Автоматизация построения тематических классификаторов с использованием алгоритмов машинного обучения♣

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

Нижегородский государственный университет им. Н.И. Лобачевского
fedorvb@gmail.com, druzhkov.paul@gmail.com, alexey.polovinkin@gmail.com

Аннотация

В работе рассматривается задача классификации научных публикаций с использованием современного аппарата машинного обучения. Проведено экспериментальное сравнение различных широко распространенных алгоритмов обучения с учителем на базе статей Вестника Нижегородского государственного университета.

1. Введение

В настоящее время в различных хранилищах знаний (электронных и традиционных) накоплены огромные массивы информации. При этом по причине ее больших объемов, а также слабой структурированности и отсутствия представления в электронном виде, зачастую поиск актуальной и полной информации по конкретной теме является достаточно сложным. В этом случае большая часть накопленных информационных ресурсов может оказаться бесполезной из-за их необозримости.

Можно отметить, что зачастую решение конкретной научной задачи требует высоких трудозатрат по поиску и анализу информации по теме. Поэтому, в связи с вышесказанным, возникает задача эффективного структурирования, хранения, обработки и поиска в информационных массивах.

Одним из традиционных подходов к решению данной задачи является классификационный поиск, к которому относится поиск с использованием различных тематических классификаторов, рубрикаторов, электронных каталогов, что позволяет искать (автоматически или вручную) документы в небольшом подмножестве исходной коллекции документов по интересующей пользователя тематике. Рубрикатор (электронный каталог) обычно представляет собой множество рубрик, объединенных в иерархию. К каждой рубрике приписываются соответствующие ее тематике документы. Несмотря на то, что традиционные каталоги (классификаторы) имеют фиксированную структуру и зачастую не поддерживают высоких темпов развития различных областей знаний в науке и технике, а также требуют высоких временных затрат на адаптацию классификаторов, и классификацию по ним документов, данный тип информационного поиска остается широко распространенным. В работе рассматривается один из возможных подходов к построению электронных каталогов, основанный на автоматической классификации научных публикаций по тематическим категориям, который заключается в сведении данной проблемы к задаче обучения с учителем. Рассматриваемые в статье исследования проводились в рамках актуальной практической проблемы автоматизации обработки и хранения статей Вестника ННГУ в виде тематического каталога и использования тематической информации при поиске по научным статьям. Тематическую информацию автоматически построенного классификатора планируется использовать в качестве механизма поддержки добавления новых статей в научную коллекцию существующего каталога [9], а также использовать ее для улучшения качества поиска по ключевым словам, реализованного в каталоге.

2. Задача обучения с учителем

В рамках этой задачи дано некоторое множество *объектов* X. Каждому объекту $x \in X$ поставлена в соответствие величина y, называемая *выходом*, или *ответом*, и принадлежащая множеству допустимых ответов Y. Упорядоченная пара «объект-ответ» (x, y), где $x \in X$, $y \in Y$ называется *прецедентом*. Требуется восстановить зависимость между входом и выходом, основываясь на данных о конечном наборе прецедентов, называемом *обучающей выборкой*: $\{(x_i, y_i) \mid x_i \in X, y_i \in Y, i = \overline{1, N}\}$. Другими словами, задача состоит в построении *модели* (функции) f, которая, получив на вход x, предсказала бы значение ответа y. Процесс нахождения f называется *обучением*, или *настройкой* модели. Если Y конечно, говорят о *задаче классификации*, если $Y = \mathbf{R}$ – *задаче восстановления регрессии* [1].

Труды 13[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.

Таблица 1. Распределение базы документов по рубрикам.

Рубрика	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Обучающая выборка	188	122	17	6	70	16	11	55	96	72	13	201	248	103	29	165	105	338
Тестовая выборка	71	54	7	4	31	8	6	24	42	32	6	87	108	44	13	71	46	146

Основным требованием, предъявляемым к решению, является высокая обобщающая способность, т. е. обученная модель должна выдавать достаточно точные предсказания на новых (не входящих в обучающую выборку) прецедентах. Таким образом, оптимальное решение задачи индуктивного обучения должно удовлетворять условию:

$$f^* = \arg\min_{f \in K} M_{y,x} L(y, f(x)),$$

где $L(y, f(x))$ – неотрицательная функция потерь (штрафа), K – некоторое множество моделей. Однако данный критерий не применим в случае конечного набора известных данных, и обычно заменяется на условие:

$$f^* = \arg\min_{F \in K} \sum_{i=1}^{N} L(y_i, F(x_i)),$$

где прецеденты (x_i, y_i), $i = \overline{1, N}$, составляют обучающую выборку.

