Нижегородский государственный университет им. Н.И.Лобачевского

Факультет Вычислительной математики и кибернетики

Автоматизация построения тематических классификаторов с использованием алгоритмов машинного обучения

Борисюк Ф.В., Дружков П.Н.,
Половинкин А.Н.

Работа выполнена при поддержке федеральной целевой программы «Научные и научно-педагогические кадры инновационной России», госконтракт 02.740.11.5131

Задача обучения с учителем

$\mathcal{X}$ – множество допустимых входов

$\mathcal{Y}$ – множество допустимых выходов

$$\mathcal{X} = \mathbb{R}^{p_1} \times A_1 \times A_2 \times \ldots \times A_{p_2},$$

$$\mathcal{Y} = \mathbb{R} \text{ или } B$$

где $A_1, A_2, \ldots, A_{p_2}, B$ – конечные множества.

Каждому элементу из $\mathcal{X}$ поставлен в соответствие элемент из $\mathcal{Y}$.

Задача обучения с учителем

$\{(x_i, y_i) \mid x_i \in \mathcal{X},\ y_i \in \mathcal{Y},\ i = \overline{1, N}\}$ – обучающая выборка

Требуется восстановить зависимость между входом и выходом

Некоторые подходы к решению

- Машина опорных векторов (SVM)
  - Строит гиперплоскость, наилучшим образом разделяющую точки разных классов
  - При решении многоклассовых задач возможны подходы «один против всех» (one-against-all) и «каждый против каждого» (one-against-one)
- Бэггинг
  - Строится множество независимых классификаторов
  - Решение принимается голосованием
- Бустинг
  - Итеративно строит множество классификаторов так, что каждый следующий исправляет ошибки предыдущих

Задача классификации научных публикаций

- Каждому документу ставится в соответствие вектор булевых признаков встречаемости слов (присутствует слово, или нет) из заранее построенного словаря
- При этом игнорируются стоп-слова (предлоги, союзы, частицы, местоимения, вводные слова)
- Все слова приводятся к основе с помощью алгоритма Портера

База публикаций журнала «Вестник ННГУ»

- Включает в себя статьи по 18 категориям («Радиофизика», «Химия», «Биология», «Механика», «Математика», «Филология», …)
- Общее число документов в базе – 2655

Описание эксперимента

- Вся имеющаяся база была разделена случайным образом на обучающую и тестовую выборки – 1855 и 800 статей, соответственно
- Объем используемого словаря (размерность пространства признаков) – 181822
- Использовались следующие модели
  - Линейная машина опорных векторов (SVM)
  - Случайные деревья (RT)
  - Полностью случайные деревья (ERT)
  - Градиентный бустинг деревьев решений (GBT)

Используемая инфраструктура

- Компилятор: Microsoft C/C++ Compiler Version 15.00.30729 (x64).
- Процессор: 2 двухъядерных процессора Intel Xeon 5150 (2.66 GHz).
- Память: 4 GB.

Экспериментальные результаты (точность)

	RT	ERT	SVM	GBT
1	0.44	0.59	0.95	0.97
2	1	1	0.8	0.87
3	—	—	1	0.88
4	—	—	—	0.33
5	—	—	0.75	0.7
6	—	—	—	1
7	—	—	—	0.8
8	—	—	0.63	0.82
9	1	0.92	0.73	0.81

	RT	ERT	SVM	GBT
10	1	1	0.89	0.96
11	—	—	0.5	0.5
12	0.87	0.8	0.87	0.85
13	0.58	0.65	0.77	0.71
14	—	0.63	0.73	0.93
15	—	—	1	0.78
16	0.69	0.69	0.73	0.84
17	1	1	0.96	0.9
18	0.45	0.62	0.94	0.88

Экспериментальные результаты (полнота)

	RT	ERT	SVM	GBT
1	0.97	0.99	0.97	0.94
2	0.04	0.41	0.8	0.87
3	0	0	0.43	1
4	0	0	0	0.25
5	0	0	0.58	0.52
6	0	0	0	0.38
7	0	0	0	0.67
8	0	0	0.79	0.75
9	0.05	0.26	0.69	0.69

	RT	ERT	SVM	GBT
10	0.69	0.69	0.78	0.81
11	0	0	0.17	0.33
12	0.67	0.78	0.85	0.87
13	0.79	0.81	0.85	0.84
14	0	0.68	1	0.93
15	0	0	0.38	0.54
16	0.61	0.82	0.86	0.87
17	0.33	0.85	0.96	0.93
18	1	0.99	0.95	0.93

Экспериментальные результаты (F-мера)

	RT	ERT	SVM	GBT
1	0.61	0.74	0.96	0.96
2	0.07	0.58	0.8	0.87
3	—	—	0.6	0.93
4	—	—	—	0.29
5	—	—	0.65	0.59
6	—	—	—	0.55
7	—	—	—	0.73
8	—	—	0.7	0.78
9	0.09	0.41	0.71	0.74

	RT	ERT	SVM	GBT
10	0.81	0.81	0.83	0.88
11	—	—	0.25	0.4
12	0.75	0.79	0.86	0.86
13	0.67	0.72	0.81	0.77
14	—	0.65	0.85	0.93
15	—	—	0.56	0.64
16	0.65	0.75	0.79	0.86
17	0.49	0.92	0.96	0.91
18	0.62	0.76	0.95	0.91

Экспериментальные результаты (время работы)

Алгоритм	Ошибка на тестовой выборке, %	Время обучения, сек. (ч.)	Среднее время классификации одного образца, сек.
RT	44.75	150381 (41.8)	0.55
ERT	30.88	312446 (86.8)	0.57
SVM	16.75	782 (0.2)	0.47
GBT	15.5	149114 (41.4)	0.29

Спасибо за внимание