УДК 519.712.2

Применение метода K-Means в задаче оценки характеристик процесса применительно к веб приложениям

Евстратов В.В. (СПбГУ),

Научный руководитель – кандидат физико-математических наук, доцент Ананьевский М.С.

(СПбГУ)

В последние годы методы машинного обучения уже не раз продемонстрировали свой потенциал в различных веб приложениях. Авторами была рассмотрена задача прогнозирования характеристик пользователей, и предложен подход на основе кластеризации методом K-means. Полученные аналитические результаты были подтверждены на практике.

Рассматривается задача оценки характеристик процесса в дискретном времени. Такая задача может быть формализована несколькими способами - например, как задача линейной регрессии, классификации, или как задача кластеризации. Каждый из этих способов уже применялся в отечественной и зарубежной практике, и выбор конкретного алгоритма, а также выбор представления данных позволяет каждый раз исследовать задачу с разных сторон.

Авторы предлагают решение задачи прогнозирования характеристик процесса в дискретном времени с помощью подхода на основе кластеризации. В данном подходе используется гипотеза о том, что процессы, схожие на первых отсчётах, будут обладать близкими характеристиками и в дальнейшем. В качестве алгоритма кластеризации предлагается использовать K-Means.

Предложенное решение было протестировано в задаче прогнозирования трат пользователей веб-приложения на тридцатый день после регистрации по данным об их активности в первые пять дней. То есть, действия отдельно взятого пользователя рассматривались как процесс, дискретный по времени (с периодом дискретизации 1 день), а траты пользователя выступали в качестве оцениваемой характеристики.

Каждый пользователь был представлен как десятимерный вектор, содержащий отсчёты времени, проведённого в приложении и трат в каждый из первых пяти дней жизни. Далее была проведена кластеризация построенного таким образом множества векторов методом K-means, и проверена гипотеза о том, что пользователи, попавшие в один кластер, к тридцатому дню жизни окажутся также близки по своим финансовым характеристикам.

Предложенное решение хорошо себя показало на исторических данных, и было применено на практике, также оправдав своё использование. В ходе экспериментов были выявлены некоторые слабые места решения, которые будут исследованы в дальнейшей работе.

Евстратов В.В. (автор)

Ананьевский М.С. (научный руководитель)