Unit 5: Cluster Analysis

5.1 Basic Concepts of Cluster Analysis, Clustering Structures

Basic Concepts of Cluster Analysis:

1. **Cluster Analysis Definition:** Cluster analysis is a data exploration technique used to identify groups of similar objects or data points within a dataset. The goal is to partition the data into distinct clusters, where objects within the same cluster are more similar to each other compared to those in different clusters.
2. **Similarity and Dissimilarity Measures:** Cluster analysis relies on measuring the similarity or dissimilarity between objects or data points. Various distance metrics, such as Euclidean distance, Manhattan distance, or cosine similarity, are commonly used to quantify the similarity/dissimilarity between objects.
3. **Unsupervised Learning:** Cluster analysis is an unsupervised learning technique, meaning it does not rely on predefined class labels. Instead, it discovers patterns and structures in the data based on the inherent similarities or dissimilarities among the objects.
4. **Data Representation:** Cluster analysis can be performed on different types of data, including numerical, categorical, or mixed attribute data. The appropriate representation and choice of distance measure depend on the nature of the data.

Clustering Structures:

1. **Hierarchical Clustering**: Hierarchical clustering builds a hierarchy of clusters by successively merging or splitting existing clusters based on a specified criterion. It results in a tree-like structure called a dendrogram, which shows the relationships and similarities between clusters at different levels.
2. **Partitional Clustering:** Partitional clustering aims to partition the data into a fixed number of clusters, where each data point belongs to exactly one cluster. Algorithms like k-means, k-medoids, and Gaussian mixture models are commonly used for partitional clustering.
3. **Density-Based Clustering:** Density-based clustering identifies clusters based on the density of data points. It groups together data points that are closely packed and have a higher density than their surroundings. Density-based spatial clustering of applications with noise (DBSCAN) is a popular density-based clustering algorithm.
4. **Model-Based Clustering:** Model-based clustering assumes that the data is generated from a mixture of probability distributions. It involves fitting statistical models, such as Gaussian mixture models, to the data and identifying clusters based on the estimated model parameters.
5. **Fuzzy Clustering:** Fuzzy clustering allows data points to belong to multiple clusters with varying degrees of membership. It assigns a membership value to each data point indicating its degree of association with each cluster. Fuzzy c-means is a well-known fuzzy clustering algorithm.

Understanding these basic concepts and clustering structures provides a foundation for exploring and applying cluster analysis techniques in various domains.
For detailed study visit at the url

5.2 Major Clustering Approaches, Partitioning, Methods, Hierarchical Methods, Density-Based Methods, Model-Based Clustering

Major Clustering Approaches:

1. **Partitioning Methods:** Partitioning methods aim to divide the dataset into a specified number of non-overlapping clusters. They optimize an objective function to find the best partitioning. Examples include k-means, k-medoids, and CLARA (Clustering Large Applications).
2. **Hierarchical Methods:** Hierarchical methods create a hierarchical decomposition of the dataset by successively merging or splitting clusters based on a similarity criterion. They can be agglomerative (bottom-up) or divisive (top-down). Examples include agglomerative hierarchical clustering and divisive hierarchical clustering.
3. **Density-Based Methods:** Density-based methods discover clusters based on the density of data points. They group together data points that have higher densities than their surroundings. Density-based spatial clustering of applications with noise (DBSCAN) and OPTICS (Ordering Points to Identify the Clustering Structure) are common density-based clustering algorithms.
4. **Model-Based Clustering:** Model-based clustering assumes that the data is generated from a mixture of probability distributions. It involves fitting statistical models, such as Gaussian mixture models, to the data and identifying clusters based on the estimated model parameters. Expectation-Maximization (EM) algorithm is often used for model-based clustering.

Partitioning Methods:

1. **K-means:** K-means is a widely used partitioning algorithm that aims to minimize the sum of squared distances between data points and their cluster centroids. It starts by randomly initializing cluster centroids and iteratively updates them until convergence.
2. **K-medoids:** K-medoids is similar to k-means but instead of using centroids, it selects actual data points (medoids) as representatives of clusters. The algorithm finds the optimal medoids by minimizing the total dissimilarity between data points and their assigned medoids.

Hierarchical Methods:

1. **Agglomerative Hierarchical Clustering**: Agglomerative clustering starts with each data point as a separate cluster and merges the closest pairs of clusters based on a linkage criterion, such as single linkage, complete linkage, or average linkage. It continues merging until a desired number of clusters is reached.
2. **Divisive Hierarchical Clustering**: Divisive clustering begins with all data points in a single cluster and recursively splits clusters based on a similarity criterion. It continues dividing until each data point is in its own cluster or until a termination condition is met.

Density-Based Methods:

1. **DBSCAN:** Density-Based Spatial Clustering of Applications with Noise (DBSCAN) identifies dense regions as clusters based on the density of data points. It defines core points, border points, and noise points based on a specified radius and minimum number of neighboring points.

Model-Based Clustering:

1. **Gaussian Mixture Models (GMM)**: Gaussian mixture models assume that the data is a mixture of Gaussian distributions. The algorithm estimates the parameters of the Gaussian components (mean, covariance matrix) and assigns data points to clusters based on their probabilities of belonging to each component.

These approaches provide different perspectives and techniques for clustering data based on their inherent structures and characteristics. Each method has its strengths and limitations, and the choice of clustering approach depends on the specific dataset and problem at hand.

Visit at E-resources to following E-resources

Partitioning Methods (K-means, K-medoids, CLARA)

Hierarchical Methods (Agglomerative, Divisive)

Density-Based Methods (DBSCAN, OPTICS)

?

