УДК 621.3.087.42

С. К. Демченко, А.С. Увайсова, Дао Ань Куан, Фам Лэ Куок Хань, Ф.Ф. Иванов, Нгуен Вьет Данг

Анализ современных методов кластеризации и классификации

За последние 20 лет методы машинного обучения прошли серьезный этап своего развития и сейчас представляют собой весьма функциональный инструмент для анализа данных. В настоящее время они применяются в любой сфере деятельности человека и помогают решать разнообразные задачи – от кредитного скоринга и прогнозирования цен на товары до распознавания номерных знаков автомобилей и синтеза речи. Наиболее актуальными из них являются задачи классификации и кластеризации объектов. В статье приведен обзор и анализ современных методов, способных решать данные задачи.

Ключевые слова: классификация, кластеризация, машинное обучение, искусственный интеллект, нейронные сети, алгоритмы классификации

Об авторах

Демченко Сергей Константинович – аспирант 2-го года обучения кафедры конструирования и производства радиоэлектронных средств РТУ МИРЭА. *E-mail:* skdemchenko@ya.ru. Г. Москва, проспект Вернадского, дом 78.

Увайсова Аида Сайгидовна – аспирант 4-го года обучения кафедры конструирования и производства радиоэлектронных средств РТУ МИРЭА.

Куан Дао Ань – аспирант 3-го года обучения кафедры конструирования и производства радиоэлектронных средств РТУ МИРЭА.

Хань Фам Лэ Куок – аспирант 4-го года обучения кафедры конструирования и производства радиоэлектронных средств РТУ МИРЭА.

Данг Нгуен Вьет – аспирант 2-го года обучения кафедры конструирования и производства радиоэлектронных средств РТУ МИРЭА.

Иванов Федор Федорович – к.т.н., профессор кафедры АСОИУ БУ ВО «Сургутский государственный университет».

Машинное обучение и искусственный интеллект – самое популярное словосочетание в современном научном мире, а также в сфере экономики. Во многом этому способствовало совершенствование технологий и алгоритмов, позволивших с успехом решать самые разнообразные задачи. Сейчас сложно представить область человеческой деятельности, в которой не формируются и не анализируются данные, а направление “Data Science” (наука о данных) стремительно развивается во многих технических вузах и мировых компаниях.

В машинном обучении наиболее развитыми считаются два направления – обучение с учителем и обучение без учителя. Обучение с учителем направлено на решение следующей задачи – существует некоторое множество объектов X, а также некоторое множество всевозможных ответов Y. Между ними существует неизвестная зависимость, т.е. зависимость между объектами и ответами. Также имеется выборка, т.е. множество пар вида «объект – ответ» (такую выборку чаще всего называют обучающей). И на основе этой обучающей выборки требуется восстановить неизвестную зависимость. Таким образом, требуется придумать алгоритм, который для каждого объекта сможет выдать достаточно точный ответ из множества ответов. Здесь точность ответа определяется некоторым функционалом качества, заданным заранее – например, в качестве такого функционала может выступать энтропия.

Задачи обучения с учителем, в зависимости от типа ответа, делятся на задачи классификации и задачи регрессии. В задаче классификации всегда конечное множество ответов. В задаче регрессии ответами являются действительные числа или векторы, состоящие из действительных чисел. Например, задача выявления наличия рака у человека – задача классификации, т.к. множество ответов ограничено (болен/не болен), а задача предсказания цены на квартиру – задача регрессии,

где в качестве ответа выступает действительное число, цена на квартиру.

Отличительной особенностью задачи обучения без учителя является то, что в ней известны только сами объекты, ответов нет, при этом решением задачи является нахождение внутренних взаимосвязей и закономерностей, существующих между объектами. Самой распространенной задачей обучения без учителя является *кластеризация*, для решения которой требуется разбить выборку на некоторые непересекающиеся подмножества объектов таким образом, чтобы в каждом из них находились схожие друг с другом объекты, а объекты разных подмножеств сильно отличались. Такие подмножества называются кластерами, что и дало название данному типу задач.

Классификация

Формальная постановка задачи классификации следующая: пусть X – множество объектов, Y – множество меток класса. Существует некоторая зависимость $\alpha: X \to Y$, а также обучающая выборка $T = \{(x_1, y_1), \dots, (x_n, y_n)\}$. Требуется построить такой алгоритм $p: X \to Y$, который сможет классифицировать произвольный объект $x \in X$.

