УДК 004.89, 004.048

Е. Н. Каруна, П. В. Соколов
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Нейросетевой классификатор текстовой информации

Рассматриваются теоретические основы машинной классификации текстовой информации. В последнее время отмечается рост интереса к данной тематике. В статье выделены основные этапы и главные сложности решения задач данного направления, представлены данные, полученные в результате работы простого алгоритма классификации текстовой информации. Обсуждены предварительная фильтрация текстов, формирование векторов признаков, структура и принципы обучения нейронной сети. Для оценки результатов используется F-мера. Проведено сравнение результатов трех коллекций текстов для различных параметров предварительного фильтра, числа нейронов в скрытом слое и времени обучения сети. Предложенная модель классификатора позволяет решить задачу классификации с точностью более 80 %, при этом решающий вклад в точность классификации вносит качество обучающих данных. Сделаны выводы о качестве полученных результатов и представлены варианты дальнейших исследований по данной теме.

Классификация, машинное обучение, тематический анализ, нейронная сеть, стемминг

С появлением открытого доступа к большим объемам текстовой информации, доступной в электронном виде, растет потребность в классификации этой информации по разным категориям и в выявлении различных закономерностей, присущих некоторой группе текстовых данных из определенной выборки. Несмотря на резкий рост интереса к задачам подобного рода в последнее время, разработка новых высокоэффективных методов и средств классификации высоко актуальна.

К наиболее важным направлениям в обработке естественного языка относится тематический анализ текстовой информации. Тематический анализ позволяет разделять текстовые данные на категории, например для быстрой классификации текстовой базы, для упрощения работы человека с этими текстами, для систем информационного поиска, а также для некоторых задач диалоговых систем. В данной статье описываются актуальные проблемы в области анализа текстовой информации и делается обзор существующих подходов и разработок в данной области.

Общая структура алгоритма тематического анализа представлена на рис. 1. В [1] подробно рассматриваются алгоритмы машинного обучения для задач естественной обработки языка. К основным элементам структуры относятся: текстовая база данных – исходный корпус текстов

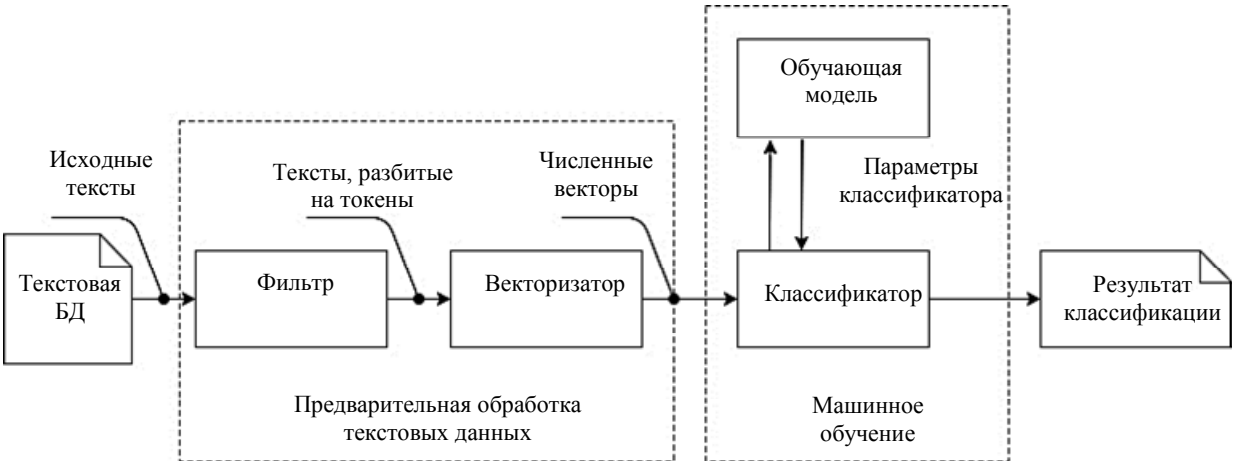

Рис. 1

для обучения и тестирования классификатора, фильтр – алгоритм удаления небуквенных символов, общеупотребительных слов, приведение слов к единому регистру, приведение к основе слова. Векторизатор – алгоритм формирования численного массива с набором признаков, характерных для каждого текста. Классификатор – алгоритм присваивания определенного класса конкретным текстам. Обучающая модель – алгоритм поиска общих закономерностей между входными данными и настройка классификатора на корректное разделение входных данных по соответствующим классам.

