DOI 10.37882/2223–2966.2020.06.19

ПРОБЛЕМА АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ КОЛИЧЕСТВА КЛАСТЕРОВ В ЗАДАЧЕ КЛАСТЕРИЗАЦИИ КОНТЕКСТОВ

THE PROBLEM OF AUTOMATIC DETERMINATION THE NUMBER OF CLUSTERS IN THE CLUSTERING CONTEXTS TASK

A. Kapitanov
V. Troyanovsky

Summary. When solving the task of clustering contexts, we face to the problem of automatically determining the number of clusters. Clustering of contexts allows us to effectively solve the problem of homonymy, which in turn leads to an increase in the quality of several problems in computational linguistics. Using the example of the text document classification problem, we will try to calculate the required number of clusters to increase the percentage of recognized documents. In the course of work, based on the DBSCAN density algorithm, we were able to calculate the number of clusters, then, based on agglomerative hierarchical clustering, break down homonymous contexts into clusters and remove homonymy. After that, we checked the quality of classification based on the naive Bayesian classifier algorithm and made sure that the percentage of correctly recognized documents increased.

Keywords: hierarchical clustering, cluster analysis, classification, polysemy, DBSCAN.

Капитанов Андрей Иванович
Ассистент, Национальный исследовательский университет «МИЭТ»
andrey@kapitanov.me
Трояновский Владимир Михайлович
Д.т.н., профессор, Национальный исследовательский университет «МИЭТ»
troy40@mail.ru

Аннотация. При решении задачи кластеризации контекстов возникает проблема автоматического определения количества кластеров. Кластеризация контекстов позволяет эффективно разрешать проблему омонимии, что в свою очередь приводит к повышению качества ряда задач компьютерной лингвистики. На примере задачи классификации текстовых документов мы попытаемся вычислить необходимое количество кластеров для повышения доли распознанных документов. В ходе работы на основе плотностного алгоритма DBSCAN нам удалось вычислить количество кластеров, далее на основе агломеративной иерархической кластеризации разбить омонимичные контексты на кластеры и снять омонимию. После этого мы проверили качество классификации на основе алгоритма наивного байесовского классификатора и убедились в увеличении доли верно распознанных документов.

Ключевые слова: иерархическая кластеризация, кластерный анализ, классификация, полисемия, DBSCAN.

Введение

В настоящее время в связи с непрекращающимся ростом количества информации становится всё более актуальной задачи, связанные с обработкой естественного языка. В рамках данной работы мы не будем затрагивать задачи, связанные с обработкой речи, а остановимся только на обработке текстовой информации на естественном языке.

Важнейшим этапом решения таких задач является предварительная обработка текстовых данных. Идея данного этапа заключается в приведении текстового документа к единому формату, поиске термов (например, поиск морфем, именованных сущностей или устойчивых словосочетаний) и разрешении кореференций (определения, к каким частям текста относятся те или иные слова и обороты).

На данный момент не существует единого формата текстового документа, который мог бы эффективно использоваться для целого класса задач обработки текстовых документов. Существует ряд эмпирически выявленных рекомендаций, которые позволяют повысить качество обработки текстовых документов, например, снятие лексической многозначности [11, с. 74; 3, с. 382; 5, с. 27]. Такое решение позволяет отделить омонимичные слова друг от друга, что особенно важно для построения моделей языка и задач классификации на основе семантической близости.

В предыдущей работе [8, с. 1862] мы рассматривали кластеризацию контекстов как метод разрешения полисемии в задаче классификации текстовых документов. Однако ограничение размерности класса лежало полностью на разработчике, либо регулировалось вручную. В рамках данной работы мы рассмотрим метод автома-

