УДК 004.942

МУЛЬТИМОДАЛЬНАЯ ГЕНЕРАЦИЯ ПЕРСОНАЛЬНЫХ СИНТЕТИЧЕСКИХ ДАННЫХ

Лысенко А.В. (Университет ИТМО)
Деева И.Ю. (Университет ИТМО)
Шиков Е.Н. (Университет ИТМО)
Научный руководитель – к.т.н., Боченина К.О.
(Университет ИТМО)

В данном исследовании рассмотрены нейросетевые подходы генерации мультимодальных синтетических данных на основе вариационных автоэнкодеров для решения задач генерации и восстановления данных. В качестве данных рассматривались транзакционные данные пользователей вместе с профилем социальной сети и вектором интересов, построенном на основе принадлежности пользователя определенным группам социальной сети. Проведенные эксперименты показывают эффективность применения таких моделей в рассматриваемых задачах, а также показывают возможность генерации данных с заданными характеристиками.

Генерация синтетических персональных данных в настоящее время используется в нескольких приложениях с проблемами конфиденциальности, таких как обучение и тестирование систем для анализа поведения пользователей социальных сетей или клиентов банка. Очень часто личные данные являются сложными и описывают различные аспекты личности (демографические данные, интересы, данные о транзакциях и другие), некоторые из которых могут отсутствовать в некоторых записях, что затрудняет работу с ними. В этой работе рассмотрена задача создания синтетических персональных данных как проблема нескольких модальностей с отсутствующими модальностями.

В качестве рассматриваемых решений использовались мультимодальные вариационные автоэнкодеры, объединяющие вектора скрытого пространства с помощью операции Product of Experts (PoE). Такие модели могут обучаться на данных с пропусками, что позволяет решать следующие задачи: 1) обучение на данных с отсутствующими модальностями; 2) генерация реалистичных синтетических данных профилей пользователей в социальных сетях; 3) восстановление отсутствующих модальностей пользователя; 4) генерация реалистичных транзакционных данных. Также рассмотрена возможность генерации пользователей с заданными характеристиками, что позволяет генерировать разнообразные профили для моделирования пользователей разных типов.

Была проведена серия экспериментов, показывающая эффективность рассматриваемых моделей для генерации как отдельно взятых модальностей, так и совместно генерируемых, восстановления пропущенных модальностей и генерации данных с заданными характеристиками.

Лысенко А.В. (автор)

Боченина К.О. (научный руководитель)