Текстовая база данных представляет собой набор текстов, при обучении они должны быть маркированы по тематикам. Вся группа текстов делится на обучающую и тестовую выборки. В статье используется 3 корпуса текстов, на которых и происходит апробация алгоритмов классификации.

Коллекция Reuters-21578 содержит около 20 тыс. текстов. В данном наборе тексты неравномерно распределены по тематикам. Для опыта были отобраны тексты на 10 наиболее популярных тем.

Корпус текстов 20NewsGroups содержит по 1000 текстов на 20 различных тематик. Для опыта были отобраны 10 случайных тем. Одна из сложностей классификации – наличие большого количества лишней информации в текстах, не связанных с их тематикой.

Коллекция myCorp – группа из 500 текстов на 10 различных тематик, собранная вручную. Тексты в этой выборке достаточно легко разделить по темам, так как темы практически не пересекаются между собой по смыслу. Недостаток коллекции – ее небольшой размер по сравнению с другими выборками.

Вектор признаков составляется на основе частоты встречаемости слов в тексте. В рамках разработки программы классификации текстовых данных был реализован алгоритм их первичной обработки. Алгоритм выполняет парсинг текста, при этом игнорируются служебные символы, встречающиеся в нем. Для полученных слов необходимо выполнить операцию нахождения основы слова (она позволит избавиться от ситуации, когда одни и те же слова, но написанные с использованием разных окончаний, воспринимаются как разные слова), для этого используется достаточно распространенный алгоритм нахождения основы слова – стеммер Портера [2].

Затем происходит наполнение локальных словарей каждого текста. На этом этапе отфильтровываются общеупотребительные слова английского языка, для чего используется набор из 200 наиболее популярных слов, которые не имеют привязки к конкретным областям. Это позволяет убрать из рассмотрения слова, наличие которых никак не способствует пониманию тематики текста. К ним относятся: союзы, предлоги, местоимения, служебные слова и т. д. [3]. Все остальные слова попадают в локальные словари. После завершения анализа всех текстов происходит формирование глобального словаря, который содержит все слова из кластера текстов. Для дополнительного уменьшения размерности словаря выполняется следующий этап фильтрации слов: удаляются слова, суммарная встречаемость которых внутри всей выборки текстов ниже порогового значения ε. Данный вид фильтрации позволяет значительно уменьшить размерность вектора признаков за счет удаления редко встречающихся слов. Вклад подобных слов в качество работы классификатора мал, так как значение каждого из них внутри каждого вектора признаков будет крайне низким и, в результате, малоинформативным. Также это позволит удалить слова, в которых по различным причинам допущены орфографические ошибки. Результаты фильтрации можно увидеть в табл. 1.

Таблица 1

Параметры фильтрации	Размерность вектора признаков для разных коллекций текстов		
	Reuters	20NewsGroups	myCorp
Кол-во текстов	7447	9931	500
До фильтрации	44 821	53 692	26 382
ε = 5	6489	18248	7488
ε = 10	3830	11633	4782
ε = 20	2316	7583	3037
ε = 30	1776	5815	2335
ε = 50	1227	4128	1567
ε = 100	743	2572	827
ε = 200	431	1500	371
ε = 400	211	779	127

После получения текстовых данных, разбитых на отдельные слова, был реализован алгоритм создания вектора признаков для каждого текста. Размерность этого вектора напрямую зависит от размера словаря, принадлежащего данной коллекции текстов. Вектор признаков состоит из значений, каждое из которых позволяет характеризовать частоту встречаемости слов в тексте, согласно формуле