5.3 Importance of Outlier, Identifying and handling of outliers, Outlier Detection Techniques

Importance of Outliers:

Outliers are data points that deviate significantly from the majority of the data in a dataset. They can occur due to various reasons such as measurement errors, data corruption, or genuine rare events. Understanding and handling outliers is crucial in data analysis and modeling for the following reasons:

1. **Data Quality and Accuracy:** Outliers can significantly impact the quality and accuracy of data analysis. Including outliers in the analysis can lead to biased results and incorrect conclusions. Identifying and handling outliers helps ensure the integrity and reliability of the data.
2. **Data Interpretation:** Outliers can distort the interpretation of patterns, trends, and relationships within the data. By removing or appropriately handling outliers, we can obtain a more accurate understanding of the underlying patterns and relationships in the data.
3. **Model Performance:** Outliers can have a strong influence on statistical models. In some cases, outliers can disproportionately affect model parameters and lead to poor model performance. Identifying and managing outliers can improve the robustness and accuracy of the models.

Identifying and Handling Outliers:

1. **Visual Inspection:** Data visualization techniques, such as scatter plots, box plots, or histograms, can help visually identify outliers. Unusual data points that fall far outside the expected range can be flagged as potential outliers for further investigation.
2. **Statistical Methods**: Statistical techniques like z-score, modified z-score, or Tukey's fences can be used to identify outliers based on their deviation from the mean or median of the data. Data points beyond a certain threshold can be considered outliers.
3. **Domain Knowledge**: Subject matter experts or domain knowledge can play a crucial role in identifying outliers. Experts who have a deep understanding of the data and its context can recognize unusual or unexpected values that may indicate outliers.

Outlier Detection Techniques:

1. **Extreme Value Analysis:** This technique focuses on identifying data points that fall in the tails of the distribution. Methods like the box plot, percentiles, or quantiles can help detect extreme values that could be potential outliers.
2. **Distance-based Methods**: These methods measure the distance between data points and use thresholds to identify outliers. Examples include the k-nearest neighbors (kNN) approach and the Local Outlier Factor (LOF) algorithm.
3. **Statistical Models**: Statistical models, such as the Gaussian distribution or mixture models, can be used to estimate the likelihood of data points and identify those with significantly low or high probabilities. This approach is commonly used in Gaussian Mixture Models (GMM) for outlier detection.
4. **Machine Learning Techniques**: Machine learning algorithms, such as Isolation Forest, One-Class SVM (Support Vector Machines), or Autoencoders, can be trained on normal data and then used to identify data points that deviate from the learned patterns. These techniques are effective for unsupervised outlier detection.

It's important to note that the choice of outlier detection technique depends on the specific characteristics of the data and the goals of the analysis. Different techniques may be more suitable for different scenarios, and a combination of methods may be used for comprehensive outlier detection and handling.

Visit at E-resources to the following :

Outliers Detection for Temporal Data

5.4 WEB MINING: Basic concepts, of web mining, Different types of web mining, PAGE RANK Algorithm, HITS Algorithm

Web Mining:

Web mining is the process of extracting useful information and knowledge from web data, including web pages, hyperlinks, user interactions, and web logs. It involves applying data mining and machine learning techniques to understand web content, structure, and usage patterns. Web mining enables tasks such as web search, recommendation systems, user behavior analysis, and web content optimization.

Different Types of Web Mining:

1. **Web Content Mining:** This type of web mining focuses on extracting information from web pages, such as text, images, videos, and other multimedia content. Techniques like text mining, information extraction, and sentiment analysis are used to analyze and understand web content.
2. **Web Structure Mining:** Web structure mining involves analyzing the link structure of the web, including hyperlinks between web pages. It aims to discover patterns, relationships, and properties of web pages and websites. Techniques like link analysis, graph mining, and clustering are used in web structure mining.
3. **Web Usage Mining**: Web usage mining focuses on analyzing user interactions and behavior on the web. It involves studying web server logs, clickstream data, and user sessions to understand user preferences, navigation patterns, and user satisfaction. Techniques like sessionization, sequential pattern mining, and recommendation systems are used in web usage mining.

PageRank Algorithm:

PageRank is an algorithm used by Google to determine the importance or relevance of web pages. It assigns a numerical value (PageRank score) to each web page based on the number and quality of incoming links from other pages. The PageRank algorithm follows the principle that a page is important if it is linked by other important pages.

The PageRank algorithm works by iteratively calculating the PageRank score for each page in a web graph. It takes into account the damping factor (probability of following a link), the number of outgoing links from each page, and the PageRank scores of the linking pages. The algorithm converges when the PageRank scores stabilize.

HITS Algorithm:

HITS (Hyperlink-Induced Topic Search) is an algorithm used for link analysis and ranking web pages based on authority and hub scores. It assumes that authoritative pages are linked to by many hub pages, and hub pages link to many authoritative pages. The HITS algorithm iteratively computes authority and hub scores for each web page.

In the HITS algorithm, authority scores represent the importance or quality of a web page, while hub scores represent the page's ability to link to authoritative pages. The algorithm starts with initial values for authority and hub scores and iteratively updates them based on the link structure of the web. The process continues until convergence is reached.

These algorithms, PageRank and HITS, are fundamental in web mining for ranking web pages based on their importance and analyzing the link structure of the web. They are widely used in search engines, recommendation systems, and various web analytics applications.

Visit to the following E-resources:

The PageRank Citation Ranking

HITS: A Generalized Link Analysis Algorithm

WEB MINING