ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 519.6 DOI 10.25730/VSU.0536.19.003

Исследование методов выбора оптимального количества признаков для решения задачи определения точки зрения автора текста*

С. В. Вычегжанин[1], Е. В. Котельников[2], Е. В. Разова[3]

[1]аспирант кафедры прикладной математики и информатики, Вятский государственный университет. Россия, г. Киров. E-mail: vychegzhaninsv@gmail.com

[2]кандидат технических наук, доцент кафедры прикладной математики и информатики, Вятский государственный университет. Россия, г. Киров. E-mail: kotelnikov.ev@gmail.com

[3]кандидат педагогических наук, доцент кафедры прикладной математики и информатики, Вятский государственный университет. Россия, г. Киров. E-mail: razova.ev@gmail.com

Аннотация. При выполнении процедуры отбора признаков (feature selection) на этапе предварительной обработки данных в машинном обучении возникает проблема выбора количества признаков, которые должны входить в результирующее множество. Существуют различные подходы к выбору количества признаков, позволяющие получить качество, близкое к оптимальному. В настоящей работе исследуются несколько методов выбора оптимального количества признаков для решения задачи определения точки зрения автора текста. Эксперименты проводятся с использованием трех текстовых корпусов, составленных из русскоязычных сообщений пользователей интернет-форумов. По результатам экспериментов наилучшим среди рассмотренных методов оказался метод максимума качества, позволивший сократить количество признаков в среднем на 62,6% от их общего числа, сохранив при этом качество на прежнем уровне.

Ключевые слова: определение точки зрения автора текста, методы отбора признаков, оптимальное множество признаков.

Введение

Каждый год в России наблюдается рост популярности социальных медиа, таких как социальные сети, блоги и микроблоги, форумы и сайты отзывов. По результатам исследований, регулярно проводимых компанией Brand Analytics, в мае 2017 года в социальных сетях было зафиксировано 38 млн активных авторов и 670 млн сгенерированных ими сообщений, а в октябре 2018 года – 46 млн активных авторов и 1,8 млрд сообщений. В этом текстовом массиве содержится ценная информация для маркетологов, социологов, политологов и других специалистов, анализирующих мнения людей. Для обработки такого огромного объема информации требуется создание автоматических средств анализа данных.

В настоящей статье рассматривается задача определения точки зрения автора текстового документа (англ. stance detection). Данная задача состоит в выявлении позиции, которой придерживается автор текста, по отношению к объекту (или объектам) обсуждения [5]. Выделяют два основных класса позиций [7]:

1. «За» – по тексту можно определить, что автор высказывается в поддержку целевого объекта. Например, для целевого объекта *прививки детям* позиция *за* выражена в тексте: «Прививаю. Иммунолог – подруга семьи, так что проблем нет. Опять же, на мой взгляд, если придумали прививки, то не просто так».
2. «Против» – по тексту можно определить, что автор высказывается против целевого объекта. Например, для целевого объекта *ЕГЭ в школе* позиция *против* выражена в тексте: «И ещё утверждают, что ЕГЭ лучше советских экзаменов. Не было в СССРе такого безобразия – НЕ БЫЛО...»

Также в литературе выделяют классы «нейтрально», «невозможность определения точки зрения» и «согласие с предыдущей точкой зрения».

В настоящей статье проводится экспериментальное исследование методов выбора оптимального количества признаков для решения задачи определения точки зрения автора текстового документа.

** Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ, государственное задание ВятГУ № 34.2092.2017/4.6.*

Обзор методов выбора оптимального количества признаков

В машинном обучении на этапе предварительной обработки данных с целью удаления нерелевантных признаков осуществляется процедура отбора признаков. Такой подход позволяет сократить размерность пространства признаков и повысить производительность методов машинного обучения.

Методы отбора признаков можно поделить на три категории: методы-фильтры (filters), методы-обертки (wrappers) и встроенные методы (embedded). Методы-фильтры не зависят от алгоритма обучения и используют информацию, полученную из обучающей выборки. Они являются наиболее вычислительно эффективными методами. Методы-обертки взаимодействуют с алгоритмом обучения и могут учитывать зависимости в признаках, но являются вычислительно дорогими. Кроме этого, имеется риск их переобучения. Встроенные методы вычислительно более эффективные, чем методы-обертки, но ограничены конкретными алгоритмами обучения, например, деревьями решений.