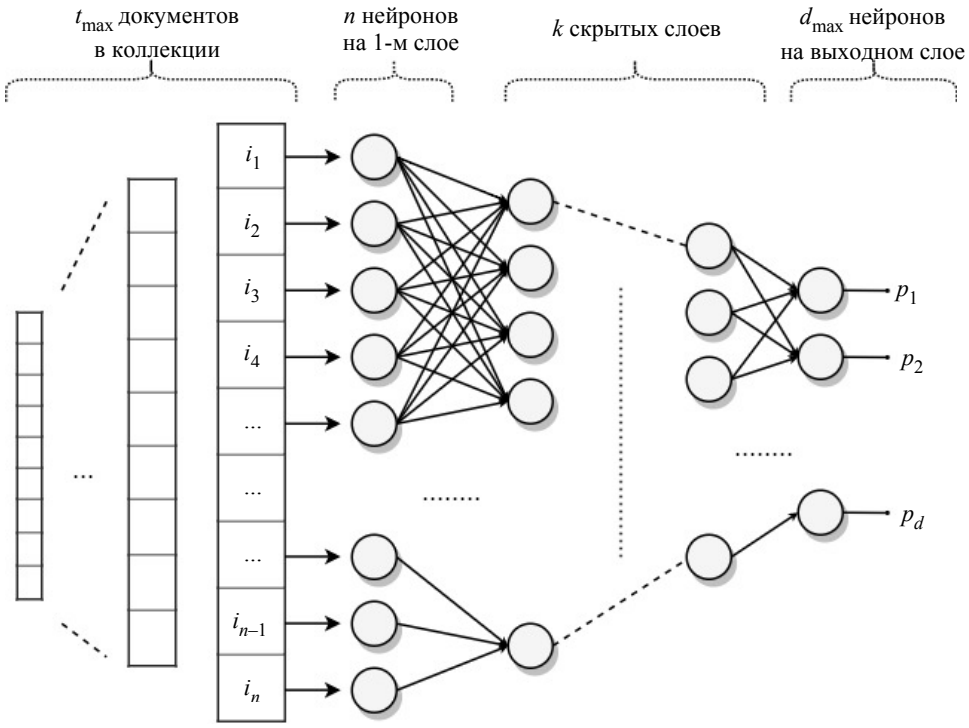

Рис. 2

$$tf_i = \frac{n_{id}}{\sum_{j=1}^{i_{max}} n_{jd}},$$

где tf_i – частота употребления слова i в документе d; n_{id} – количество слов i в документе d; i_{max} – размер словаря; $\sum_{j=1}^{i_{max}} n_{jd}$ – количество слов в документе d.

Нейронная сеть создавалась при помощи библиотеки Neuroph для языка Java. С помощью отдельных объектов из этой библиотеки, – нейрон, слой, функция активации, правила обучения и др., – разрабатывалась нейронная сеть, изображенная на рис. 2.

Нейронная сеть представлена сетью прямого распространения. Для обучения сети использовался метод обратного распространения ошибок. Каждый ее узел представлен искусственным нейроном, который служит аналогом модели естественного нейрона. Поступившая на вход линейная комбинация всех входных сигналов проходит через сигмоидальную функцию активации – передаточную функцию для отправки сигнала на нейроны последующего слоя сети [4]. Размер входного слоя сети зависит от размера словаря того корпуса, на котором происходит обучение сети. Количество скрытых слоев k может быть любым включая 0, и оптимальное значение необходимо определять эмпирически. Размер выходного слоя зависит от количества тем – в рассматриваемом примере эта величина всегда равняется 10. Значения на выходных нейронах характеризуют вероятность принадлежности текста к каждому из классов. Так, при определении принадлежности текста находится нейрон с наибольшим выходным значением – номер этого нейрона и будет определять номер класса.