3. Обзор алгоритмов обучения с учителем

Исторически первым алгоритмом такого рода считается персептрон Розенблата. В дальнейшем было изобретено множество оригинальных методов [4]: среди наиболее известных следует отметить нейронные сети, деревья решений, машину опорных векторов [2, 8] и метод ближайших соседей, а также различные виды восстановления регрессии.

Огромным прорывом в данной области явилась идея об использовании комбинирования моделей для улучшения обобщающей способности алгоритма. Можно выделить две основные конкурирующие идеи данного подхода – бэггинг (*bagging* от *Bootstrap Aggregating*) и бустинг (*boosting*). Первая из них состоит в построении множества независимых (между собой) моделей с дальнейшим принятием решения путем голосования в случае задачи классификации и усреднения в случае регрессии. Данный подход реализован в алгоритме случайных деревьев (*random trees* или *random forest*). Основной сложностью применения этой идеи является обеспечение независимости построенных моделей. Бустинг, в противоположность бэггингу, обучает каждую следующую модель с использованием данных об ошибках предыдущих моделей [4].

4. Сведение задачи классификации текста к задаче обучения с учителем

Основная идея подхода сведения задачи классификации текста к задаче обучения с учителем заключается в построении признакового описания каждого рассматриваемого документа, которое представляет собой вектор булевых признаков встречаемости слов (присутствует слово, или нет) в документе из заранее построенного словаря. При построении вектора описания документа предварительно из списка слов документа исключаются стоп-слова (предлоги, союзы, частицы, местоимения, вводные слова). Каждое слово документа приводится к основе слова (основа определяется как часть слова без окончания и формообразующих суффиксов), таким образом, несколько вариаций слова отображается в одну компоненту вектора описания документа. Построение основы слова в рассматриваемом документе производилось с использование алгоритма Портера [7], реализованного для русского языка.

5. Описание вычислительного эксперимента

В работе был проведен вычислительный эксперимент на базе публикаций журнала «Вестник ННГУ». Данная база включает в себя статьи по следующим тематикам: «Инновации в образовании», «Радиофизика», «Химия», «Биология», «Механика», «Математика», «Математическое моделирование», «Оптимальное управление», «Информационные технологии», «Филология», «Социология. Психология. Философия», «Экономика» и ряд других дисциплин (всего рассматривается 18 категорий). Общее число документов в базе – 2655, которое было разбито случайным образом на обучающую и тестовую выборки (с числом объектов, равным 1855 и 800, соответственно), количество документов каждой категории, попавших в обучающую и тестовую выборки, указано в табл. 1. Для построения классификаторов и предсказания использовалась библиотека с открытым кодом OpenCV [6], реализация алгоритма Gradient Boosting Trees в которой принадлежит авторам работы [9]. Все вычислительные эксперименты проводились с использованием следующей инфраструктуры:

- Компилятор: Microsoft C/C++ Compiler Version 15.00.30729 (x64).

- Процессор: 2 двухъядерных процессора Intel Xeon 5150 (2.66 GHz).
- Память: 4 GB.

Рассматривались следующие классификаторы:

1) Random Trees (RT) [1] (глубина деревьев – 10, количество деревьев – 10000).
2) Extremely Randomized Trees (ERT) [5] (глубина деревьев – 15, количество деревьев – 10000).
3) Gradient Boosting Trees (GBT) [3] (глубина деревьев – 5, количество деревьев – 1000, shrinkage – 0.005, subsampling отключен).
4) Support Vector Machine (SVM) [2, 8] (линейное ядро, параметр, определяющий штраф за нарушение ограничений, $C = 0.1$).

С целью оценки качества классификации для каждой модели были вычислены значения точности, полноты и F-меры, приведенные в табл. 2, 3 и 4 соответственно. В некоторых случаях значения точности, а, следовательно, и F-меры оказались не определены, т.к. ни один документ из тестовой выборки не был отнесен системой к соответствующему классу. Стоит отметить, что лишь модель градиентного бустинга деревьев решений позволила произвести категоризацию документов тестовой выборки на все 18 категорий, несмотря на малое количество документов данных классов в обучающей выборке. При этом качество классификации модели GBT сопоставимо с качеством SVM-классификатора. Применение случайных деревьев (RT) и экстремально случайных деревьев (ERT) дает худшие результаты в обоих аспектах. В табл. 5 приведена общая по всем классам тестовая ошибка (процент неверно классифицированных документов от общего числа объектов тестовой выборки), а также время тренировки модели, и среднее время категоризации одного документа. Из полученных результатов видно, что время обучения моделей, представляющих собой ансамбли деревьев решений (RT, ERT, GBT), значительно превышает время настройки SVM, однако, GBT дает наилучшие показатели по времени классификации.

