УДК 004.942

ИССЛЕДОВАНИЕ МЕТОДОВ ГЕНЕРАЦИИ МНОГОМЕРНЫХ ВЫБОРОК ДЛЯ ЗАДАЧ АНАЛИЗА ИЗМЕНЧИВОСТИ ПОГОДНЫХ ХАРАКТЕРИСТИК

Плесовская Е.П. (Университет ИТМО)
Научный руководитель – к.т.н. Иванов С.В.
(Университет ИТМО)

Работа посвящена проблеме генерации многомерных выборок на основе ограниченного набора исходных данных. В докладе будет представлен сравнительный анализ синтетических выборок, сгенерированных на основе различных генеративных моделей. В качестве эмпирических данных использовался набор погодных характеристик, содержащих информацию о состоянии облачности.

Генерация синтетических погодных данных может использоваться как для анализа их изменчивости, так и для генерации целевых переменных, для которых установлена зависимость от погодных характеристик. Ранее авторами доклада была разработана модель, которая на основе данных о состоянии облачности во время сумерек позволяет сгенерировать значения естественной освещенности. В результате при наличии достаточного объема данных о состоянии облачности можно оценить распределение значений естественной освещенности и получить интервальные оценки для оптимального графика работы осветительного оборудования. Однако при генерации факторов облачности возникает ряд трудностей, связанных с ограниченным объемом доступной выборки (до 1000 набл.) при относительно большом количестве переменных (6) и достаточно большим объемом наблюдений (до 50%) с краевыми дискретными значениями (0 или 1) при непрерывности распределений переменных.

В данной работе задача генерации многомерных выборок решается с помощью обучения модели, которая аппроксимирует распределение исходных данных. Одним из наиболее простых и эффективных методов обучения генеративной модели на основе небольшой выборки является ядерная оценка плотности. При использовании данного метода результат оценки существенно зависит от параметра сглаживания. В данном исследовании сравниваются синтетические выборки, сгенерированные на основе моделей, использующих различные методы подбора параметра сглаживания: кросс-валидацию на основе метода максимального правдоподобия, кросс-валидацию на основе метода наименьших квадратов, адаптивный подбор параметра сглаживания. Кроме того, предлагается подход, учитывающий наличие краевых дискретных значений в исходных данных при генерации значений с помощью ядерной оценки плотности. Для оценки качества генерируемых выборок использовались тесты, сравнивающие исходную и синтетическую выборки на основе оценки расстояния между их совместными, маргинальными распределениями, а также на основе классификации обучающей и синтетической выборок.

В ходе исследования был проведен сравнительный анализ синтетических выборок, полученных на основе различных генеративных моделей, обученных по выборке с небольшим объемом (до 1000 набл.). По его результатам были описаны достоинства и недостатки методов на основе ядерной оценки плотности для целей генерации погодных характеристик. Также были выделены тесты для сравнения исходной и синтетической выборок, которые наилучшим образом подходят для решаемой задачи.