При использовании процедуры отбора признаков важным вопросом является выбор количества признаков, которые должны содержаться в результирующем множестве. Для этого применяются эвристические приемы, которые часто позволяют получить решение, близкое к оптимальному. В экспериментальной части настоящей статьи исследуются четыре подхода к определению оптимального количества признаков:

1. Константные значения

Существуют работы, в которых пороговые значения количества признаков принимаются постоянными и составляют фиксированный процент от общего числа признаков [2; 1]. Недостатком данного подхода является то, что он не учитывает влияние обучающих данных на выбор оптимального количества признаков. В настоящей работе пороговые значения приняты равными 10%, 25% и 50% от общего числа признаков.

2. Значение, определяемое функцией от общего числа признаков

В статье [2] в качестве функции, определяющей пороговое значение количества признаков, применяется $log_2 N$.

3. Метод максимума качества (MAX)

В работе [3] предлагается подход с использованием процедуры перекрестной проверки. Пусть N – общее количество признаков на обучающих данных. Тогда оптимальное число признаков определяется по следующему алгоритму:

1) Выполняется ранжирование всех N признаков с помощью некоторого метода-фильтра.

2) Строится модель с использованием первых k признаков и оценивается качество этой модели на тестовом множестве с использованием некоторой метрики качества. Параметр k изменяется от 1 до N.

3) Изображается график зависимости качества модели от количества признаков. Количество признаков, соответствующее максимальному качеству, принимается за оптимальное.

4. Отбор признаков, основанный на корреляции (Correlation-based Feature Selection, CFS).

Данный подход, предложенный в работе [4], относится к методам-фильтрам, но является полностью автоматическим, не требующим задания количества признаков, необходимых для выбора. Метод CFS оценивает подмножества признаков, основываясь на гипотезе о том, что хорошее подмножество содержит признаки, сильно коррелируемые с метками классов и слабо коррелируемые друг с другом. Согласно такой гипотезе нерелевантные признаки имеют низкую корреляцию с метками классов и игнорируются алгоритмом.

Методы и инструменты

Экспериментальное исследование проводилось с использованием языка программирования Python. На этапе предварительной обработки текстов все слова приводились к начальной форме на основе модуля *pymystem3*. Ранжирование слов в текстовых документах осуществлялось с помощью метода-фильтра Индекс Джини (Gini Index, GI), наиболее производительного по результатам исследования [9]. В процессе проведения экспериментов была использована реализация метода GI из библиотеки *scikit-feature*, реализация метода CFS – из библиотеки *Weka*, написанной на языке программирования Java. Для классификации текстов применялся метод опорных векторов (Support Vector Machine), реализованный в библиотеке *scikit-learn* [6]. Для получения объективных оценок использовалась процедура 5-кратной перекрестной проверки (5-fold cross-validation). Качество классификации оценивалось с помощью F1-меры.

Эксперименты проводились на основе трех текстовых корпусов, составленных из русскоязычных сообщений пользователей интернет-форумов. Характеристики корпусов представлены в табл. 1.

Таблица 1

Характеристики текстовых корпусов

Название корпуса	Метка текста	Количество текстов	Общее количество слов	Средняя длина текста, слов	Размер словаря
Прививки детям	за	500	35 326	70	5 100
	против	500	34 167	68	
ЕГЭ в школе	за	600	35 410	59	5 943
	против	800	40 453	51	
Клонирование человека	за	450	18 860	42	4 990
	против	650	27 405	42	

Результаты экспериментов

В табл. 2 представлено количество признаков, выбранных в каждом блоке процедуры 5-кратной перекрестной проверки, в табл. 3 – соответствующие значения F1-меры.