Таблица 2. Значение точности для каждого класса, полученное на тестовой выборке.

Метод	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
RT	.44	1	—	—	—	—	—	—	1	1	—	.87	.58	—	—	.69	1	.45
ERT	.59	1	—	—	—	—	—	—	.92	1	—	.80	.65	.63	—	.69	1	.62
SVM	.95	.80	1	—	.75	—	—	.63	.73	.89	.50	.87	.77	.73	1	.73	.96	.94
GBT	.97	.87	.88	.33	.70	1	.80	.82	.81	.96	.50	.85	.71	.93	.78	.84	.90	.88

Таблица 3. Значение полноты для каждого класса, полученное на тестовой выборке.

Метод	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
RT	.97	.04	0	0	0	0	0	0	.05	.69	0	.67	.79	0	0	.61	.33	1
ERT	.99	.41	0	0	0	0	0	0	.26	.69	0	.78	.81	.68	0	.82	.85	.99
SVM	.97	.80	.43	0	.58	0	0	.79	.69	.78	.17	.85	.85	1	.38	.86	.96	.95
GBT	.94	.87	1	.25	.52	.38	.67	.75	.69	.81	.33	.87	.84	.93	.54	.87	.93	.93

Таблица 4. Значение F-меры для каждого класса, полученное на тестовой выборке.

Метод	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
RT	.61	.07	—	—	—	—	—	—	.09	.81	—	.75	.67	—	—	.65	.49	.62
ERT	.74	.58	—	—	—	—	—	—	.41	.81	—	.79	.72	.65	—	.75	.92	.76
SVM	.96	.80	.60	—	.65	—	—	.70	.71	.83	.25	.86	.81	.85	.56	.79	.96	.95
GBT	.96	.87	.93	.29	.59	.55	.73	.78	.74	.88	.40	.86	.77	.93	.64	.86	.91	.91

Таблица 5. Тестовая ошибка и временные показатели.

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

Таким образом, алгоритм градиентного бустинга деревьев решений (GBT) при решении задачи текстовой классификации показал себя серьезным конкурентом машине опорных векторов (SVM), традиционно применяемой в задачах такого рода, позволив добиться увеличения качества классификации, особенно для малочисленных категорий документов.

6. Заключение

В работе рассмотрен метод автоматического построения тематических классификаторов (каталогов) путем сведения данной проблемы к задаче классификации текстов с использованием современного аппарата машинного обучения.

Проведено комплексное экспериментальное сравнение различных широко распространенных алгоритмов обучения с учителем на базе статей Вестника Нижегородского государственного университета для решения задачи классификации текстовых документов.

С использованием разработанной авторами статьи реализации перспективного метода машинного обучения Gradient Boosting Trees [10] получено лучшее по качеству решение поставленной задачи текстовой классификации по сравнению с широко известными методами классификации, что позволяет говорить о научной и практической важности полученных результатов.

Полученные результаты демонстрируют возможность использования рассматриваемого метода для автоматического построения тематических каталогов в электронных библиотеках.

Литература

[1] L. Breiman. Random Forests // Machine Learning. – 2001. – Vol.45, No.1. – P. 5–32.

[2] C. Cortes and V. Vapnik. Support-Vector Networks // Machine Learning. – 1995. – Vol.20, No.3. – P. 273–297.

[3] J.H. Friedman. Greedy function approximation: a gradient boosting machine // The Annals of Statistics. – 2001. – Vol.29, No.5. – P. 1189–1232.

[4] T. Hastie, R Tibshirani, J. Friedman. The Elements of Statistical Learning – Springer – 2001.

[5] P. Geurts, D. Ernst, L. Wehenkel. Extremely Randomized Trees // Machine Learning. – 2006. – Vol.36, No.1. – P. 3–42.

[6] OpenCV Wiki. - http://opencv.willowgarage.com/wiki/

[7] M.F. Porter. An Algorithm for Suffix Stripping // Program. – 1980. – Vol.14, No.3. – P. 130–137.

[8] Вапник В. Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979.

[9] Вестник ННГУ. http://www.unn.ru/e-library/vestnik.html

[10] Дружков П. Н., Золотых Н. Ю., Половинкин А. Н. Программная реализация алгоритма градиентного бустинга деревьев решений // Вестник Нижегородского государственного университета им. Н. И. Лобачевского. – 2011. – № 1. – С. 193–200.

Automated Construction of Subject Classifiers Using Machine Learning Algorithms

© F.V. Borisyuk, P.N. Druzhkov, A.N. Polovinkin

Classification problem of scientific publications applying modern machine learning approaches has been examined. Experimental comparison results of wide spread supervised learning algorithms on the Vestnik UNN journal papers dataset is given.

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