COMPSCI 762 Advanced ML

Sunday, 14 June 2020 10:22 AM

Association Rule Mining

- Frequent itemset generation (support ≥ minsup)
    - Apriori [generate-and-test]
        - supersets of an infrequent itemset must be infrequent (pruned)
    - FP-Tree & FP-Growth [projection-based]
        - scan1: order each row by support, remove infrequent single items
        - scan2: construct FP-tree (compression), build header table
        - for each item in header table
            - construct its conditional pattern base
            - construct its conditional FP-tree, remove infrequent nodes
            - generate frequent patterns
- Rules generation (confidence ≥ minconf)
    - Apriori
        - for same itemset, confidence = / ↓ when #RHS ↑ (pruned)
- Pattern evaluation
    - Interestingness measures
        - confidence, lift / interest, PS, ...
    - Simpson's paradox

Clustering

- Partitional clustering
    - K-means
    - DBSCAN [density-based]
- Hierarchical clustering (dendrogram)
    - Agglomerative
        - Inter-cluster distance
            - min (single link)
            - max (complete link)
            - group average
            - between centroids
            - ward's method (use squared error between point and centroid)
    - Divisive
- Measures of cluster validity
    - entropy (with labels)
    - SSE
    - cohesion - within cluster sum of squares
    - separation - between cluster sum of squares
    - silhouette coefficient

Data Preprocessing

- Data cleaning
    - Missing data
        - imputation
            - global constant
            - average
            - class average
            - most probable value (train a model to predict)
            - matrix decomposition, expectation maximization, ...
    - Noise
        - binning
        - regression
    - Outliers
        - clustering
- Data reduction
    - Dimensionality reduction
        - PCA (only for numeric data)
        - feature selection
            - correlation
                - nominal data: chi-squared test
                - numeric data: pearson's correlation coefficient
            - heuristic search (greedy)
                - Relief (considers all attribtues)
- Data transformation
    - Normalization
        - min-max, z-score, decimal scaling, ...
    - Discretization (continuous → intervals)
        - binning
        - histogram analysis
        - clustering
        - correlation analysis (chi-squared)
        - decision tree
- Imbalanced data
    - undersample majority class
    - oversample minority class
    - cluster-based oversampling
    - SMOTE (create new minority class instance)

Instance-based Learning

- K-Nearest Neighbour (kNN)
    - voronoi diagram
- Support Vector Machine (SVM)

    - complexity parameter C (for noise)
    - nonlinear
        - Kernel trick
            - calculate distance between points as if transformation is done
    - can be applied to regression

Bayesian Learning

- Maximum A Posteriori (MAP)
    - $\arg\max_{h \in H} P(h|D) \approx \arg\max_{h \in H} P(D|h)P(h)$
- Maximum Likelihood (ML)
    - when every hypothesis is equally probable apriori
    - $\arg\max_{h \in H} P(D|h)$
- Minimum description length
    - $\arg\max_{h \in H} P(D|h)P(h)$
      $= \arg\min_{h \in H} - \log_2 P(h) - \log_2 P(D|h)$
      = arg min (length of optimal code for H + ... for D)
      = arg min (complexity of h + #errors made by h)
    - shorter hypothesis is preferred (occam's razor)
    - trade off complexity for #errors -> prevent overfitting
- Bayes optimal classifier (like ensemble)
- Naive bayes classifier
    - laplace smoothing
    - gaussian NBC
    - multinomial NBC
- Bayesian network
    - a directed acyclic graph showing dependencies between attributes

Reinforcement Learning

- Markov decision process (S, A, R, P, γ) - memoryless
    - S: possible states
    - A: possible actions
    - R: reward for state & action
    - P: gives the next state for current state & action
    - γ: discount factor ($0 \leq \gamma < 1$)
- Goal: find optimal action policy $\pi^*: S \rightarrow A$
  ⇔ find optimal value function V* (gives value of the state)
- Value Iteration (for learning V*)
- Q-learning (without knowing P)
    - iteratively simulate markov decision process & update Q matrix

Data Stream

- Sampling
  - reservoir sampling
- Find frequent items
  - lossy counting
  - counting within a sliding window
  - exponential histogram
- Concept drift
  - drift detector
    - CUSUM - alarm when mean differs from 0
    - DDM - monitor error rate
    - ADWIN - based on exponential histogram
    - DDD - ensemble of detectors
- Performance evaluation
  - holdout
  - test-then-train
  - prequential
- Algorithm
  - Hoeffding tree
    - VFDT - with a tie threshold
    - CVFDT - deal with concept drift
  - Hoeffding adaptive tree
    - Hoeffding tree + ADWIN
  - OzaBag
  - Adaptive random forest

Anomaly Detection

- Outliers vs noise
  - Noise: random error or variance, remove first
  - Outliers: violates the normal mechanism that generates the normal data and is interesting
- Statistical (model-based)
  - assume the normal data follow a statistical model (a stochastic model)
  - parametric method
    - univariate data
    - multivariate data
  - non-parametric method
- Proximity-based
  - an object is an outlier if the nearest neighbours are far away
  - distance-based
    - distance/fraction threshold

    - density-based
        - Local Outlier Factor (LOF)
- Clustering-based
    - not belong to any cluster
        - density-based clustering, e.g. DBSCAN
    - far from its closest cluster
        - k-means
    - belong to a small/sparse cluster
        - Cluster-based Local Outlier Factor (CBLOF)
- Classification
    - One-class model, e.g. SVM
    - Semi-supervised
        - clustering-based + one-class model
- Unsupervised
    - Isolation forest
        - Assume anomalies are more easily isolated (shorter path from root)
- Evaluation
    - accuracy is not sufficient
    - use ROC curve (recall vs. FP)