УДК 004.9

ЗАДАЧА МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

Рогов Д.Е., студент гр. ИИм-221, 2 курс
Кузбасский государственный технический университет имени Т.Ф. Горбачева, г. Кемерово

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 применяется для повышения скорости работы и эффективности [3, 4]. Однако в процессе использования машинного обучения 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]. Для основной задачи классификации эти не несут особой пользы. Так как вокруг шумов плотность распределения мала, 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 $P(y)$, плотность распределения x при известном классе y.

Априорную вероятность класса $P(y)$ можно оценить, если известно количество запросов в поисковых системах, 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, для того чтобы определить априорную вероятность класса $P(y)$ возьмем за эталон по одной статье из каждой предметной области [8]:

Тогда, 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 выступают несколько величин:

1. Принцип наименьшей достоверности (3):

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

Таким образом, чем меньше значение $p_1(x)$, тем больше распределение вероятностей похоже на равномерное, а 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. $p_1(x) \approx p_2(x) \approx p_3(x)$.

Используем принципы и получим следующие данные (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» - к 2 классу (Математика). И слово «люди» – к 1 классу (Геометрия). Для других же слов, тяжело определить к какому классу их отнести, так как, 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. Свидетельство о государственной регистрации программы для ЭВМ № 2023680124 Российская Федерация. BrainPower : № 2023669010 : заявл. 16.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QXBJIM.
2. Математические и программные методы построения моделей глубокого обучения : Учебное пособие / А. В. Протодьяконов и др. –

Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 176 с. – ISBN 978-5-9729-1484-5. – EDN PZLUAH.

3. Свидетельство о государственной регистрации программы для ЭВМ № 2023680335 Российская Федерация. Maitak Intelligence Natural Language Processing Module : № 2023669704 : заявл. 27.09.2023 : опубл. 28.09.2023 / Р. В. Майтак.
4. Методы восстановления непараметрической регрессии в условиях несбалансированных данных / А. Д. Салычева и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 192 с. – ISBN 978-5-9729-1856-0. – EDN AAJATW.
5. Свидетельство о государственной регистрации программы для ЭВМ № 2023684619 Российская Федерация. Efficient Network: № 2023684038: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
6. Свидетельство о государственной регистрации программы для ЭВМ № 2023680070 Российская Федерация. Модернизированная модель DBSCAN для определения скрытых взаимосвязей : № 2023668841 : заявл. 13.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN KQUUKF.
7. Асимптотический анализ поведения прикладных моделей машинного обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.
8. Свидетельство о государственной регистрации программы для ЭВМ № 2023684621 Российская Федерация. Destructed Deep Random Forest: № 2023684050: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
9. Свидетельство о государственной регистрации программы для ЭВМ № 2023684622 Российская Федерация. Mask Made AI: № 2023684042: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
10. Свидетельство о государственной регистрации программы для ЭВМ № 2023680103 Российская Федерация. Cognitive Solution : № 2023669189 : заявл. 19.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QEMFJA.
11. Свидетельство о государственной регистрации программы для ЭВМ № 2023684624 Российская Федерация. Программа автоматического распознавания лиц в видеопотоке: № 2023684236: заявл. 15.11.2023: опубл. 16.11.2023 / П. А. Пылов.