Таблица 2

Количество признаков

Название корпуса	10%	25%	50%	100%	$\log_2 N$	MAX	CFS
Прививки детям	510	1275	2550	5100	12	3480	49
	510	1275	2550	5100	12	1160	62
	510	1275	2550	5100	12	3970	64
	510	1275	2550	5100	12	850	62
	510	1275	2550	5100	12	3370	59
ЕГЭ в школе	595	1487	2975	5943	12	4020	53
	595	1487	2975	5943	12	3900	104
	595	1487	2975	5943	12	3490	63
	595	1487	2975	5943	12	4650	110
	595	1487	2975	5943	12	5900	54
Клонирование человека	499	1247	2495	4990	12	2920	34
	499	1247	2495	4990	12	2100	91
	499	1247	2495	4990	12	3320	92
	499	1247	2495	4990	12	3180	79
	499	1247	2495	4990	12	3890	84
Среднее значение	**535**	**1336**	**2673**	**5345**	**12**	**3347**	**71**

Таблица 3

Значение F1-меры, %

Название корпуса	10%	25%	50%	100%	$\log_2 N$	MAX	CFS
Прививки детям	78,5	79,0	79,5	78,4	67,5	80,5	72,0
	75,0	72,4	73,0	73,9	69,9	72,4	76,0
	75,9	78,5	78,5	77,0	62,2	76,5	65,5
	79,0	76,5	76,5	78,0	66,8	78,5	71,0
	74,5	77,0	75,5	74,0	66,4	75,5	71,5
ЕГЭ в школе	71,3	72,0	71,6	75,4	63,1	71,6	65,6
	75,0	77,9	78,2	77,1	58,7	77,5	70,2
	74,1	74,4	72,0	73,3	64,2	72,8	71,5
	75,9	73,4	74,4	76,4	62,7	75,0	73,2
	66,0	67,8	68,6	69,4	59,6	69,8	69,3
Клонирование человека	71,5	75,1	74,0	73,5	68,2	74,1	68,0
	72,1	72,5	75,8	75,0	62,5	75,3	72,2
	73,9	75,8	73,5	74,1	64,8	75,4	69,0
	71,2	73,2	72,9	75,8	62,6	74,0	66,2
	71,2	74,9	73,9	75,3	64,9	74,3	63,9
Среднее значение	**73,7**	**74,7**	**74,5**	**75,1**	**64,3**	**74,9**	**69,7**

На основании результатов из табл. 3 можно сделать вывод, что лучшее качество классификации достигается при использовании 100% признаков. Среди четырех рассмотренных подходов к выбору оптимального количества признаков лучшим оказался метод MAX, который позволил получить близкое к наилучшему качество при использовании в среднем 62,6% признаков от их общего числа. Применение в качестве порогового значения констант, равных 25% и 50% признаков, незначительно уступает методу MAX по F1-мере на 0,2% и 0,4% соответственно. При этом проверка ста-

тистической значимости результатов с использованием критерия знаковых рангов Уилкоксона (Wilcoxon signed rank test) [8] показала, что результаты для методов 25%, 50% и MAX статистически незначимо отличаются от результатов для 100% признаков на уровне значимости p=0,05. Однако для методов 10%, $\log_2 N$ и CFS отличия статистически значимы. Это объясняется малым количеством выбранных признаков.

Заключение

Таким образом, при сокращении размерности пространства признаков в проведенном исследовании лучшим из всех методов по F1-мере оказался MAX. Кроме этого, хорошие результаты дает метод константных значений, который трудно превзойти более сложными методами.

