УДК 316.6

РОГОВ Д. Е., студент учебной группы 3272
Военная академия связи имени С.М. Будённого
г. Санкт-Петербург

ЗАДАЧА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

Одним из ключевых компонентов современных систем управления в организациях является аналитическая работа. Она представляет собой исследовательскую деятельность, нацеленную на выявление взаимосвязей между происходящими событиями, трендами и закономерностями в соответствующей сфере [1], необходимых для обоснования принятых управленческих решений и оценки эффективности функционирования используемой модели управления [2].

Машинное обучение сегодня часто применяется для повышения скорости работы и эффективности [3, 4]. Однако в процессе использования машинного обучения регулярно требуется маркировка данных, то есть соотнесение определенных фактов с их значениями. Например, мы можем разработать модель для классификации статей по определенной тематике [5].

Для начала из исходных данных нужно убрать шумы, а именно — союзы, вводные слова и предлоги, потому что эти элементы являются нетипичными и могут сбивать с толку модель [6, 7]. К тому же для основной задачи классификации они не несут особой пользы. Так как вокруг шумов плотность распределения мала, в этом случае используется эвристика «взвешивание по плотности», где приоритет отдается объектам с более высокой плотностью распределения. Продемонстрируем этот процесс формирования выборки на примере.

Для каждого объекта, не являющегося размеченным, рассчитаем дискретное распределение x по классам $P(y|x)$. Для этого оценим априорную вероятность каждого класса $P(y)$, а также плотность распределения x при известном классе y.

Априорную вероятность класса $P(y)$ можно оценить, если известно количество запросов в поисковых системах, а также количество книг, статей, учебников на соответствующую тему («геометрия», «математика», «искусственный интеллект»). Но в таком случае для нахождения каждого значения $p(x|y)$ для каждого слова необходимо совершить анализ всего объема вышеперечисленных текстовых материалов, что не представляется возможным за адекватное время.

Следовательно, придётся пойти другим путём: для того, чтобы определить априорную вероятность класса $P(y)$, мы возьмем за эталон по одной статье из предметных областей [8] «Геометрия», «Математика» и «Искусственный интеллект».

В нашем случае априорные вероятности будут такие:

$$P_{y1} = \frac{2132}{7511} = 0{,}284;\ P_{y2} = \frac{1848}{7511} = 0{,}246; P_{y3} = \frac{3531}{7511} = 0{,}47;$$

Для примера выберем следующие слова (рисунок 1).

Слово	Количество	Статья по геометрии	Статья по алгебре	Статья по ИИ
новост	122	0	0	0
находят	29	10	20	15
люди	29	2	0	0
контент	23	3	4	5
информация	30	0	0	0
социальн	32	0	0	0
эффект	18	8	0	5
избегающие	27	3	3	3
образом	15	4	0	1
людей	13	2	2	2
потребление	13	4	1	4
отказ	12	0	0	0
сетях	12	0	0	1
случае	12	3	4	4
позволяет	11	5	5	5
медиа	10	0	0	1
например	10	7	3	1
событиях	9	7	5	9
более	9	7	5	6
человека	9	0	0	5
потребления	9	3	2	1

Рисунок 1. Список слов

Исключим из выборки цифры, предлоги, а также те слова, которых нет ни в одном из классов [9], после чего найдем апостериорные вероятности для каждого класса [10].

Py 1	P(x|y) 1	Py 1 * P(x|y) 1	P(y1|x)
0,284	0	0	0
	0,004690432	0,001332083	0,153846154
	0,000938086	0,000266417	0,030769231
	0,001407129	0,000399625	0,046153846
	0	0	0
	0	0	0
	0,003752345	0,001065666	0,123076923
	0,001407129	0,000399625	0,046153846
	0,001876173	0,000532833	0,061538462
	0,000938086	0,000266417	0,030769231
	0,001876173	0,000532833	0,061538462
	0	0	0
	0	0	0
	0,001407129	0,000399625	0,046153846
	0,002345216	0,000666041	0,076923077
	0	0	0
	0,001876173	0,000532833	0,061538462
	0,003283302	0,000932458	0,107692308
	0,003283302	0,000932458	0,107692308
	0	0	0
	0,001407129	0,000399625	0,046153846

Рисунок 2. Поиск апостериорных вероятностей для 1 класса (Геометрия)

Метриками точности в этом случае выступают несколько величин:

1. Принцип наименьшей достоверности (3):

$$u_i = \arg\min_{u \in U} p_1(u) \qquad (3)$$

То есть чем меньше значение $p_1(x)$, тем больше распределение вероятностей похоже на равномерное. Это значит, что объект должен попасть в выборку, направляемую для оценки [11].

2. Принцип наименьшей разности отступов (4):

$$u_i = \arg\min_{u \in U}(p_1(u) - p_2(u)) \quad (4)$$

То есть чем меньше разница между $p_1(x)$ и $p_2(x)$, тем ближе объект находится к границе классов, поэтому объект должен попасть в выборку, направляемую для оценки.

3. Принцип максимума энтропии (5):

$$u_i = \arg\min_{x \in U} \sum_m p_m(u) \ln p_m(u) \quad (5)$$

Оценка проводится по энтропии распределения вероятностей. Она будет минимальна при равномерном распределении, т.е. $p_1(x) \approx p_2(x) \approx p_3(x)$.

