Разработка и оценка сложности алгоритмов big-data

Соколов А.П.
м.н.с., к.ф.-м.н, каф. МаТИС
sokolov@intsys.msu.ru

1.1. Градиентный бустинг решающими деревьями

- За 20 лет с момент своего появления алгоритмы градиентного бустинга решающими деревьями стали «золотым стандартом» для обработки структурированных данных в задачах ML (GDBT, XGBoost, LightGBM и др.);

- Тем не менее, за это время они существенно эволюционировали:

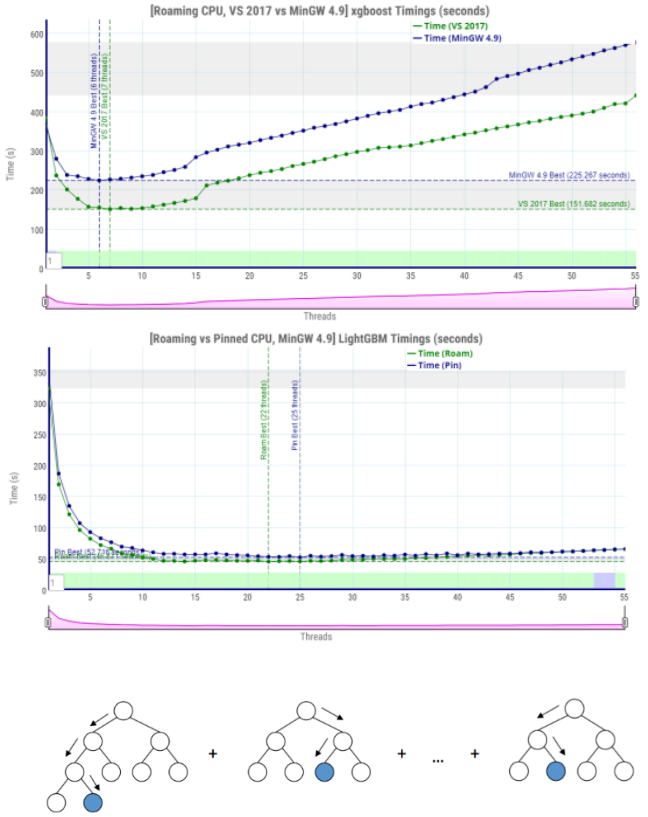
  - B. Panda, J. Herbach, S. Basu, and R. Bayardo. Planet: Massively parallel learning of tree ensembles with mapreduce. Proceedings of the Very Large Database Endowment, 2(2):1426–1437, 2009.
  - Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794. ACM, 2016.
  - Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boostingdecision tree. In Advances in Neural Information Processing Systems, pages 3149–3157, 2017.
  - Shi Yu, et. al. – Gradient Boosting with PL Regression Trees, 2019.

1.2. Некоторые постановки задач

- Разработка и эффективного распределенного алгоритма обучения ансамбля решающих деревьев;
- Разработка и оценка сложности алгоритма обучения ансамбля решающих деревьев на основе пороговых функций:

$$f_s(x_i) = sign\left(b_s + \sum_{j=1}^{m_s} b_{s,j} x_{i,k_{s,j}}\right)$$

$b_{s,j} \in B$, B – допустимое множество коэффициентов.;