Основными методами решения данной задачи являются метод k-ближайших соседей, логистическая регрессия, метод опорных векторов [1], дерево решений [2], различные виды бэггингов, среди которых наиболее известным и стабильным считается случайный лес [3], бустинги, среди которых наиболее популярными являются *xgboost* [4], *lightgbm* [5] и *catboost*, а также нейронные сети, которые используются в задачах компьютерного зрения и автоматической обработки текстов.

Проведем сравнительный анализ всех вышеописанных методов на нескольких наборах данных. Данные [9] содержат информацию о пассажирах «Титаника», требуется предсказать, выжил ли пассажир при крушении. Пример входных данных приведен ниже.

Сравнение алгоритмов друг с другом осуществляется по метрике точности, т.е. по доле правильных предсказаний из всей выборки (табл. 2).

В качестве гиперпараметров использовались стандартные гиперпараметры каждой из моделей. Из приведенных выше результатов видно, что самые современные алгоритмы (бустинги) показывают наибольшую точность по сравнению с остальными.

Теперь рассмотрим задачу классификации рукописных изображений от 0 до 9 [10]. Визуализируем входные данные (рисунок).

В табл. 3 приведены результаты исследования.

Таблица 1. **Пример входных данных по одному пассажиру**

Passenger Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
3	1	3	Heikkenen, Miss. Laina	Female	26	0	0	STON/O2. 3101282	7.9250	NaN	S

Таблица 2. **Результаты работы различных алгоритмов по метрике точность**

Алгоритм	Точность
5 ближайших соседей	0.74
Логистическая регрессия	0.75
Дерево решений	0.7
Случайный лес	0.75
LGBM	**0.76**
XGBoost	**0.76**
Catboost	**0.77**
Нейронная сеть	0.74
Метод опорных векторов	0.72

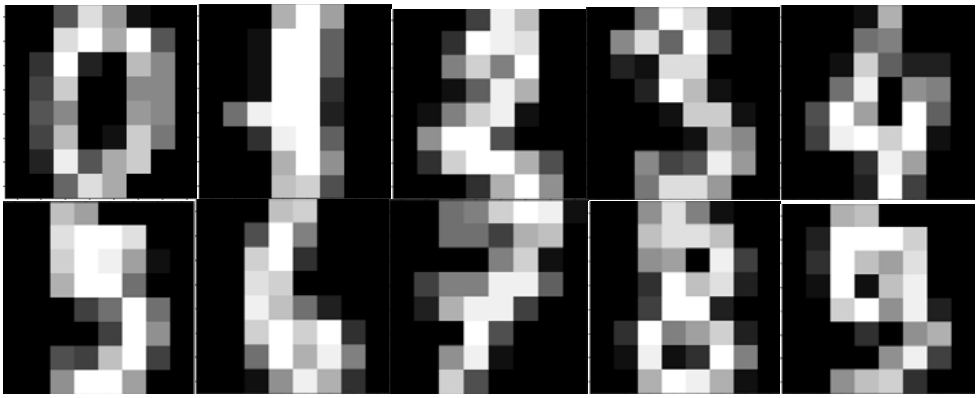

Рисунок. Входные данные для задачи классификации ручных изображений цифр

Таблица 3. **Результаты работы различных алгоритмов по метрике точность**

Алгоритм	Точность
5 ближайших соседей	0.97
Логистическая регрессия	0.96
Дерево решений	0.84
Случайный лес	0.97
LGBM	0.98
XGBoost	0.95
Catboost	0.98
Нейронная сеть	**0.99**
Метод опорных векторов	0.94

Как и ожидалось, при обработке изображений лучшие результаты показали алгоритмы, реализованные в нейронных сетях. Их точность составила 0,99. На втором месте с точностью 0,98 оказались бустинговые алгоритмы, которые качественнее классифицируют объекты, чем более ранние алгоритмы, такие как логистическая регрессия, метод опорных векторов, деревья решений.

Кластеризация

Формальная постановка задачи кластеризации отличается от задачи классификации тем, что множество ответов Y неизвестно. Таким образом, есть множество объектов X, множество кластеров Y и задана некоторая функция расстояния $\rho(x, \hat{x})$ для любых $x, \hat{x} \in X$. Требуется разбить обучающую выборку на кластеры таким образом, чтобы в каждом кластере были близкие объекты по метрике ρ, а в разных кластерах были далекие друг от друга объекты. Как правило, количество кластеров заранее неизвестно, поэтому решается задача определения оптимального количества кластеров при заданном критерии качества кластеризации.