Коллекция текстов Reuters содержит неравномерное распределение текстов по темам, из-за чего стандартная оценка правильности алгоритма, зависящая от соотношения правильно угаданных тем к общему количеству текстов, не подходит, так как при сильном смещении текстов обучающей выборки в сторону одного класса классификатор может принимать адекватные решения только для этого класса. Это приводит к тому, что классификатор может выдавать хорошую общую оценку точности, но при этом точность определения отдельных классов будет практически нулевой. Для оценки необходимо использовать F-меру, так как эта метрика учитывает средние показатели качества классификации по каждому классу, а не их суммарную точность классификации. Вычисление F-меры для оценки решения задачи классификации при количестве классов больше двух требует построения матрицы ошибок M размером $N \times N$, где N – это количество классов.

Столбцы этой матрицы отвечают за номер класса, которому принадлежит документ d, строки – за номер класса, к которому классификатор причислил этот документ. Тогда при решении задачи классификации для каждого документа происходит увеличение на единицу элемента на пересечении необходимых строки и столбца – полученная матрица позволяет наглядно увидеть результаты работы классификатора [5].

Необходимо рассчитать усредненные точность $\bar{P}$ и полноту $\bar{R}$ по всей матрице ошибок:

$$\bar{P} = \sum_{i=1}^{N} \frac{M_{i,i}}{\sum_{k=1}^{N} M_{i,k}};$$

$$\bar{R} = \sum_{i=1}^{N} \frac{M_{i,i}}{\sum_{k=1}^{N} M_{k,i}},$$

где i – номер строки, для которой вычисляется точность $\bar{P}$, или номер столбца, для которого вычисляется полнота $\bar{R}$ матрицы ошибок; M – матрица ошибок; k – каждый элемент строки при вычислении точности $\bar{P}$ или каждый элемент столбца при вычислении полноты $\bar{R}$.

F-мера вычисляется как гармоническое среднее между точностью и полнотой:

$$F = 2\frac{\bar{P} \times \bar{R}}{\bar{P} + \bar{R}}.$$

Сначала необходимо определить влияние параметра фильтра ε на качество работы алгоритма классификации. Для этого была проведена серия опытов с разными наборами данных на сети без скрытых слоев, при этом нейронная сеть была представлена только входным и выходным слоями, размеры которых зависят от размерности входного вектора и количества тем соответственно. Результаты экспериментов отражены в табл. 2. Согласно таблице: $t_{\text{обуч}}$ – время обучения сети; n – точность работы классификатора по методу оценки на основе соотношения количества правильно угаданных текстов к суммарному числу текстов в корпусе; F – точность классификатора по F-мере; ε – параметр фильтра.

Таблица 2

Параметры алгоритма	Результаты классификации для разных коллекций документов								
	Reuters-21578			20NewsGroups			myCorp		
ε	n	F	$t_{\text{обуч}}$	n	F	$t_{\text{обуч}}$	n	F	$t_{\text{обуч}}$
5	–	–	–	–	–	–	0.85	0.87	5:02
10	–	–	–	–	–	–	0.91	0.92	2:12
20	0.85	0.7	5:24	0.6	0.68	2:04:40	0.9	0.91	59 с
30	0.87	0.76	3:24	0.72	0.75	22:32	**0.93**	**0.94**	**51 с**
50	0.9	0.82	2:19	0.7	0.76	17:57	0.89	0.89	35 с
100	**0.9**	**0.83**	**1:47**	**0.82**	**0.83**	**6:55**	0.88	0.82	10 с
200	0.9	0.82	1:12	0.82	0.82	4:30	0.7	0.55	4 с
400	0.89	0.78	20 с	0.8	0.8	3:01	0.57	0.51	2 с