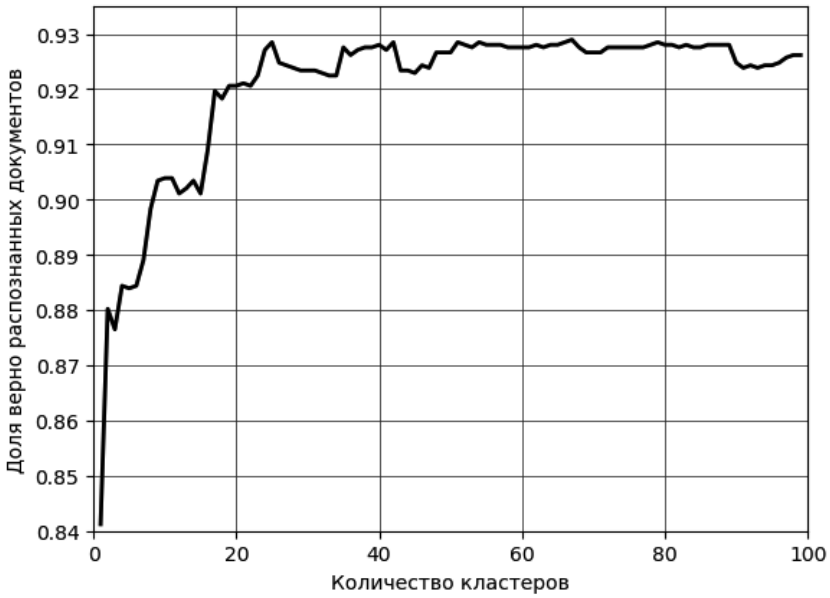

Рис. 1. Доля верно распознанных документов в зависимости от количества кластеров

тического определения размерности класса и сделаем оценку качества классификатора в зависимости от полученных кластеров.

Идентификация и обработка многозначных слов

Можно заметить, что большинство омонимичных слов имеют иерархическую структуру, например, более широкий контекст может быть уточнен различными узкими контекстами. Так, например, в предыдущей работе [8, с. 1864] мы рассматривали слово «звезда», которое имеет несколько широких контекстов, один из них — это «популярная личность», который в свою очередь разделяется на несколько подмножеств: «популярная личность в спорте», «популярная личность в кино», и т.д.

Исходя из данной структуры мы использовали агломеративную иерархическую кластеризацию [14, с. 58; 6, с. 24; 4, с. 41; 1, с. 115] с фиксированный количеством кластеров. Параметрами такой модели являлись: ширина окна, весовая функция, минимальная частота термов и размерность вектора. В качестве метода объединения точек для поиска ближайших кластеров был выбран Complete linkage — максимум попарных расстояний между точками из двух кластеров:

$$\max\{d(a,b): a \in A, b \in B\} \quad (1)$$

Расстояние между точками определяется по формуле косинусной близости:

$$sim = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (2)$$

Для того чтобы проанализировать влияние полисемии на качество классификации, необходимо выбрать некоторое множество документов, с которым мы будем работать. В качестве исходных данных возьмем информационные сообщения одного из крупнейших в мире международных агентств новостей и финансовой информации Reuters.

В качестве классификатора мы выбрали наивный байесовский классификатор, который можно использовать в режиме реального времени, благодаря простоте его реализации и скорости работы [13, с. 5; 2, с. 30; 15, с. 7]. В одной из предыдущих работ мы рассматривали проблему нулевой вероятности данного алгоритма и способы её решения [2, с. 30], в рамках данной работы мы будем брать классическую реализацию данного классификатора (модуль MultinomialNB библиотеки scikit-learn языка Python).

Для упрощения будем применять операцию снятия омонимии с помощью кластеризации контекстов только для одного омонимичного слова, например, для слова «пост». Отберем для классификатора только те документы, в которых присутствует данное омонимичное слово. Результаты классификации в зависимости от количества кластеров представлены на рис. 1.

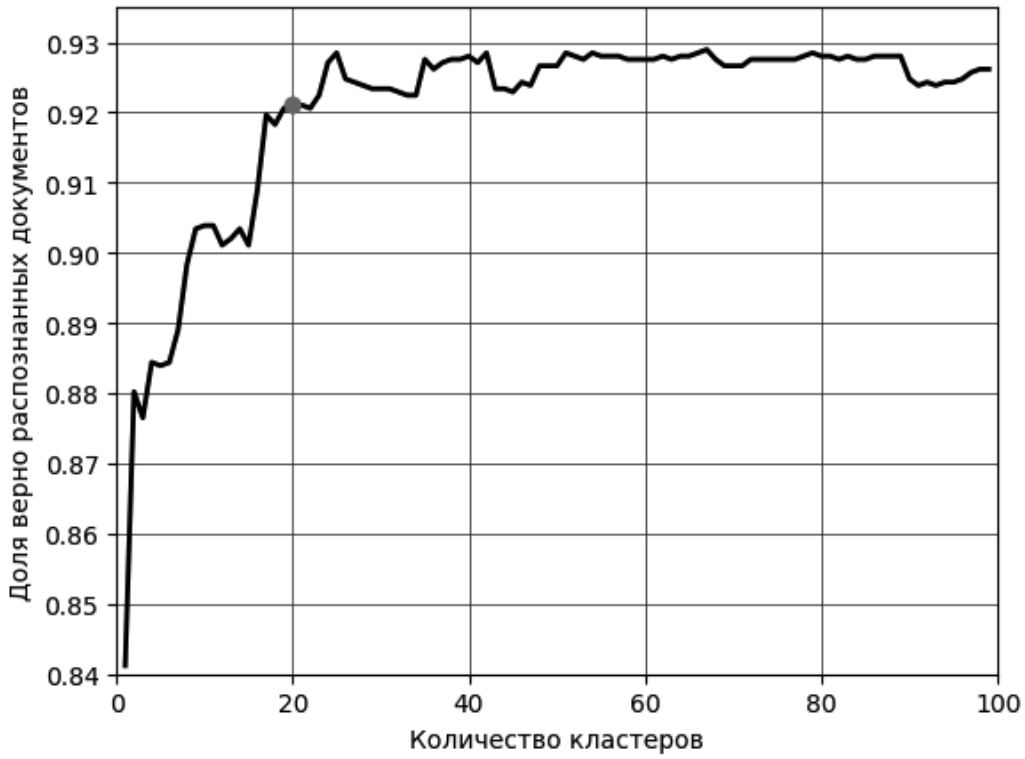

Рис. 2. Снятие омонимии для слова «пост».

