УДК 004.85

Г.И. Гаптуллазянова, ст. преп., Д.С. Осипова, лаборант
(КНИТУ-КАИ им. А.Н.Туполева, г. Казань, Россия)

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ

В современном мире освоение новых навыков не представляет особой сложности, однако для получения глубоких и основательных знаний необходимо обучение в высших учебных заведениях. Основной целью любого вуза является подготовка сильных специалистов, способных значительно повысить качество производства в экономике. И успеваемость обучающегося в определенной степени показывает уровень освоения им этих знаний.

Однако стоит понимать, что на успеваемость могут влиять и внешние факторы (работа, наличие семьи и пр.), а не только само выполнение учебных программ. Поэтому необходимо проанализировать все возможные факторы для точного прогнозирования успеваемости обучающихся.

В целом задача прогнозирования – это процесс, подразумевающий обработку больших объемов данных и использование различных аналитических методов для предсказания требуемых показателей. В наше время задача прогнозирования в полной мере решается средствами машинного обучения, которое предоставляет большой спектр методов интеллектуального анализа данных.

Рассмотрим некоторые методы, которые часто используются для задачи прогнозирования.

Метод k-ближайших соседей. Суть данного алгоритма заключается в том, что прогнозирование значений новых данных основано на их близости к уже маркированным данным в обучающем наборе.

Достоинства метода:

– прост в реализации;

– легко обосновать результаты прогнозирования, ссылаясь на схожесть с уже известными решениями;

– не требует обучения, так как лишь сравнивает с уже известными значениями [1].

Недостатки:

– требует много памяти, т.к. необходимо хранить все объекты, которые составляют параметры метода;

– при большом объеме выборки значительно увеличивается вычислительная нагрузка, потому что вычисляются расстояния каждого объекта относительно всех других;

– проклятье размерности [1].

Данный метод очень хорошо подходит для систем с динамическим поступлением информации, т.к. в качестве обучающей выборке нужно учитывать только те объекты, которые мы недавно пронаблюдали.

Метод опорных векторов. Подразумевает, построение гиперплоскостей (границ), которые оптимально разделяют пространство данных выборки. Опорные векторы – ближайшие точки данных к гиперплоскости, которые играют решающую роль при выборе гиперплоскости и зазора (расстояние между опорными векторами и гиперплоскостью) [2].

Достоинства:

– высокая точность прогнозов за счет оптимального разделения гиперплоскостями;

– возможность обработки многомерных данных без их предварительного преобразования или понижения размерности.

Недостатки:

– неустойчивость к шуму: выбросы становятся опорными объектами-нарушителями и напрямую влияют на построение разделяющей гиперплоскости.

– низкая скорость работы на данных большого размера [3].

Дерево решений. Цель состоит в том, чтобы создать иерархическую древовидную структуру, состоящей из элементов двух типов – узлов и листьев. В каждом узле алгоритм выбирает признак и соответствующий порог, который максимизирует критерий разделения на листья. За счет обучающего множества правила разделения генерируются автоматически в процессе обучения.

Достоинства деревьев решений:

– простой в реализации;

– возможность визуализировать деревья для определения правил разделения;

– способно работать как с числовыми, так и с категориальными данными.

Недостатки метода:

– обучающие деревья решений могут создавать слишком сложные деревья, которые плохо обобщают данные;

– деревья решений могут быть нестабильными, поскольку небольшие изменения в данных могут привести к созданию совершенно другого дерева [4].

Наивный байесовский классификатор. Данный метод является вероятностным классификатором на основе формулы Байеса со стро-

гим (наивным) предположением о независимости признаков между собой при заданном классе

Преимущества:

– простой в реализации;

– высокая скорость работы и точность прогнозов во многих ситуациях;

– имеет относительно хорошую устойчивость к шуму и выбросам, поскольку основан на вероятностных распределениях и наивном предположении о независимости признаков.

Недостатки:

– в случае нарушения предположения о независимости признаков, точность прогнозов может значительно снизиться;

– может отдавать предпочтение к классам с большим количеством образцов в случае несбалансированных данных [5].

Нейронные сети. Суть алгоритма заключается в имитации работы человеческого мозга для решения разнообразных задач. Нейрон –единичный простой вычислительный процессор, способный воспринимать, преобразовывать и распространять сигналы. Нейросеть состоит из трёх типов слоёв: входного, скрытого, и выходного. На вход подаётся исходный набор данных; в скрытых слоях происходит каждый нейрон обрабатывает входящую информацию и передаёт её дальше; на выходном слое формируется конечный результат или прогноз [6].

Достоинства метода:

– нейросети способны выявить скрытые закономерности в данных;

– самообучаемость. Алгоритм самостоятельно принимает решения о том, как выполнить заданную задачу;

– нейронные сети быстро адаптируются к возможным переменам во входных данных и продолжают работать.

