# Исследование методов и алгоритмов для решения задач предиктивной аналитики

К.К. Симовин, А.В. Благов

Самарский университет, Самара, Россия

**Обоснование.** Предиктивная аналитика позволяет прогнозировать настоящее и формировать будущее, основываясь на накопленных данных с применением методов интеллектуального анализа. В процессе обработки и анализа данных выделяются факторы, которые приводят к получению новой (спрогнозированной) информации [1].

**Цель** — исследовать этапы построения предиктивной модели с использованием методов интеллектуального анализа данных.

**Методы.** Данные, необходимые для проведения предиктивной аналитики, принято делить на внутренние и внешние. К первым относятся данные, собранные внутри той или иной исследуемой системы различными методами. Ко второму типу данных относятся все данные за пределами исследуемой системы. Такие данные добываются с использованием парсеров — программ, автоматизирующих добычу данных. Популярными методами добычи внешних данных являются веб-скрапинг, API и запросы к базам данных.

Производится очистка собранных данных от ошибок в столбцах и строках. Типичными ошибками (аномалиями) в столбцах являются недопустимые значения, пропущенные значения, орфографические ошибки, многозначность, перестановка слов и вложенные значения. К ошибкам в строках можно отнести нарушение уникальности, дублирование записей, противоречивость записей и нарушение логических связей между признаками.

Далее следует выбор методов построения предиктивных моделей. Так как предиктивный анализ бывает связан с машинным обучением, можно выделить основные типы аналитики: контролируемое и неконтролируемое обучение [1]. При первом типе аналитики выстраиваются прогнозы, которые основаны на входных данных и предпочтительных выходах — «учителях». Целью контролируемого обучения является изучение общего паттерна, по которому входные данные сопоставляются с выходными. Данный тип разделяется на две категории: регрессию для количественных ответов и классификацию для бинарных [1]. Регрессия позволяет прогнозировать зависимую переменную на основе значений факторов. Основой предиктивной модели является связь между предполагаемым результатом и предикторными переменными. Классификация позволяет разбивать данные на категории, например, по уровню дохода — высокий, средний и низкий

