УДК 004.051

ПРОБЛЕМЫ ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ

П.В. ДМИТРИЧЕНКО
(Представлено: канд. физ.-мат. наук, доц. Ю.Ф. ПАСТУХОВ)

Изучены наиболее важные технические проблемы, с которыми сталкивается машинное обучение как прикладная наука при внедрении машинного обучения в существующие информационные системы.

Введение. Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Различают два типа обучения. Обучение по прецедентам, или индуктивное обучение, основано на выявлении общих закономерностей по частным эмпирическим данным. Дедуктивное обучение предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний. Дедуктивное обучение принято относить к области экспертных систем, поэтому термины машинное обучение и обучение по прецедентам можно считать синонимами.

Машинное обучение находится на стыке математической статистики, методов оптимизации и классических математических дисциплин, но имеет также и собственную специфику, связанную с проблемами вычислительной эффективности и переобучения. Многие методы индуктивного обучения разрабатывались как альтернатива классическим статистическим подходам. Многие методы тесно связаны с извлечением информации и интеллектуальным анализом данных (Data Mining).

При разработке информационных продуктов, использующих технологии машинного обучения, возникают проблемы, большая часть которых нетипичная для экспертных систем или их аналогов, которые машинное обучение и призвано заменить.

Основная часть. Методология сбора данных – представляет собой все методы, связанные со сбором данных фильтраций заведомо недостоверных данных и условия в которых эти данные, собираются. Может оказать существенное влияние на наблюдаемые результаты так, например, социологические исследования, которые проводятся в более комфортных для участников условиях могут показывать большую субъективную удовлетворённость продуктом в сравнение с исследованиями, проводимыми в некомфортных. Для получения качественного набора данных в котором присутствуют субъективные признаки или возможно влияние человеческого фактора необходимо сохранять постоянными условия их сбора или максимально диверсифицировать их. Для некоторых находящихся в открытом доступе наборов данных методологии их сбора описаны крайне кратко либо вовсе не описаны. Возможным решением проблемы является сбор данных в условиях максимально приближенных к тем, в которых система будет применяться.

Недостающие детали в данных либо сомнительная точность этих деталей – при построении сложных моделей машинного обучения большую роль играют детали, например, для модели оценки риска раковых заболеваний очень важным является рацион питания человека, однако не все пользователи такой модели ведут журнал потребляемых продуктов, а те люди, которые его ведут используют разный уровень детализации, например некоторые люди могут не вносить данные о продуктах которыми они перекусывали в виде шоколадных батончиков, печенья и т.п. . Сомнительность предоставленных деталей может быть вызвана личными убеждениями, что выливается в попадание в данные только тех деталей, которые характеризуют человека с положительной стороны или выражают его личные убеждения, например, человек может не вносить в базу данных продукты, съеденные на банкетах, которые он считает явно нездоровыми. Исходя из этого можно сделать вывод о необходимости отбора данных с достоверно точными деталями или снижение глубины детализации используемых данных.

Недостаток данных – в зависимости от предметной области получение достаточного количества данных может быть затруднено или вообще недоступно. Очень часто сочетается с недостаточным количеством деталей в существующих данных. Один из путей решения консолидация данных из нескольких источников или собранных с использованием разных методик который в свою очередь приводит к неравномерности их распределения и разному уровню детализации.

Неравномерность распределения – большинство моделей машинного обучения используют классификацию в качестве конечной или промежуточной целей, исключением являются только простые модели типа линейной регрессии. Так модели классификации могут использовать классификацию на промежуточных этапах (например, классификация животное не животное в модели классификатора кот/собака) а модели регрессии могут использовать классификаторы для предварительного разделения данных на группы и применения разных регрессионных моделей к каждой из групп, например, модели типа дерево или лес. Для таких моделей ещё более важным является относительно равномерное распределение данных для предотвращения переобучения для одной группы данных в сочетании с низкой глубиной формируемых

структур для других. В зависимости от используемого алгоритма обучения в лучшем случае результатом обучения на неравномерных данных будет являться узкоспециализированная система.

Ценность данных и их подготовки – не все данные представляют одинаковую ценность для тренировки моделей машинного обучения. При этом данных которые лишены недостатков, описанных выше может быть недостаточно. Подготовка данных является важным процессом, на котором как правило убираются данные достоверность которых вызывает сомнения, а также могут восстанавливаться недостающие детали. Этот процесс довольно трудоёмок и при этом плохо поддаётся автоматизации необходимость использования, которой обусловлена большим объёмом данных.

Проблема отсутствия значимых корреляций. Не все данные могут объяснять видимый результат, а некоторые проблемы могут быть либо детерминированными, либо стохастическими по природе. При этом при наличии больших объёмов данных модели машинного обучения могут демонстрировать точность превышающую точность генератора случайных чисел на 5% что является достаточным для публикации научного исследования, причём цели, преследуемые авторами, могут быть как личными, в виде повышения количества научных статей и публикаций, так и банальный недостаток квалификации и глубины анализа проблемы, чтобы точно установить является ли найденная корреляция случайной или присутствуют слабые причинно-следственные связи. Также возможна ситуация, когда наблюдаемые корреляции существуют не между учитываемыми факторами и наблюдаемыми результатами, а между неучтенным фактором, который может влиять на учтённый в определённых условиях и наблюдаемым результатом. Сложность формируемых моделей при использовании больших данных также усложняет их анализ для определения значимости зависимостей.

Интерпретируемость результатов – для практического применения должна быть четко доказана эффективность иначе бизнес не готов инвестировать в разработку таких систем. Однако учитывая вышенаписанное научное доказательство эффективности сформированных моделей машинного обучения представляется либо очень трудоёмкой, либо вообще невыполнимой задачей (возможно лишь эмпирическое доказательство). Одна из возникающих при этом проблем в то что в отличие от экспертных систем, которые как минимум систематизируют уже имеющие знания специалистов, машинное обучение не опирается на них и может демонстрировать более низкую эффективность, которая может быть измерена только с учётом определённого набора данных, а не объяснена логически.

Вывод. Несмотря на наличие большого количества проблем и ограничений машинное обучение прочно входит в повседневную жизнь и тенденции показывает только вероятность более плотной интеграции его в большинство продуктов сферы информационных технологий. Несмотря на существенные различия методологии в сравнении с точными науками машинное обучение позволяет решать проблемы, которые недоступны, например, простому статистическому анализу из-за низких коэффициентов корреляции и/или сложности связей.

ЛИТЕРАТУРА

1. Ограничения машинного обучения [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/462365. – Дата доступа:15.09.2020.
2. Проблемы применения машинного обучения для решения реальных задач [Электронный ресурс]. – Режим доступа: http://robocraft.ru/blog/machinelearning/3710.html. – Дата доступа:15.09.2020.
3. 4 способа решить проблемы внедрения Machine Learning в стартапы [Электронный ресурс]. – Режим доступа: https://datastart.ru/blog/read/4-sposoba-reshit-problemy-vnedreniya-machine-learning-v-startapy. – Дата доступа:15.09.2020.
4. Major Challenges for Machine Learning Projects [Электронный ресурс]. – Режим доступа: https://www.topbots.com/major-challenges-machine-learning-projects. – Дата доступа:15.09.2020.
5. Machine learning: 9 challenges [Электронный ресурс]. – Режим доступа: https://www.kaspersky.com/blog/machine-learning-nine-challenges/23553. – Дата доступа:15.09.2020.