3A KDD Cup 2009 customer relationship prediction

Introduction of competition

In this task we chose a competition, which was held in year 2009, on SIGKDD's website. As described on KDD's website[1], the database comes from French Telecom company Orange to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable. The interesting part of data is that all data given to developers were scrambled. This gave a great challenge on how to clean and restructure data. For all teams, the submitted models for different tasks were evaluated by Area Under the ROC Curve (AUC) performance, then get the average AUC as finally ranking score.

It had two track, the first one is slow track with 5 days limitation, and another one is slow track with longer extension and a reduced version of data set is given. For there is no precise score info on slow track we will only discuss about Fast Track and the rank of competition is given below.

Rank	Team Name	Method	Churn	Appetency	Upselling	Score
1	IBM Research	Final Submission	0.7611	0.8830	0.9038	0.8493
5	Financial Engineering Group	boosting	0.7498	0.8732	0.9057	0.8429

Competition winner (IBM Research) related algorithm[2]

Their overall strategy is to solve these problems was Ensemble Selection (Caruana and Niculescu-Mizil, 2004)[3]. The idea came from Statistical ensemble, it is not a particular Machine Learning method or Model, but a mechanics that group multiple models with different weights and generate a new model to solve the target problem. For different model could have different strengths and weaknesses, Ensemble Selection can take advantage of them and get better predictions.

They included 13 different base level models, like Random Forests, Boosted Decision Trees, Regularized Logistic Regression, SVM and so on and also tested different model's prediction result on 3 problems. They finally found that, the best model on churn were boosted trees, the best single method on appetency was random forests, and the best single method on up-selling were boosted trees. In order to all models have the same grading skill they applied post training calibration using Platt Scaling .(Niculescu-Mizil and Caruana, 2005)[4].

Then they did ensemble selection, by using greedy forward stepwise classifier selection and use two validation folds to test the performance.

Compare to non-winning methods

When we view non-winning methods most of them used boost trees methods, which is also described in IBM Research's paper. Compared to IBM Research's group Multi-model structure took all different single model's advantage, which is a brilliant methods, and Of course better than single model.

Then we comes to boost trees[5], we found that it also have relevant grouping advantage of sub-models thinking in implementation. It aims to iteratively build an ensemble of weak learners, in an attempt to generate a strong overall model. From this these idea, we can see that grouping base level's existing classifiers in proper way is an important and strong method in machine learning

[1] https://www.kdd.org/kdd-cup/view/kdd-cup-2009

[2] Alexandru Niculescu-Mizil, Claudia Perlich, Grzegorz Swirszcz, Vikas Sindhwani, Yan Liu, Prem Melville, Dong Wang, Jing Xiao, Jianying Hu, Moninder Singh, Wei Xiong Shang, Yan Feng Zhu, Winning the KDD Cup Orange Challenge with Ensemble Selection, 2009

[3] Rich Caruana and Alexandru Niculescu-Mizil. Ensemble selection from libraries of models. In
Proceedings of the 21st International Conference on Machine Learning (ICML'04), 2004.

[4] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learn- ing. In Proceedings of the 22nd International Conference on Machine Learning (ICML'05), 2005.

[5] SACHIN JOGLEKAR,A (small) introduction to Boosting, 2016