Список литературы

1. *Bolon-Canedo V., Alonso-Betanzos A.* Recent Advances in Ensembles for Feature Selection. 2018. 205 p.
2. Ensemble feature selection: Homogeneous and heterogeneous approaches / B. Seijo-Pardo et al. // Knowledge-Based Systems. 2017. Vol. 118. P. 124–139.
3. Finding the optimal number of features based on mutual information / P. Chen et al. // Proceedings of EUSFLAT-2017. 2018. Vol. 641. P. 477–486.
4. *Hall M. A.* Correlation-based Feature Selection for Machine Learning: PhD dissertation Department of Computer Science. Waikato University, Hamilton, NZ, 1999. 198 p.
5. Joint Models of Disagreement and Stance in Online Debate / D. Sridhar et al. // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. P. 116–125.
6. Scikit-learn: Machine Learning in Python / F. Pedregosa et al. // JMLR. 2011. Vol. 12. P. 2825–2830.
7. *Vychegzhanin S. V., Kotelnikov E. V.* Stance Detection in Russian: a Feature Selection and Machine Learning Based Approach // Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017) July 2017. 2017. Vol. 1975. P. 166–179.
8. *Wilcoxon F.* Individual comparisons by ranking methods // Biometrics Bulletin. 1945. Vol. 1, № 6. P. 80–83.
9. *Вычегжанин С. В., Котельников Е. В.* Экспериментальное исследование методов отбора признаков для решения задачи определения точки зрения автора текста // Информатика: проблемы, методология, технологии: материалы XIX Международной научно-методической конференции. Воронеж: Издательский дом ВГУ, 2019.

Study of methods for selecting the optimal number of features to solve the stance detection task

S. V. Vychegzhanin[1], E. V. Kotelnikov[2], E. V. Razova[3]

[1]post-graduate student of the Department of applied Mathematics and Computer Science, Vyatka State University. Russia, Kirov. E-mail: vychegzhaninsv@gmail.com

[2]PhD of technical sciences, associate professor of applied Mathematics and Computer Science, Vyatka State University. Russia, Kirov. E-mail: kotelnikov.ev@gmail.com

[3]PhD of pedagogical sciences, associate professor of applied Mathematics and Computer Science, Vyatka State University. Russia, Kirov. E-mail: razova.ev@gmail.com

Abstract. When performing the feature selection procedure at the stage of data preprocessing in machine learning, there is a problem of selecting the number of features that should be included in the result set. There are different approaches to the choice of the number of features, allowing to obtain a quality close to optimal. In this paper, we study several methods for selecting the optimal number of features to solve the stance detection task. The experiments are carried out using three text corpora made up of Russian-language messages from users of Internet forums. According to the results of the experiments, the best among the methods considered was the method of maximum quality, which allowed to reduce the number of features on average by 62.6% of their total number, while maintaining the quality at the same level.

Keywords: stance detection, methods of feature selection, optimal feature set.

References

1. *Bolon-Canedo V., Alonso-Betanzos A.* Recent Advances in Ensembles for Feature Selection. 2018. 205 p.
2. Ensemble feature selection: Homogeneous and heterogeneous approaches / B. Seijo-Pardo et al. // Knowledge-Based Systems. 2017. Vol. 118. P. 124–139.
3. Finding the optimal number of features based on mutual information / P. Chen et al. // Proceedings of EUSFLAT-2017. 2018. Vol. 641. P. 477–486.
4. *Hall M. A.* Correlation-based Feature Selection for Machine Learning: PhD dissertation Department of Computer Science. Waikato University, Hamilton, NZ, 1999. 198 p.

5. Joint Models of Disagreement and Stance in Online Debate / D. Sridhar et al. // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015. P. 116–125.

6. Scikit-learn: Machine Learning in Python / F. Pedregosa et al. // JMLR. 2011. Vol. 12. P. 2825–2830.

7. *Vychegzhanin S. V., Kotelnikov E. V.* Stance Detection in Russian: a Feature Selection and Machine Learning Based Approach // Supplementary Proceedings of the Sixth International Conference on Analysis of Images, Social Networks and Texts (AIST 2017) July 2017. 2017. Vol. 1975. P. 166–179.

8. *Wilcoxon F.* Individual comparisons by ranking methods // Biometrics Bulletin. 1945. Vol. 1, № 6. P. 80–83.

9. *Vychegzhanin S. V., Kotel`nikov E. V. Eksperimental`noe issledovanie metodov otbora priznakov dlya resheniya zadachi opredeleniya tochki zreniya avtora teksta* [Experimental study of methods of selection of characteristics for the solution of the problem of determining the point of view of the author of the text // *Informatika: problemy, metodologiya, tehnologii: materialy XIX Mezhdunarodnoj nauchno-metodicheskoj konferentsii* – Informatics: problems, methodology, technology: materials of XIX International scientific-methodical conference. Voronezh. Publishing house of VSU. 2019.