УДК 338.2

МЕТОДИЧЕСКИЕ И ПРАКТИЧЕСКИЕ АСПЕКТЫ КЛАСТЕРИЗАЦИИ КЛИЕНТОВ ПРЕДПРИЯТИЯ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ BI-СИСТЕМ И МАШИННОГО ОБУЧЕНИЯ

Ярская А.А. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – д.э.н., профессор Цуканова О.А.
(Университет ИТМО, г. Санкт-Петербург)

Аннотация

Работа посвящена анализу методических аспектов и разработке практических рекомендаций для автоматизированной кластеризации клиентов предприятия на основе интеграции BI-систем и алгоритмов машинного обучения. В исследовании раскрыты возможности автоматизации алгоритма кластеризации через интеграцию BI-систем с алгоритмами машинного обучения на языке Python. В методической части работы также определены особенности разных алгоритмов кластеризации: K-means, DBSCAN, агломеративная кластеризация. В практической части работы выполнена кластеризация клиентов службы доставки ресторанов в г. Санкт-Петербурге на основе рассмотренных методов, раскрыты достоинства и ограничения каждого подхода.

Введение. Business Intelligence – системы уже стали стандартными инструментами в работе большинства российских компаний. Гибкость BI-систем, легкость масштабирования и внедрения позволяет им удовлетворять потребности в аналитике как гигантов рынков, так и малого бизнеса в любых отраслях экономики. В настоящее время BI-система – это не только инструмент для аналитиков: ориентация вендоров на self-service аналитику позволила вовлечь в процесс создания отчетности сотрудников компании независимо от их должностей и направлений работы. BI-инструменты позволяют сократить время на сбор и обработку данных, что так важно для экономики «в режиме реального времени», где выделение среди конкурентов основано уже не на качестве самого продукта, а на скорости его доставки и умении быстро реагировать на различное поведение клиентов.

Следующим этапом развития BI-систем стало внедрение в них искусственного интеллекта. Именно на этом направлении и сконцентрировали усилия большинство разработчиков BI-платформ. Искусственный интеллект и машинное обучение, в частности, позволяют автоматизировать продвинутую аналитику данных, проводить более глубокий анализ данных и процессов организаций. Если раньше потребности компаний реального сектора экономики в проектах машинного обучения были редки и даже единичны, то сейчас задачи машинного обучения становятся повседневной потребностью для решения задач оперативного управления или повышения эффективности работы компании в целом.

Практические аспекты автоматизации аналитики данных на основе инструментов машинного обучения и BI-систем раскрыты в работе на примере разработки соответствующего решения для службы доставки сети ресторанов, расположенной в г. Санкт-Петербурге.

Цель – раскрыть методические и практические аспекты задачи кластеризации на основе использования BI-систем и алгоритмов машинного обучения.

Предмет исследования - процесс автоматизации анализа данных.

Объект исследования - BI – системы третьего поколения с поддержкой искусственного интеллекта.

Основная часть. Для решения задачи были рассмотрены несколько различных алгоритмов кластеризации – K-means, DBSCAN, алгомеративная (иерархическая) кластеризация. Все алгоритмы тестировались на одной и той же выборке, состоящей из 13 454 объектов. Кластеризация выполнялась по четырем признакам: общее количество заказов клиентом, средний чек клиента в рублях, количество дней с даты последнего заказа, разница между датой последнего заказа клиента и датой его первого заказа в днях. Наилучшие результаты показал алгоритм K-means. Было выделено четыре кластера – постоянные

клиенты, новые клиенты, потерянные клиенты, клиенты с высоким средним чеком. Алгоритм K-means не только выделил кластеры, интерпретируемые с точки зрения бизнес-метрик, но и имел преимущество перед другими алгоритмами по сложности и скорости вычислений. Алгоритм K-средних хорошо работает с кластерами высокой плотности, вытянутыми и расположенными близко друг к другу. Именно поэтому он был выбран в качестве итогового алгоритма для реализации проекта.

Выводы. По результат моделирования в BI-системе с поддержкой искусственного интеллекта был выбран алгоритм кластеризации K-means. Алгоритмы агломеративной кластеризации и DBSCAN не выделили кластеры потерянных и новых клиентов, так же они чувствительны к выбросам, что является неблагоприятным фактором для кластеризации в автоматическом режиме, так заранее предугадать, какие выбросы могут появиться в сырых данных невозможно.

Ограничением проекта является особенность метода кластеризации K-means – количество кластеров в нем задается исследователем в виде параметра на основе предварительного анализа данных и экспертной оценки. Риск состоит в том, что данные и количество кластеров на их основе могут изменяться в процессе функционирования компании. Однако распределение объектов претерпевает существенные изменения только при наличии сильного внешнего воздействия – например, полной переориентации компании и перехода на другой сегмент рынка. В таких случаях необходимо проводить переобучение построенной модели.

Ярская А.А. (автор) Подпись

Цуканова О.А. (научный руководитель) Подпись