Используя эти принципы, получим следующие данные (см. рисунок 3).

	P(y1|x)	P(y2|x)	P(y3|x)	Принцип 1	Принцип 2	Принцип 3	Класс
новост	0	0	0	-	-	-	-
находят	0,175438596	**0,363636364**	0,214285714	0,3636364	0,1493506	-1,0032952	2
люди	**0,035087719**	0	0	0,0350877	-		1
контент	0,052631579	**0,072727273**	0,071428571	0,0727273	0,0012987	-0,5340956	-
информация	0	0	0	-	-	-	-
социальн	0	0	0	-	-	-	-
эффект	**0,140350877**	0	0,071428571	0,1403509	0,0689223	-	-
избегающие	0,052631579	**0,054545455**	0,042857143	0,0545455	0,0019139	-0,448623	-
образом	**0,070175439**	0	0,014285714	0,0701754	0,0558897	-	-
людей	0,035087719	**0,036363636**	0,028571429	0,0363636	0,0012759	-0,3396377	-
потребление	**0,070175439**	0,018181818	0,057142857	0,0701754	0,0130326	-0,422854	-
отказ	0	0	0	-	-	-	-
сетях	0	0	**0,014285714**	0,0142857	-	-	3
случае	0,052631579	**0,072727273**	0,057142857	0,0727273	0,0155844	-0,5091458	-
позволяет	0,087719298	**0,090909091**	0,071428571	0,0909091	0,0031898	-0,6199694	-
медиа	0	0	**0,014285714**	0,0142857	-	-	3
например	0,01754386	0,036363636	**0,085714286**	0,0857143	0,0493506	-0,4020239	-
событиях	0,035087719	**0,127272727**	0,085714286	0,1272727	0,0415584	-0,5904808	-
более	**0,122807018**	0,090909091	0,085714286	0,122807	0,0318979	-0,6861115	-
человека	0	0	**0,071428571**	0,0714286	-	-	3
потребления	**0,052631579**	0,036363636	0,014285714	0,0526316	0,0162679	-0,3361791	-

Рисунок 3. Использование метрик точности

Без сомнений можно сказать, что слова «сетях», «медиа», «человека» принадлежат к 3 классу (Искусственный интеллект), слово «находят» - ко 2 классу (Математика), а слово «люди» – к 1 классу (Геометрия). Что касается других слов, то можно сказать, что довольно тяжело определить, к какому классу их отнести: исходя из принципа 4, разница между апостериорными вероятностями классов крайне мала. Следовательно, эти слова попадают в выборку по неуверенности. Также в данную выборку попадают те слова, которые ни разу не повторились в наших статьях.

Список литературы:

1. *Томас Кормен, Чарльз Лейзерсон.* Алгоритмы: построение и анализ, 3-е издание – М.: ООО И.Д. Вильямс. 2013. – 1328 с.

2. *Пылов, П. А.* Человек управляет искусственным интеллектом или искусственный интеллект управляет человеком? / П. А. Пылов, И. В. Кудаева // Россия молодая : Сборник материалов XIII Всероссийской научно-практической конференции с международным участием, Кемерово, 20–23 апреля 2021 года / Редколлегия: К.С. Костиков (отв. ред.) [и др.]. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2021. – С. 94703.1-94703.5. – EDN LSBOPZ.
3. L. Graesser, W. L. Keng. Foundations of Deep Reinforcement Learning: Theory and Practice in Python. Addison-Wesley Professional, 2019.
4. *Пылов, П. А.* Идентификация рукописных чисел в цифровом формате средствами искусственного интеллекта / П. А. Пылов, А. В. Протодьяконов // Инновации в информационных технологиях, машиностроении и автотранспорте (ИИТМА-2020) : сборник материалов IV Международной научно-практической конференции с онлайн-участием, Кемерово, 07–10 декабря 2020 года. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2020. – С. 189-191. – EDN FGKEFO.
5. I. Isaev, S. Dolenko. Group Determination of Parameters and Training with Noise Addition: Joint Application to Improve the Resilience of the Neural Network Solution of a Model Inverse Problem to Noise in Data. Advances in Intelligent Systems and Computing, 2019, V.848, pp. 138-144. Springer, Cham. DOI: 10.1007/978-3-319-99316-4_18 (дата обращения: 21.08.2023).
6. *Пылов, П. А.* Единичная оценка в сравнении с упаковочными алгоритмами: смещение смещения дисперсии / П. А. Пылов, А. В. Протодьяконов // Инновации в информационных технологиях, машиностроении и автотранспорте (ИИТМА-2020) : сборник материалов IV Международной научно-практической конференции с онлайн-участием, Кемерово, 07–10 декабря 2020 года. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2020. – С. 192-194. – EDN YEZZUM.
7. Свидетельство о государственной регистрации программы для ЭВМ № 2023688086 Российская Федерация. Программа латентной модели согласованности : № 2023687442 : заявл. 12.12.2023 : опубл. 20.12.2023 / Р. В. Майтак. – EDN RFFALN.