Таблица 3

Параметры алгоритма		Коллекции текстов								
		Reuters-21578			20NewsGroups			myCorp		
ε	Нейронов в скрытом слое	n	F	$t_{\text{обуч}}$	n	F	$t_{\text{обуч}}$	n	F	$t_{\text{обуч}}$
30	5	–	–	–	–	–	–	0.78	0.73	19
30	10	–	–	–	–	–	–	**0.94**	**0.93**	**36 с**
30	20	–	–	–	–	–	–	**0.94**	**0.93**	**1:55**
30	40	–	–	–	–	–	–	0.9	0.91	8:44
100	10	**0.91**	**0.84**	**1:55**	0.8	0.83	13:04	0.91	0.91	11 с
100	20	0.91	0.83	4:08	**0.84**	**0.85**	**25:24**	0.91	0.91	24 с
100	50	0.89	0.81	14:01	0.83	0.84	1:28:54	0.87	0.88	1:29

По табл. 2 видно, что ε сильно влияет как на скорость работы сети, так и на качество классификации. Уменьшение ε приводит к тому, что нейронная сеть извлекает неверные признаки, соответствующие текстам, из-за чего сеть может продолжать улучшать свою производительность на обучающих данных, но хуже справляться с тестовой выборкой из-за ухудшения обобщающих качеств сети.

Значение ε подбирается в зависимости от размера выборки данных. Так, для набора текстов Reuters и 20NewsGroups при отсечении всех слов, встречаемость которых ниже 100 во всем корпусе, достигается наиболее высокий уровень производительности для нейронной сети без промежуточных скрытых слоев.

Далее выполняется серия опытов с фиксированным значением ε и переменным количеством нейронов в скрытом слое.

Сравнение изменения точности алгоритмов в процессе обучения сети при наилучших параметрах сети для каждой выборки данных представлено на рис. 3.

По графикам видно, что качество процесса обучения сильно зависит от исходного набора данных. Так, в выборке с неравномерным распределением тем наблюдается достаточно высокая частота угадываний при самом старте алгоритма, но при этом из-за низкой точности определения любого класса, которому принадлежит небольшое количество документов, наблюдается низкий показатель *F*-меры. При этом наилучший показатель оказывается именно в самой маленькой выборке, и связано это с тем, что данный корпус текстов обладает наименее зашумленными исходными данными и равномерным распределением помеченных данных по темам.

Для корпуса текстов Reuters-21578, который характеризуется неравномерным распределением документов по тематикам при сильном смещении текстов обучающей выборки в сторону одного класса, результаты классификации оказываются на уровне 91 % точности по соотношению правильно угаданных текстов к общему числу текстов и 84 % точности по *F*-мере. Оценки показывают, что для каждого класса наблюдаются различные результаты классификации и наилучший способ оценки – для самого большого класса текстов.

Корпус 20NewsGroups – это самая большая выборка данных, более того, в этих документах содержится большое количество данных, не имеющих отношения к конкретным тематикам, что сильно усложняет задачу классификации. Наилучшие оценки классификации – это 84 % точности и 85 % по *F*-метрике, наблюдаются при ε = 100 и 20 нейронах в скрытом слое.

Классификация на основе приготовленного корпуса данных из 500 текстов показала наибольшую эффективность алгоритма, обеспечивая точность более 93 % по двум метрикам при ε = 30 и 10 или 20 нейронах в скрытом слое.

Исходя из полученных результатов, можно сделать следующие выводы.

Фильтр редких слов, характеризуемый величиной ε, показал свою эффективность, но требует дополнительных статистических исследований. Значение ε должно зависеть от размера корпуса данных.

Решающий вклад в точность классификации вносит качество обучающих данных, причем качество оказалось важнее количества этих данных, так как обучение на выборке из 500 текстов показало себя лучше, чем обучение на выборках, размер которых в 14…20 раз больше.