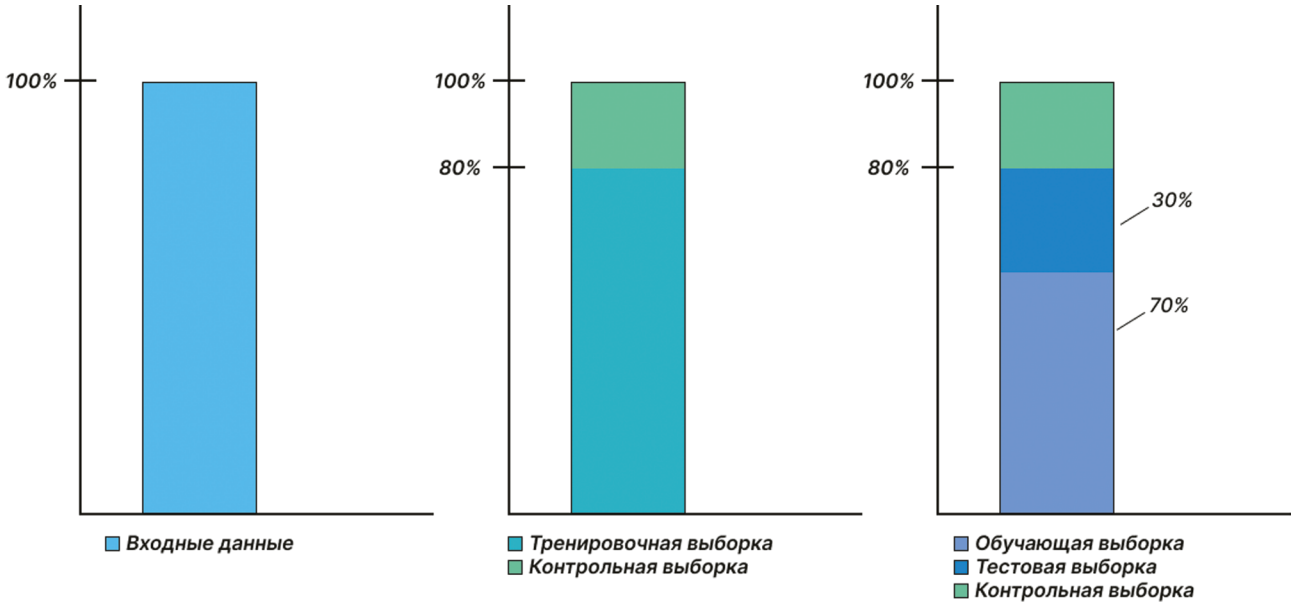


**Рис. 1.** Визуализация разбиения входных данных на выборки

доход. Классификатор исследует набор данных, в котором каждое наблюдение содержит информацию о переменной ответа и предикторах [2]. Главной задачей неконтролируемого обучения является обнаружение общих паттернов во входных данных. Для решения этой задачи используется кластеризация, так как с ее помощью возможно выявление взаимосвязей между наборами данных. Процесс кластеризации иногда называют сегментацией, потому что он сегментирует данные по категориям, чтобы идентифицировать кластер коррелирующих результатов.

Входные данные разбиваются на три типа выборок: тестовую, обучающую и валидационную (контрольную). По тестовой выборке будет оцениваться качество построенной модели, она не должна пересекаться с обучающей. Обучающая выборка используется для обучения модели. По контрольной выборке происходит выбор наилучшей модели. Она также не пересекается с обучающей.

Массив входных данных делится на выборки в следующем отношении (рис. 1):

– все данные делятся в случайном порядке в соотношении 80/20 (20 % отводится под контрольную выборку, 80 % — под тренировочную);
– тренировочная выборка делится на обучающую и тестовую выборки в соотношении 70/30.

Финальный этап построения модели — ее оценка. Сначала проверяется точность построенной модели на тестовой выборке. Если точность устраивает, происходит финальная проверка на контрольной выборке на наборе данных, которые модель до этого не использовала. Окончательно модель будет построена тогда, когда точность модели по тестовой и контрольной выборкам совпадет.

**Результаты.** Определены три основных метода сбора данных; выявлены аномалии при очистке данных; описаны методы построения предиктивных моделей; описаны выборки, на которые разбиваются исходные данные; изучен процесс оценки построенной модели.

**Выводы.** В данной работе были рассмотрены этапы построения предиктивной модели.

**Ключевые слова:** построение предиктивной модели; кластеризация; регрессия; классификация; предиктивная аналитика.

## Список литературы

1. Калытюк И.С., Французова Г.А., Гунько А.В. К вопросу выбора методов предиктивного анализа данных социальных медиа // Автоматика и программная инженерия. 2019. № 4. С. 9–17.
2. Соборнов Т.И., Ковалев И.О. Актуальность и возможности предиктивной аналитики // Научный Лидер. 2022. № 47. С. 20–21.
3. Акулин Е.В., Свиридова Л.Е. Машинное обучение // Сборник статей по итогам Международной научно-практической конференции: «Проблемы современных интеграционных процессов и поиск»; Март, 4, 2022; Стерлитамак. Стерлитамак: АМИ, 2022. С. 59–61.
4. Черкасов Д.Ю., Иванов В.В. Машинное обучение // Наука, техника и образование. 2018. № 5. С. 85–87.

*Сведения об авторах:*

**Кирилл Константинович Симовин** — студент, группа 6307-010302D, институт информатики и кибернетики; Самарский университет, Самара, Россия. E-mail: ksimovin@bk.ru

**Александр Владимирович Благов** — научный руководитель, кандидат технических наук, доцент; Самарский университет, Самара, Россия. E-mail: blagov@ssau.ru