LightGBM: A Highly Efficient Gradient Boosting Decision Tree

Prepared by:

Dursun Karaca ERDEMİR

İrem BİGAT

Topics

- Gradient Boosting Decision Tree
- LightGBM
- Gradient-based One-side Sampling
- Exclusive Feature Bundling
- Experiments
- Conclusion

Gradient Boosting Decision Tree (GBDT)

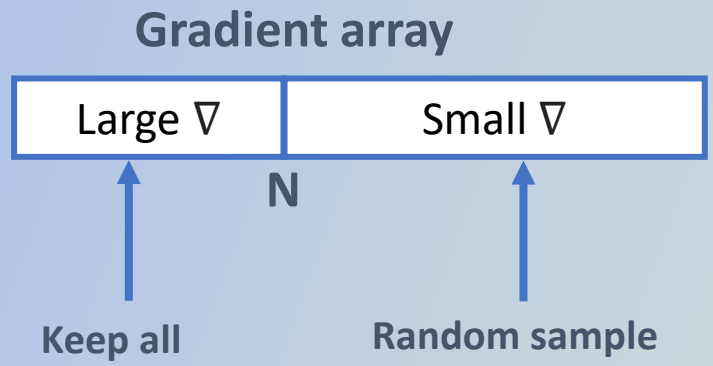
- a powerful machine learning algorithm
- used for regression, classification, ranking
- finds split points of each feature
  - Gradient Boosting = Gradient + Boosting
  - many adaptations (XGBoost, SGB etc.)
    - ensemble of weak learners (decision trees)
    - performs excellent with small data
    - performs poorly with big data and high dimensions

LightGBM

- improved GBDT
- reduced computing cost -> same accuracy
- less data & features
- two novel algorithms (GOSS & EFB)

Gradient-based One-side Sampling (GOSS)

1) Calculate error gradient
2) Sort descending order
3) Keep N largest
4) Random sample from the rest
5) Assign weight for balance
6) Reduced data size but same accuracy



Exclusive Feature Bundling (EFB)

- Made for sparse high-dimensional data
- Near-lossless
- Bundles (merges) mutually exclusive features
- Significantly speeds up training

- Two steps involved:

Greedy bundling

- Optimal bundling problem
- Reduced to graph coloring
- NP-hard problems
- Bundles features with small conflict

Merge exclusive features

- Merges each bundle
- Add offsets for value separation
- Many sparse features -> few dense features

Experiments

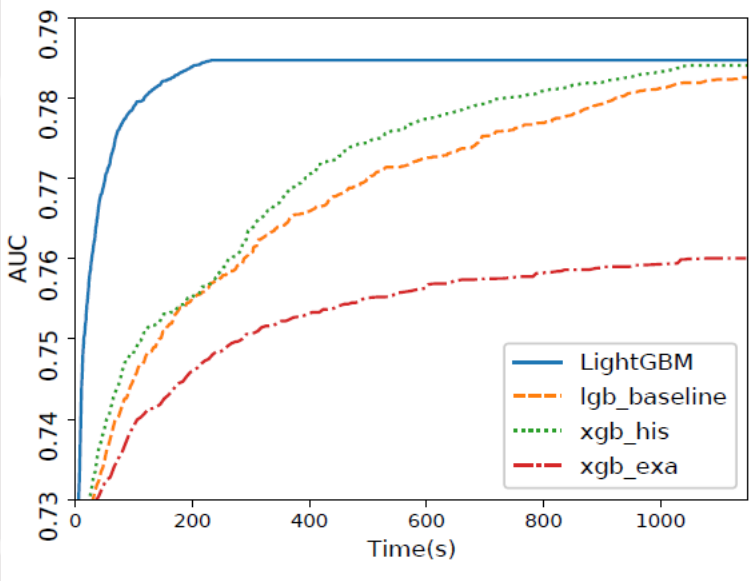


Table 1: Datasets used in the experiments.

Name	$\#data$	$\#feature$	Description	Task	Metric
Allstate	12 M	4228	Sparse	Binary classification	AUC
Flight Delay	10 M	700	Sparse	Binary classification	AUC
LETOR	2M	136	Dense	Ranking	NDCG

Table 2: Average time cost per iteration

	xgb_exa	xgb_his	lgb_baseline	EFB_only	**LightGBM**
Allstate	10.85	2.63	6.07	0.71	**0.28**
Flight Delay	5.94	1.05	1.39	0.27	**0.22**
LETOR	5.55	0.63	0.49	0.46	**0.31**

Table 3: Overall accuracy comparison

	xgb_exa	xgb_his	lgb_baseline	SGB	**LightGBM**
Allstate	0.6070	0.6089	0.6093	0.6064 ± 7e-4	**0.6093 ± 9e-5**
Flight Delay	0.7601	0.7840	0.7847	0.7780 ± 8e-4	**0.7846 ± 4e-5**
LETOR	0.4977	0.4982	0.5277	0.5239 ± 6e-4	**0.5275 ± 5e-4**

Conclusion

- GBDT -> very costly with large-scale data
- LightGBM -> faster computing, less data, same accuracy
- Downsampling (GOSS)
- Dimension reduction (EFB)
- Less space and time complexity than SGB & XGBoost

THE END

Thanks for your time