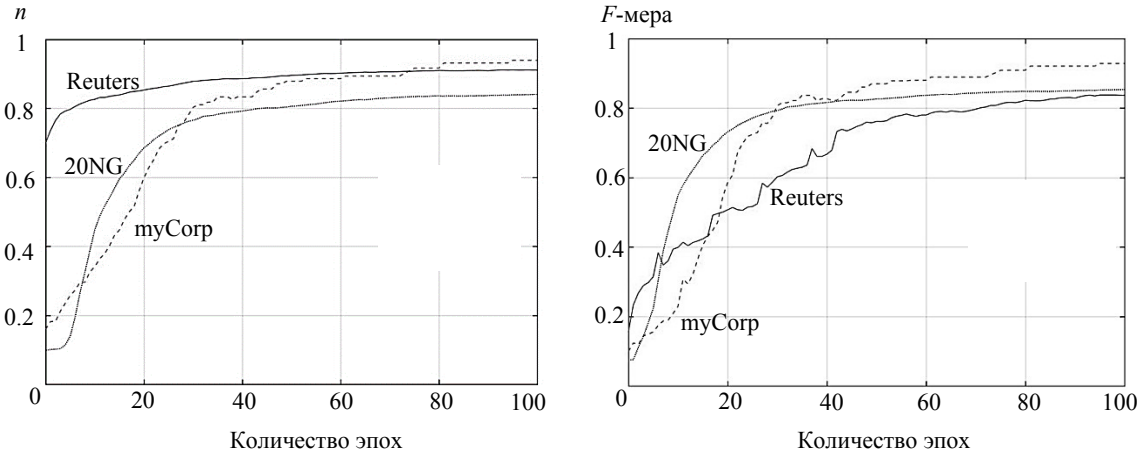

Рис. 3

Предложенная модель классификатора на основе нейронной сети позволяет решать задачу классификации с точностью больше 80 % при разных способах оценки. Определено, что точность во многом зависит от корпуса текстов, на котором будет происходить обучение сети.

Полученные в настоящей статье результаты станут основой для создания системы кластерного анализа данных, что позволит исключить необходимость ручной маркировки каждого текста в ситуации, когда есть большой корпус текстов, который необходимо разбить по темам.

СПИСОК ЛИТЕРАТУРЫ

1. Melnikov A. V., Botov D. S., Klenin J. D. On usage of machine learning for natural language processing tasks as illustrated by educational content mining // Ontology of designing. 2017. № 7. P. 34–47.

2. Porter M. F. An algorithm for suffix stripping // Program: electronic library and information systems. Vol. 14, № 3. P. 130–137.

3. Fox C. A Stop List for General Text // SIGIR Forum. 1990. Vol. 24, № 1–2. P. 19–35.

4. Стремоухов В. Д. Practical usage of artificial neural networks for the tasks of classification and image compression // Науч.-техн. вестн. информационных технологий, механики и оптики. 2006. № 27. С. 122–128.

5. Stapor K. Evaluating and comparing classifiers: Review, some recommendations and limitations // Proc. of the 10th Intern. Conf. on Computer Recognition Systems and Computing. URL: https://link.springer.com/book/10.1007/978-3-319-59162-9 (дата обращения 26.07.2020).

E. N. Karuna, P. V. Sokolov
Saint Petersburg Electrotechnical University

NEURAL NETWORK CLASSIFIER OF TEXT INFORMATION

The theoretical foundations of machine classification of text information are considered. Recently, there has been an increase in interest in this topic. The paper highlights the main stages and main difficulties in solving problems of this direction, presents the data obtained as a result of the work of a simple algorithm for the classification of text information. The preliminary filtering of texts, the formation of feature vectors, the structure and principles of training a neural network are discussed. The F-measure is used to evaluate the results. The comparison of the results for three collections of texts for different parameters of the preliminary filter, the number of neurons in the hidden layer and the training time of the network is carried out. The proposed model of the classifier allows solving the classification problem with an accuracy of more than 80% percent. In this case, the quality of the training data makes a decisive contribution to the classification accuracy. Conclusions about the quality of the results and options for further research on this topic are presented.

Classification, machine learning, thematic analysis, neural network, stemming