into k clusters where k < n.
- Uses iterative relocation.
- Examples: k-means, k-medoids.

Hierarchical Method

- Decomposes data into hierarchical clusters.
- Represented by a Dendrogram.
- Two approaches:
  - **Agglomerative (Bottom-up)**
  - **Divisive (Top-down)**

Density-Based Method

- Clusters based on local density.
- **Features**:
  - Discovers arbitrary shape clusters.
  - Handles noise data.
  - Examines local regions for density.

    - Requires density parameters.
- **Types**:
    - **Density Based Connectivity**: DBSCAN, DBCLASD.
    - **Density Based Function**: DENCLUE.

Grid-Based Method

- Uses multilevel grid structures.
- Efficient with complexity O(N).
- Examples: STING, CLIQUE.
- **Issue**: Deciding grid size, depends on user experience.

Model-Based Clustering Method

- Assumes data is generated by a mixture of probability distributions.
- Optimizes fit between data and model.
- Examples: Statistical approach, neural network approach.
- **Challenges**:
    - Choosing a suitable model for unknown data distributions.
    - High computational cost for large datasets.

Constraint-Based Method

- Partitions data based on certain constraints.
- Uses supervised machine learning techniques.
- **Constraints**: Desired properties of clustering results (e.g., number of clusters, cluster size, important dimensions).
- Examples: Decision Trees, Random Forest, Gradient Boosting.
- **Process**:
    - Tree is constructed by splitting without constraints.
    - Leaf nodes are combined into clusters with constraints using suitable algorithms.

Partitioning Method

- Popular choice for analysts to create clusters
- Also known as Supervised Clustering method
- Requires specifying the number of clusters
- Iterative process to reassign data points based on distance

k-Means Algorithm

- **Type:** Unsupervised and iterative algorithm
- **Objective:** Minimize distance between cluster and data set
- **Process:**
    i. Define the number of clusters (k) and centroids
    ii. Calculate distance from every data point to all centroids
    iii. Assign point to cluster with minimum distance

iv. Calculate new centroid for the cluster
v. Repeat until desired clusters are formed

- **Complexity:** O(tkn) where n = total data set, k = clusters, t = iterations
- **Advantages:**
  - Effortless implementation
  - Dense, spherical clusters
  - Suitable for large databases
- **Disadvantages:**
  - Inappropriate for clusters with different density and size
  - Non-equivalent results on iterative runs
  - Euclidean distance may weigh unequally
  - Unsuccessful for non-linear and categorical data
  - Difficult to handle noisy data and outliers

k-Medoids or PAM (Partitioning Around Medoids)

- **Similarity to k-means:** Process is similar but medoid must be an input data point
- **Process:**
  i. Choose m random points as initial medoids
  ii. Assign each data point to the closest medoid
  iii. Calculate swapping cost for chosen and unchosen objects
  iv. Replace if cost < 0
  v. Repeat until no change in medoids
- **Characteristics:**
  - Shift-out membership
  - Shift-in membership
  - Update current medoids
  - No change
- **Advantages:**
  - Easy to understand and implement
  - Quick and converges in few steps
  - Allows dissimilarities between objects
  - Less sensitive to outliers compared to k-means
- **Disadvantages:**
  - Initial sets of medoids can produce different results
  - Clusters may depend on units of measurement

Hierarchical Method

- Decomposes data items into a hierarchy
- Two approaches:
  - Agglomerative (Bottom-up)
  - Divisive (Top-down)

Agglomerative Approach

- **Process:**
    1. Initialize all n data points into N individual clusters
    2. Find and combine closest cluster pairs
    3. Calculate pair-wise distance between clusters
    4. Repeat until all samples are merged into a single cluster
- **Advantages:**
    - Easy to identify nested clusters
    - Better results and ease in implementation
    - Suitable for automation
    - Reduces computing time and space complexity
- **Disadvantages:**
    - Cannot undo previous steps
    - Difficulty handling different sized clusters and convex shapes
    - No direct minimization of objective function
    - Difficulty identifying the exact number of clusters

Divisive Approach

- **Process:**
    1. Start with one cluster containing all samples
    2. Select largest cluster with widest diameter
    3. Find point with minimum average similarity
    4. Add to fragment group
    5. Find element with highest average similarity to fragment group
    6. Assign data sample if average similarity is greater
    7. Repeat until each data point is separated into individual clusters
- **Advantages:**
    - More accurate hierarchies than bottom-up in some cases
- **Disadvantages:**
    - Computationally complex
    - Different distance metrics may generate different results

Density Based Method

- Clusters formed based on neighborhood density reaching a threshold
- Assumes spherical or regular shapes

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **Parameters:**
    - Eps: Maximum radius from its neighborhood
    - MinPts: Minimum points in Eps-Neighborhood
- **Definitions:**
    - **Core Point:** Lies within Eps and MinPts, surrounded by dense neighborhood
    - **Border Point:** Lies within neighborhood of core point, not densely surrounded
    - **Noise/Outlier:** Does not belong to cluster

  - **Direct Density Reachable:** Point p from q within Eps, MinPts
  - **Density Reachable:** Chain of points from q to p
- **Algorithm:**
  i. Consider a random point p
  ii. Find all points density reachable from p
    - If core point, form cluster
    - If border point, visit next point
  iii. Continue until all points are processed
- **Advantages:**
  - Identifies outliers
  - No need to specify number of clusters in advance
- **Disadvantages:**
  - Efficiency drops with changing data density
  - Not suitable for high-quality data
  - Parameters must be specified in advance

Limitations with Cluster Analysis

- Difficulty dealing with arbitrarily shaped data distributions
- High computational cost for validating clustering results
- Inefficiency of clustering algorithms on large datasets
- Exclusion of user domain knowledge in the clustering process

Outlier Analysis

- Outliers are data points that deviate significantly from the norm
- Important in identifying experimentation flaws, fraud, and new trends

Outliers in Data Mining

- Often ignored by algorithms but critical in applications like fraud detection
- **Causes:**
  - Financial fraud detection
  - Monitoring customer purchase habits
  - Typing errors
  - Troubleshooting machines and systems

Handling Outliers in Data Mining

- **Reasons:**
  - Impact on database outcomes
  - Potential for useful discoveries and patterns
  - Valuable in research
  - Essential subfield in data mining

Outlier Detection

- Defined as models far from mainstream data
- **Techniques:**
  - Numeric Outlier: Uses IQR for one-dimensional feature space
  - Z-Score: Considers Gaussian distribution of data
  - DBSCAN: Based on DBSCAN clustering method
  - Isolated Forest: Suitable for large datasets, uses isolation number

Models for Outlier Detection Analysis

- **Intensive Value Analysis:** Basic form, suitable for 1-dimensional data
- **Linear Models:** Structures data outside lower dimensional substructure
- **Probabilistic and Statistical Models:** Uses specific data distributions
- **Proximity-based Models:** Designs outliers as points of isolation
- **Information-theoretical models:** Increases minimum code length to describe data set

Uses for Detecting Outliers in Data Mining

- **Applications:**
  - Fraud Detection
  - Telecom Fraud Detection
  - Cyber Security Intrusion Detection
  - Medical Analysis
  - Environmental Monitoring (Cyclones, Tsunamis, Floods, Droughts)
  - Noticing unforeseen database entries

Check Your Progress-1

1. Describe the uses of cluster analysis in data mining.
2. Differentiate between Various Clustering Methods along with their description, advantages, disadvantages and algorithms available.
3. Briefly discuss Outlier and Outlier Detection.