CenDTect Decision Tree Algorithm

Data Preprocessing

Raw censorship data collected from Hyperquack HTTPS data is organized to add in metadata including IP ownership (ISP) and domain categories such as classifying domains i.e (social media, news) Censorship is enforced at the organizational level so data is aggregated by IP organization (ISP) now processed data is ready for analysis which is organized by domain, IP organizations and time.

Decision Tree Algorithm

Main goal : Identify groups of domains that share similar censorship patterns (blocking rules) across specific times and locations.

Process

For each domain a decision tree is created utilizing the Sklearn Gini impurity algorithm to model the censorship behavior in a specific county and time. The decision tree splits the data based on features such as time, location (IP organization) and censorship method type to identify homogeneous blocking patterns. Trees have an average depth of < 16 and max < 25 ensuring interpretability.

Distance Metric

Distance metric for clustering is used in order to group domains with similar blocking rules, CenDTect uses a custom distance metric called **cross-classification** , in which measures how well the decision tree of one domain (DTi) can predict the censorship behavior of another domain (dj). A low distance indicates similar blocking rules.

Iterative Clustering

Algorithm 1 (getDomainTreeClusters)
Identifies “innocent trees” (domains with no censorship, labeled as 0 for accessible)
Removes innocent trees to focus on censored domains.

Generates initial clusters utilizing Algorithm 2 (getInitialClusters)
Creates decision trees for all domains, utilizes DBSCAN (density based clustering algorithm) to group domains with similar decision trees based on the distance metric.Then assigns domains to clusters with user defined distance thresholds and a minimum cluster size. Groups and merges clusters with similar blocking rules, then extracts censorship events (geolocation, timespan,

blocking method) by analyzing the decision trees in each cluster. Removes data associated with identified events and repeats clustering until no new events are found.

Special Prediction Function

Handles cases where a domain is blocked differently across IP organizations. For example if a decision tree predicts a blocked domain as accessible (but the tree itself is not “innocent’) the prediction is considered successful to account for overlapping censorship patterns.

Example: If twitter is blocked in ISP A but accessible in ISP B the algorithm can still cluster it with the other domains blocked in ISP A.

Output : Clusters of domains with similar censorship patterns each associated with a blocking rule

Decision Trees : Uses hierarchical splitting to identify consistent blocking patterns.
Cross Classification : A distance metric is used to measure the similarity between domains censorship behaviors.
Iterative Clustering : Groups domains with similar blocking rules and refines clusters over many iterations to handle overlapping events.

Decision trees are useful since data is naturally partitioned into homogeneous groups ideal for identifying consistent censorship patterns.