Algorithms for learning from spatial and mobility data

Motivation

- Human and natural processes often have a spatial component
- Analysing location data - important in decision making
- Large quantity and complexity - efficient algorithms to extract patterns

Topics

Human mobility data

- People traveling in a city
- Similarity of trajectories
- Aggregating trips for transportation resources placement

Localised weather data

- Numerical weather forecasts for solar power prediction
- Accuracy influenced by spatial distribution of plants
- Spatial dimension influences machine learning algorithms differently

Sketches for trajectories

- Finding similar GPS trajectories is a fundamental problem
  - similarity - based on Hausdorff, Fréchet distances
- Challenges:
  - large data volume
  - intrinsic complexity of trajectory comparison
- Sketches for trajectory data
  - drastically reduce the computation costs associated with near neighbour search, distance estimation, clustering and classification, and subtrajectory detection

Sketches for trajectories - algorithms

- Locality sensitive hashing
  - hashes similar objects to the same bucket with high probability
- Sketches show what areas trajectories pass through
  - binary sketches - for Hausdorff
  - ordered sketches - for Fréchet
- MRTS (Multi-Resolution Trajectory Sketch) - layered data structure
  - used to solve queries such as distance estimation, clustering, subtrajectory detection

Sketches for trajectories - results

- Data: taxi traces from Porto and Rome (GPS locations)
- Distance based on sketches - **correlated** with Hausdorff, Fréchet, and DTW
- Distance correlation - **robust to data loss**
- Distance computation - much **faster** for both binary and ordered sketches
- Nearest neighbours search space - **pruned by 80%**
- **Better performance** (accuracy and storage) compared to other method

Transit network design

- Pairs of locations of the origin and destination of each trip
  - large-scale patterns
  - efficient placement of resources

- There are constraints on resources
  - construction (budget) constraint
  - transit distance constraint
  - total distance constraint

- Designing a transit network
  - challenging: non-convex optimisation problem

Transit network design - algorithm

- Cluster trips that are likely to be served by the same path
  - equivalent to covering segments with rectangles: NP-hard
  - heuristic for efficiency: project trips on lines
  - grid based clustering: start locally and increase resolution

- Match clusters to road network
  - lines are equivalent to shortest paths

- Simplify resulting graph
  - heuristics based on utility function of number of trips served and length for each edge

Transit network design - results

- Data
  - Santander rental system - bike journeys for one week
  - The current bike paths in London
- Set of 100 journeys - enough to cover **30% of 5000 new journeys**
- Generated network of similar size with cycling network based on 500 journeys
  - **20% more trips covered** with the most basic simplification heuristic

Regional factors in solar power prediction

- Day ahead solar power output is challenging to predict
    - fluctuations in power output determined by the movement and stochastic formation of clouds
- The spatial distribution of plants influences the range of fluctuations
- Previous work shows that less spatial correlation improves the accuracy of SVM models
- The same effect does not hold for other machine learning models

Regional factors in solar power prediction - model

- Ensemble of linear model (MLR), support vector machines (SVM) and neural network (GRU)
    - the result is the average of the predictions

- Input data:
    - persistence model
    - clear sky predictions
    - next day numerical weather forecasts

- Feature selection:
    - based on a function of the weights of parameters of linear regression, gradient boosting, SVM models
    - sequential backward selection for GRU model

Regional factors in solar power prediction - results

- Models consider predictions for the separate areas and aggregated
- Ensemble model has highest accuracy
- Global information - relevant in all models
- Aggregate prediction is more accurate only in the case of GRU
- Removing correlated features improves the performance of the SVM model, but not GRU or MLR

Future work

Sketches for trajectories

- Consider temporal dimension
- Applications to robotics, biological systems

Transit network design

- Design paths that minimise the number of turns
- Extend for other modes of transportation

Regional factors in solar power prediction

- Improve the prediction model by: considering snow and fog forecasts; different weighting method
- Further utilise spatial correlation by considering a graph neural networks