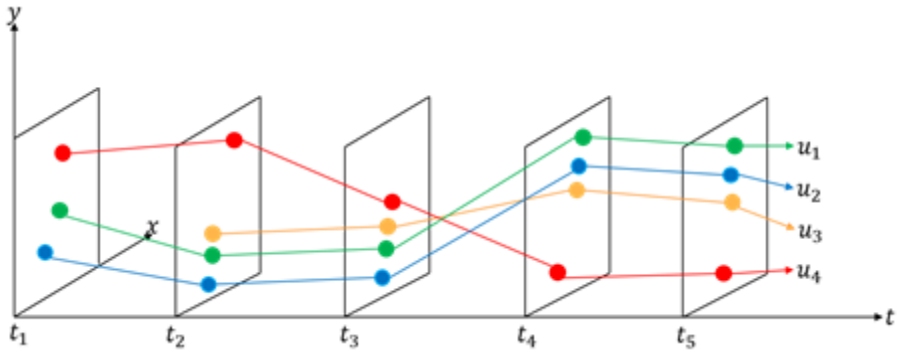


2.1. Распределенные алгоритмы обработки траекторных данных

- Сотовые телефоны и смартфоны являются источником огромных массивов траекторной информации.
- Эта информация может быть накоплена у операторов сотовых сетей и далее использована для решения множества полезных практических задач.
- Основная особенность данных задач – необходимость обработки огромных объемов информации. Как следствие – необходимость разработки алгоритмов для распределенных вычислительных систем и оценка их сложности.

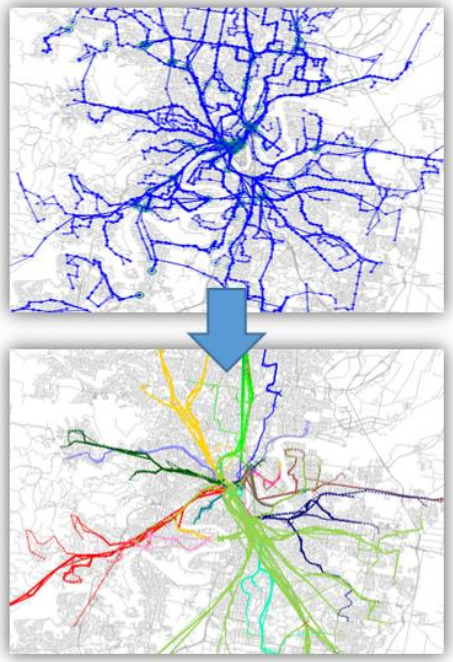
2.2. Некоторые постановки

- Поиск множеств траекторий с заданными свойствами:
- Дано:
  - множество траекторий объектов $T = \{t_1, \dots, t_N\}$ за некоторый промежуток времени.
  - Функция-предикат p на множестве пар траекторий, задающая некоторое отношение.
- Найти:
  - Все пары траекторий из T, на которых выполнен предикат p;
- В качестве предикатов могут выступать:
  - Совместное путешествие;
  - Сопровождение одним объектом другого;



2.2. Некоторые постановки (продолж.)

- Построение модели транспортной сети
- Дано:
  - множество траекторий объектов $T = \{t_1, \dots, t_N\}$ за некоторый промежуток времени.
  - модель дорожной сети (например, граф дорог);
- Найти:
  - Функцию, определяющая априорную вероятность перемещения объекта из одной вершины дорожной сети в другую;
  - Функция, которая по типу объекта, времени суток, пункту назначения и отправления, определяет наиболее ожидаемый маршрут передвижения;
  - Функция, которая по заданному маршруту передвижения определяет ожидаемое время путешествия;



2.3. Некоторые публикации по теме

- [1] Соколов А.П., Алисейчик П.А., Моисеев С.В. – Распределенный алгоритм поиска траекторий-компаньонов, Интеллектуальные системы. Теория и приложения, т.25, №1, стр. 71-90, 2021.
- [2] Li, Zhenhui, et al. "Swarm: Mining relaxed temporal moving object clusters." Proceedings of the VLDB Endowment, 3.1-2 (2010): 723-734.
- [3] Tang, Lu-An, et al. "On discovery of traveling companions from streaming trajectories." Proceedings of IEEE 28th International Conference on Data Engineering (ICDE), 186-197, 2012.
- [4] Zheng, Kai, et al. "Online discovery of gathering patterns over trajectories." IEEE Transactions on Knowledge and Data Engineering, 26.8 (2014): 1974-1988.

Спасибо за внимание!