Из постановки задачи следует, что решение задачи кластеризации неоднозначно и может давать разные результаты при разных критериях качества. Помимо этого, функция расстояния и количества кластеров зачастую определятся субъективно экспертом.

Наиболее широко методы кластеризации применяются в задачах сегментации целевой аудитории по интересам, новостей по темам, фильмов по жанру, а также символов по написанию для улучшения распознавания.

Среди наиболее популярных алгоритмов кластеризации можно выделить следующие:

1. k-Means-алгоритм итеративно находит центры кластеров, на вход требует задать количество кластеров.

2. Affinity Propagation [6]. Данный алгоритм не требует задать количество кластеров на вход.

3. Спектральная кластеризация [7]. На вход требуется задать определенное количество кластеров.

4. Агломеративная кластеризация [8]. На вход также требуется задать количество кластеров.

Сравнение представленных алгоритмов по эффективности распознавания рукописных текстов приведено в табл. 4. Для анализа применялась метрика “Adjusted Rand Index”

(ARI), которая вычисляется следующим образом: $ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$, где $RI = \frac{a+b}{n(n-1)}$; n – размер выборки; a – количество пар объектов, находящихся в одном кластере и имеющих одинаковые метки; b – число пар объектов, находящихся в разных кластерах и имеющих разные метки. Значения метрики лежат в интервале $[-1; 1]$. Чем ближе значение к «1», тем выше вероятность того, что алгоритм приблизился к истинному значению.

Таблица 4. **Результаты работы алгоритмов кластеризации по метрике *ARI***

Алгоритм	ARI
k-means алгоритм	0.67
Affinity Propagation	0.18
Спектральная кластеризация	0.76
Агломеративная кластеризация	0.79

Из результатов, представленных в табл. 4, видно, что алгоритм “Affinity Propagation” имеет наихудший результат среди остальных алгоритмов. Связано это с его большей универсальностью и меньшей специфичностью к решаемой задаче, он не требует явного указания на входе количества кластеров. Для всех других алгоритмов входная информация содержала данные о количестве кластеров, равном 10 (10 различных цифр в наборе данных).

Выводы

В статье рассмотрены результаты работы известных алгоритмов классификации и кластеризации, проведен сравнительный анализ данных методов на примере двух наборов данных – о выживших при крушении на «Титанике», а также на данных о рукописных записях цифр. Из результатов видно, что в задачах классификации, связанных с табличными данными, наиболее предпочтительным вариантом выбора алгоритмов являются бустинговые алгоритмы, такие как *xgboost*, *lightgbm* и *catboost*. А в задачах, связанных с классификацией объектов на картинках, наиболее предпочтительным вариантом является использование нейронных сетей.

Применительно к кластеризации рассмотрена работа четырех различных алгоритмов на примере задачи кластеризации изображений рукописных цифр. Алгоритм “Affinity Propagation”, не требующий на вход количество кластеров, заметно уступил в качестве по метрике *ARI* своим аналогам.

Библиографический список

1. Бишоп К.М. Распознавание образов и машинное обучение. М.: Вильямс, 2020. 960 с.

2. Andrew Y.Ng, Michael I. Jordan, Yair Weiss. On Spectral Clustering: Analysis and an algorithm // NIPS'01 : Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. 2001. P. 849–856.

3. Brendan J., Delbert Dueck. Clustering by Passing Messages Between Data Points // SCIENCE. 2007. V. 315. P. 972–976.

4. Gass S.I., John F. Magee, Assad A. Profiles in Operations Research // International Series in Operations Research & Management Science. 2011. V. 147. – https://doi.org/10.1007/978-1-4419-6281-2_33.

5. Leo Breiman. Random Forests // Machine Learning volume. 2001. 45. P. 5–32.

6. Marcel R. Ackermann Analysis of Agglomerative Clustering // Algorithmica. 2014. 69. P. 184–215.

7. Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System // KDD '16: Proceedings of the 22nd ACM SIGKDD // International Conference on Knowledge Discovery and Data Mining. 2016. P. 785–794.

Поступила в редакцию
27.11.2020