Секция «Вычислительная математика, математическое моделирование и численные методы»

Сравнительный анализ базовых методов машинного обучения

Сивков Егор Сергеевич
Студент (магистр)
Московский физико-технический институт, Москва, Россия
E-mail: sivkovegor@yandex.ru

Данное исследование является вступительной работой для написания магистерского диплома. Его целью является определение наиболее эффективного метода анализа данных для дальнейшего его использования и сравнения с более сложными алгоритмами обработки данных. Исследование было проведено с помощью программного обеспечения SAS Enterprise Miner [3].

Для построения моделей были взяты следующие методы: Decision Tree, Interactive Decision Tree, Gradient Boosting, Logistic Regression, Scorecard и Neural Network. Для анализа использовались два набора данных: German Credit Data из 1000 записей с 20 предикторами [2] и данные ОТП Банка из 15 223 записей с 50 предикторами [1]. В первом наборе данных пропущенные значения отсутствовали, поэтому для алгоритмов Logistic Regression, Scorecard и Neural Network, требующих обработки пропущенных значений на втором наборе данных, были использованы два разных способа замены пропущенных значений: методом медианы и методом построения дерева выбора. Обе выборки были разбиты на группы обучения и проверки в пропорции 55% на 45%. Качество работы алгоритмов было оценено с помощью базовых показателей AUC (area under curve) — площади под ROC-кривой, вычисляемой по контрольным данным и связанного с ним индекса-gini.

В результате моделирования было получено, что среди выбранных методов для обоих наборов данных наиболее подходящим для построения прогнозной модели является метод Gradient Boosting с показателями AUC - 0.872, Gini - 0.744, AUC - 0.713, Gini - 0.426 для групп обучения AUC - 0.774, Gini - 0.548, AUC - 0.664, Gini - 0.327 для групп проверки малого и большого наборов соответственно.

Среди алгоритмов замены пропущенных значений лучше себя проявил метод подстановки с помощью деревьев выбора. Для построения скоринговой карты ни один из методов не дает существенного преимущества, в то время как для логистической регрессии, показавшей значения AUC - 0.597, Gini - 0.195, AUC - 0.527, Gini - 0.054 на группе проверки для дерева выбора и медианы, преимущество замены построением дерева очевидно.

Выигрыш модели Gradient Boosting был ожидаем, так как за счет построения различных деревьев с вариацией предикторов и набора данных для обучения получилось более полно охватить особенности данных и взаимосвязь различных предикторов с предсказываемой величиной.

Источники и литература

1) Полигон.machinelearning.py: http://poligon.machinelearning.ru/Contests/Card.aspx?synonim
2) Репозиторий машинного обучения UCI: https://archive.ics.uci.edu/ml/datasets/Statlog+(Ge
3) Сас.ком: http://www.sas.com/en_us/industry/higher-education/on-demand-for-academics.html