– Недостатки нейронных сетей:

– нейросети требуют точной настройки параметров для корректной работы, это процесс итерационный;

– очень требовательны к качеству исходных данных, иначе возможны недообучение или тупик;

– нейронные сети являются чёрными ящиками: из них нельзя получить данные о том, как было получено решение [7].

В заключение приведем сравнительный анализ рассмотренных методов машинного обучения для прогнозирования целевого значения в виде таблицы.

Таблица – Сравнительный анализ методов

	Метод k-ближайших соседей	Метод опорных векторов	Дерево решений	Наивный байесов классификатор	Нейронные сети
Оптимальный размер выборки данных	малая	средняя	средняя	большая	большая
Сложность реализации и настройки	простой	простой	простой	простой	средний
Скорость вычислений	понижается с увеличение выборки данных	высокая	высокая	высокая	высокая
Устойчивость к шумам	не устойчив	не устойчив	не устойчив	устойчив	устойчив
Точность прогнозов при оптимальном размере выборки	высокий	высокий	высокий	высокий	высокий

Каждый из методов имеет свои преимущества и недостатки, и выбор метода прогнозирования зависит целиком и полностью от специалиста, который будет им заниматься. Перед выбором метода необходимо произвести анализ данных, оценку качества выборки и при необходимости провести балансировку данных, точно определить целевую переменную, которую необходимо спрогнозировать и только затем выбирать способ прогнозирования, причем надо опробовать несколько алгоритмов для определения наиболее подходящего способа под решаемую задачу.

ЛИТЕРАТУРА

1. Анализ метода К ближайших соседей [Электронный ресурс]. – Режим доступа: https://deepmachinelearning.ru/docs/Machine-learning/Metric-methods/KNN-analysis (дата обращения: 10.01.2025).

2. Support Vector Machine: классификация данных с помощью метода опорных векторов [Электронный ресурс]. – Режим доступа: https://blog.skillfactory.ru/svm-metod-opornyh-vektorov/ (дата обращения: 13.01.2025).

3. Метод опорных векторов (SVM). Подходы, принцип работы и реализация с нуля на Python [Электронный ресурс]. Режим доступа: https://habr.com/ru/articles/802185/ (дата обращения: 15.01.2025).

4. Деревья решений (Decision Trees) [Электронный ресурс]. – Режим доступа: https://scikit-learn.ru/stable/modules/tree.html# (дата обращения: 16.01.2025).

5. Наивный байесовский классификатор. Основная идея, модификации и реализация с нуля на Python [Электронный ресурс]. – Режим

доступа: https://habr.com/ru/articles/802435/ (дата обращения: 17.01.2025).

6. Методы искусственного интеллекта: учебное пособие / Н.В. Андреянов, Т.С. Евдокимова, А.Д. Павлов, А.С. Сытник, М.П. Шлеймович – Казань, 2023. – 342 с. – URL: https://elibs.kai.ru/_docs_file/623/HTML/index.html

7. Достоинства и недостатки нейронных сетей [Электронный ресурс]. Режим доступа: https://bewave.ru/blog/dostoinstva-i-nedostatki-neyronnykh-setey/ (дата обращения: 19.01.2025).

УДК 004.05

В.В. Гедранович, доц. (БГЭУ, г. Минск)

ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ В ТЕСТИРОВАНИИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Современные технологии анализа данных играют важную роль в оптимизации процессов тестирования. Они позволяют собирать, обрабатывать и анализировать огромные объемы информации о поведении программного обеспечения (ПО), выявлять аномалии, которые могут указывать на возможные ошибки или недостатки. Внедрение алгоритмов машинного обучения и искусственного интеллекта в процессы тестирования способствует не только автоматизации рутинных задач, но и повышению точности и эффективности выявления дефектов. Анализ данных помогает тестировщикам принимать более обоснованные решения, улучшая качество конечного продукта и сокращая время вывода его на рынок.

Рассмотрим следующие технологии анализа данных, применяемые в тестировании ПО: Big Data, Машинное обучение, Data Mining, Аналитика в режиме реального времени.

Big Data, или большие данные, представляет собой большие объемы сложных и неструктурированных данных, которые постоянно генерируются и поступают из различных источников. Они характеризуются тремя основными аспектами, известными как три «*V*»: объем (*Volume*), скорость (*Velocity*) и разнообразие (*Variety*). Также иногда добавляют четвёртое «V» – достоверность (*Veracity*), что указывает на необходимость проверки качества и надежности данных, и пятое «V» – ценность (*Value*), чтобы подчеркнуть значимость извлечения полезной информации.

Использование технологий Big Data позволяет оптимизировать тестовые процессы: