Выбор топологии нейронных сетей и их применение для классификации коротких текстов

О.С. Смирнова, В.В. Шишков

Аннотация **– В статье рассмотрена классификация текстов как подход, применяющийся для получения знаний из неструктурированных данных, в том числе и для определения тональности текстовых сообщений. С этой целью обусловлен выбор топологии нейронных сетей, описана методология создания и результаты тестирования двух классификаторов.**

Ключевые слова – ***топология нейросетей; сентимент-анализ; тональность текста; сверточные нейросети; сети Элмана; машинное обучение; предобработка текста.***

В настоящее время поток информации, приходящейся на одного человека в день и получаемой с помощью технических средств, составляет 27 мегабайт в день, это эквивалентно содержанию 176 газет, из них полезной информацией может считаться лишь небольшой процент [1, 2]. С учетом того, что возможности технических средств обработки информации, а также специализированных устройств и программ, предназначенных для создания контента и доставки его потребителю, удваиваются каждые 18 – 24 месяца, возможности успешного усвоения хотя бы 10 процентов информации для обычного человека падают с каждым годом [3].

Исторически сложилось так, что большая часть информации, полученной человеком с помощью технических средств, воспринимается в письменном виде, в виде текста на естественном языке. Это обусловлено простотой передачи и создания такого контента. Для упрощения ориентирования в таком потоке информации, начиная с 1960-х годов проводятся эксперименты по созданию все более и более продвинутых систем поддержки принятия решений в тех или иных областях. Зачастую, одной из проблем, с решением которых сталкивается такая система является извлечение знаний из неструктурированной текстовой информации. Полученные знания используются для формирования правил, на основании которых дается рекомендация действий пользователю. Это позволяет оградить рядового пользователя от переизбытка доступной информации, что оказывает положительное влияние на скорость принятия решения [4].

Среди подходов, применяющихся для получения знаний из неструктурированных данных, в том числе и определения тональности текстовых сообщений, одно из ключевых мест занимает классификация. Разделение множества элементов на различные группы не дает новой информации об исследуемых объектах напрямую, но позволяет установить связь между объектами, что позволяет экстраполировать некоторые отдельные свойства элементов на все множество, что используется, к примеру, при анализе социальных графов [5]. Применительно к текстам существует несколько подходов разбиения множества текстов на N непересекающихся множеств. К примеру, наивный байесовский классификатор – самый популярный метод классификации текстов нашел широкое применение в различных спам-фильтрах. К недостаткам такой системы относятся: ограниченный словарь значимых для системы слов и довольно большая необходимая вычислительная мощность при работе с большим объемом текста (проблема арифметического переполнения) [6].

Это не единственный подход к классификации текстов в машинном обучении. В последнее время в качестве инструмента классификации получают широкое распространение нейронные сети. Некоторые топологии нейросетей обладают специфическими свойствами, к примеру, для предсказания ввода последующего слова (предиктивный ввод) при заполнении поисковых форм используются рекуррентные нейросети с несколькими сверточными слоями, размещенных ближе к входным нейронам [7]. Такие же рекуррентные сети, но другой конфигурации возможно использовать для классификации текстов [8].

Одним из основных свойств нейросетей является возможность адаптации, то есть готовую нейросетевую модель можно обучить дополнительно, теоретически получив большую точность результатов работы. Это

Статья получена 14.07.2016 г.

Исследование выполнено федеральным государственным бюджетным образовательным учреждением высшего образования «Московский технологический университет» (МИРЭА) за счет гранта Российского фонда фундаментальных исследований (проект №16-37-00492).

О.С. Смирнова, МИРЭА (e-mail: mail.olga.smirnova@yandex.ru).

В.В. Шишков, МИРЭА (e-mail: shishkov61@gmail.com).

свойство особенно важно при использовании нейросети как классификатора зашумленных данных. Шум в данных позволяет избежать переобучения нейросети [9]. В совокупности это позволяет повысить качество распознавания данных с шумом в сравнении с более простыми способами классификации, такими как классификация с помощью каскадов правил. Вместе с тем, выбор топологии и характеристик, а также параметров обучения нейросети является нетривиальной задачей и выполняется с использованием эвристических подходов и различных «правил большого пальца», а от правильного выбора зависит релевантность получаемых результатов [10]. Время, необходимое для обучения нейросети часто велико, и для обучения классификатора может потребоваться до нескольких суток на машине с процессором Core i5 2011 года [11]. Возможно применение одновременных вычислений на графическом процессоре с помощью видеоадаптеров, поддерживающих одну из архитектур параллельных вычислений, к примеру, NVIDIA CUDA, но при этом в системах реального времени использование самообучающихся структур нецелесообразно [12].

Свойства нейронной сети практически полностью задаются выбором ее топологии. Существует порядка 25 известных на данный момент эффективных структур нейросетей [13]. При распознавании изображений лучший результат дает сверхточная нейронная сеть, в которой отсутствуют обратные связи, но за счет сверточного суммирования простых сигналов в «карты признаков» удалось сократить количество настраиваемых весов с отдельного коэффициента для каждого входного пикселя, как в перцептроне, до небольшого «ядра свертки», что позволяет производить быструю классификацию. Однако области применения топологий редко широки. Сверточная нейронная сеть не будет показывать хороших результатов при классификации текстов, однако некоторые из особенностей работы таких сетей пригодны для использования в гибридных сетях, что и будет продемонстрировано ниже [14].

Для каждой из задач, решаемых с применением нейросетей существует несколько общих правил выбора структуры сети. В 1990 году Дж. Л. Элман предлагает структуру сети, полученную из многослойного перцептрона введением обратных связей от выходов внутренних нейронов, получается, что выходной сигнал (реакция сети) зависит не только от текущего стимула, но и от предыдущего. Такие сети планировались для применения в системах управления динамическими объектами, но нашли более широкое применение при работе с текстами [15].

Для осуществления обучения сети на реальных текстах необходимо производить некоторые операции перед подачей входных векторов на обучение классификатора. Вид входных значений зависит от структуры нейросети, для типов нейросетей, упомянутых в данной статье в связи с тестированием топологий в качестве входных векторов используются кортежи натуральной чисел длиной в 30 значений, что позволяет подавать на вход классификатора отдельные предложения.

Используемые для обучения данные должны быть:

– репрезентативными, то есть всесторонне освещать моделируемый объект, при этом допустима избыточность;

– качественно перемешаны алгоритмом, использующим нормальное распределение случайных величин.

При этом для числовых некотегориальных данных необходимо проводить нормировку и центрирование, то есть входные данные должны лежать в интервале [-1;1] [16].

Как правило, при обучении нейросетей используется метод обратного распространения ошибки, но существуют более эффективные алгоритмы обучения, к примеру, BFGS (Broyden-Fletcher-Goldfarb-Shanno) или Conjugate Gradients [17, 18].

При обучении сети словам естественного языка ставятся в соответствие числовые коды по словарю, для этого морфологические вариативности в составе слова детерменируются с помощью стеммирования. Для составления словаря также должно использоваться стеммирование до установки соответствия слова натуральному числу. Стеммирование – это процесс нахождения основы слова для исходной словоформы. При этом, основа слова не всегда совпадает с морфологическим корнем слова [19].

В данной статье для тестирования на задаче определения тональности текстовых сообщений выбраны 2 нейросетевых топологии: нейросеть с двумя рекуррентными слоями (рисунок 1) и нейросеть с рекуррентным и сверточным слоями (рисунок 2). Для описания используется формальный язык фреймворка Python Keras [20].

```
model = Sequential()
model.add(Embedding(max_features, embedding_size,
input_length=maxlen))
model.add(LSTM(64, return_sequences=True))
model.add(LSTM(64))
model.add(Dropout(0.5))
model.add(Dense(1))
model.add(Activation('sigmoid'))
```

Рисунок 1 – Структура нейросети с двумя рекуррентными слоями

```
model = Sequential()
model.add(Embedding(max_features, embedding_size,
input_length=maxlen))
model.add(Dropout(0.25))
model.add(Convolution1D(nb_filter=nb_filter,
                        filter_length=filter_length,
                        border_mode='valid',
                        activation='relu',
                        subsample_length=1))
model.add(MaxPooling1D(pool_length=pool_length))
model.add(LSTM(lstm_output_size))
model.add(Dense(1))
model.add(Activation('sigmoid'))
```

Рисунок 2 – Структура нейросети с 1 рекуррентным и 1 сверточным слоями

Остальные параметры сети представлены в следующей конфигурации:

```
# Data
data_count = 10000
# Embedding
max_features = 20000
maxlen = 30
embedding_size = 128
# Convolution
filter_length = 3
nb_filter = 64
pool_length = 2
# LSTM
lstm_output_size = 70
# Training
batch_size = 30
nb_epoch = 5
```

При обучении нейросети использовались два набора размеченных вручную по тональности твитов, т.е. коротких сообщений длиной до 140 символов. В обучающую выборку попали 10000 элементов из каждого набора [21].

Перед составлением словаря проводилась стеммизация с помощью пакета pystemmer и очистка, использующая набор стоп-слов для русского языка, представленных порталом Algorithmist [22, 23].

Полученный словарь (рисунок 3) имеет мощность порядка 176 тысяч слов и хранится в формате json.


```
"5yxnkji0p0": 173855,
"обезьяны": 41902,
"иностранцу": 173856,
"60027": 173857,
"девуль": 173858,
"смысла": 6414,
"violettaya": 173860,
"nodehierarchy": 173861,
"свинарник": 62241,
"мтбанк": 173862,
"nekitoooos": 173863,
"союзнефтегаз": 173864,
"паадала": 173865,
"комиксами": 173866,
"барбуды": 173867,
"аята": 173868,
"тюремные": 173869,
"бар…": 104336,
"упущенно": 173870,
"noeolnpnal": 173871,
"программированием": 173872,
"всеровн": 26310,
"isycwl58y8": 173873,
"екатерининский": 173874,
"ftbygwjs7j": 173875,
```


Рисунок 3 – Часть словаря токенизации

Если входной вектор получается после очистки и токенизации по длине меньше 30, то недостающая длина дополняется нулями.

Для тестирования обученных моделей используется 20000 случайно выбранных элементов исходного набора, не участвующих в процессе обучения. Тестирование проводилось с помощью встроенных механизмов фреймворка Theano [24]. Для сети с двумя рекуррентными слоями количество удачно классифицированных документов составляет 69%. Для сети с рекуррентным и сверточным слоями и процент «попадания» немного выше – 71%.

```
>>> evaluate_one('посмотрели прекрасную пьесу, игра актеров превосходна')
[[    0     0     0     0     0     0     0     0     0     0     0     0
      0     0     0     0     0     0     0     0     0     0     0     0
      0   417   434   485  2927 11271]]

1/1 [==============================] - 0s
array([[ 0.93522263]])
```

Рисунок 4 – Пример работы сети с 1 рекуррентным слоем

Таким образом, в статье были частично освещены методы извлечения знаний из неструктурированных данных, при этом более пристальное внимание уделено проблеме классификации. Изложены основные принципы использования нейронных сетей при работе с текстовыми данными в качестве классификатора, рассмотрены наиболее часто используемые топологии нейронных сетей и описана методология и результаты тестирования применения двух экземпляров сети с различными структурами, использующими слои рекуррентных нейронов, для определения тональности коротких текстовых сообщений. Наиболее точный результат классификации – 71% получен при использовании сети с рекуррентным и сверточным слоями.

БИБЛИОГРАФИЯ

[1] Richard Alleyne. Welcome to the information age. Эл. ресурс: http://www.telegraph.co.uk/news/science/science-news/8316534/Welcome-to-the-information-age-174-newspapers-a-day.html (дата обращения: 17.05.2016 г.).

[2] Стратонович Р.Л. Теория информации. — М.: Советское радио, 1975.

[3] Moore, Gordon E. No Exponential is Forever: But «Forever» Can Be Delayed!. International Solid-State Circuits Conference (ISSCC) 2003 / SESSION 1 / PLENARY / 1.1.

[4] Keen, P. G. W. (1978). Decision support systems: an organizational perspective. Reading, Mass., Addison-Wesley Pub. Co. ISBN 0-201-03667-3

[5] Wikipedia. Statistical classification. Эл. ресурс: https://en.wikipedia.org/wiki/Statistical_classification (дата обращения: 17.05.2016 г.).

[6] Д.Баженов. Наивный байесовский классификатор. Эл. ресурс: http://bazhenov.me/blog/2012/06/11/naive-bayes.html (дата обращения: 23.05.2016 г.).

[7] Vincent A. Akpan Adaptive predictive control using recurrent neural network identification. Эл. ресурс: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5164515&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D5164515 (дата обращения: 17.05.2016 г.).

[8] Wikipedia. Recurrent neural network. Эл. ресурс: https://en.wikipedia.org/wiki/Recurrent_neural_network (дата обращения: 17.05.2016 г.).

[9] Richard M. Zur, Yulei Jiang, Lorenzo L. Pesce, and Karen Drukker. Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. Эл. ресурс: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2771718/ (дата обращения: 23.05.2016 г.).

[10] Wikipedia. Нейронные сети / Выбор топологии сети. Эл. ресурс: https://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть#.D0.92.D1.8B.D0.B1.D0.BE.D1.80_.D1.82.D0.BE.D0.BF.D0.BE.D0.BB.D0.BE.D0.B3.D0.B8.D0.B8_.D1.81.D0.B5.D1.82.D0.B8 (дата обращения: 25.05.2016 г.).

[11] Habrahabr. Обучение OpenCV каскада Хаара. Эл. ресурс: https://habrahabr.ru/post/208092/ (дата обращения: 25.05.2016 г.).

[12] NVIDIA Параллельные вычисления CUDA. Эл. ресурс: http://www.nvidia.ru/object/cuda-parallel-computing-ru.html (дата обращения: 27.05.2016 г.).

[13] Wikipedia. Нейронные сети / Известные типы сетей. Эл. ресурс: https://ru.wikipedia.org/wiki/Искусственная_нейронная_сеть#.D0.98.D0.B7.D0.B2.D0.B5.D1.81.D1.82.D0.BD.D1.8B.D0.B5_.D1.82.D0.B8.D0.BF.D1.8B_.D1.81.D0.B5.D1.82.D0.B5.D0.B9 (дата обращения:27.05.2016 г.).

[14] Wikipedia. Сверточная нейронная сеть. Эл. ресурс: https://ru.wikipedia.org/wiki/Сверточная_нейронная_сеть (дата обращения: 29.05.2016 г.).

[15] Wikipedia. Нейронная сеть Элмана. Эл. ресурс: https://ru.wikipedia.org/wiki/Нейронная_сеть_Элмана (дата обращения: 27.05.2016 г.).

[16] Russell C. Eberhart, Roy W. Dobbins – Neural Network PC Tools: A Practical Guide – Academic Press, 28 Jun. 2014 – pp. 90 – 134.

[17] Wikipedia. BFGS. Эл. ресурс: https://en.wikipedia.org/wiki/Broyden–Fletcher–Goldfarb–Shanno_algorithm (дата обращения: 29.05.2016 г.).

[18] Wikipedia. CG. Эл. ресурс: https://en.wikipedia.org/wiki/Conjugate_gradient_method (дата обращения: 27.05.2016 г.).

[19] Портал tartarus.org Russian stemming algorithm. Эл ресурс: http://snowball.tartarus.org/algorithms/russian/stemmer.html (дата обращения: 29.05.2016 г.).

[20] Портал keras.io Документация Keras. Эл. ресурс: http://keras.io/getting-started/faq/#why-is-the-training-loss-much-higher-than-the-testing-loss (дата обращения: 27.05.2016 г.).

[21] Ю.В.Рубцова. Построение корпуса текстов для настройки тонового классификатора. Программные продукты и системы, 2015, №1(109), – С.72-78.

[22] Github, хостинг проектов Pystemmer Эл. ресурс: https://github.com/snowballstem/pystemmer (дата обращения: 25.05.2016 г.).

[23] Портал Algorithmist. Стоп символы русского языка. Эл. ресурс: http://www.algorithmist.ru/2010/12/stop-symbols-in-russian.html (дата обращения: 27.05.2016 г.).

[24] Портал deeplearning.net. Документация. Theano Эл. ресурс: http://deeplearning.net/software/theano/ (дата обращения: 27.05.2016 г.).

The choice of the topology of neural networks and their use for the classification of small texts

O.S. Smirnova, V.V. Shishkov

Annotation – **In this article, we describe the classification of texts as the approaches used to gain knowledge from unstructured data. This approach includes determining messages tone. We describe the choice of the topology for neural networks and the methodology for classifiers creation. Our paper presents testing results for two created classifiers.**

KeyWords – **topology of neural networks; sentiment-analysis; tonality; convolutional neural networks; Ellman's network; machine learning; text preprocessing.**