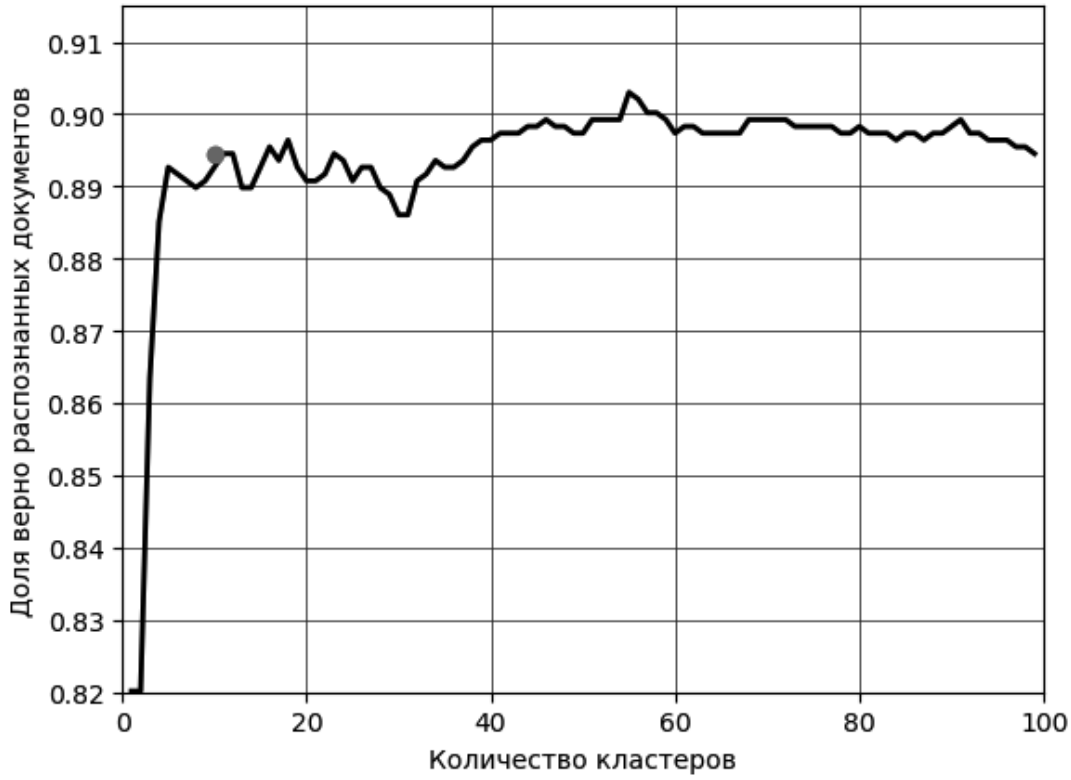

Рис. 3. Снятие омонимии для слова «звезда».

Доля верно распознанных документов:

- без кластеризации = 0,84;
- при разделении омонимичного слова на два кластера = 0,88;
- при разделении омонимичного слова на 66 контекстов = 0,928 (глобальный максимум).

Для иерархической кластеризации характерен визуальный анализ дендрограммы и определение по ней наиболее предпочтительного числа кластеров. Однако для автоматического определения количества кластеров необходимо использовать другие алгоритмы. В данной работе мы будем использовать один из наиболее популярных алгоритмов — DBSCAN. Он является плотностным алгоритмом для кластеризации пространственных данных с присутствием шума и позволяет разбивать данные на кластеры произвольной формы [7, с. 227; 9, с. 209; 10, с. 20]. На рис. 2 представлен результат определения количества кластеров на основе алгоритма DBSCAN.

Количество кластеров = 20.

Доля верно распознанных документов = 0,921.

Возьмем омонимичное слово «звезда» и повторим эксперимент. На рис. 3 представлен результат определения количества кластеров на основе алгоритма DBSCAN для слова «звезда».

Можно заметить, что автоматическое определение количества кластеров в задаче классификации текстовых документов хорошо находит локальный максимум, однако не гарантируется попадание в точку глобального максимума [12, с. 23]. Одной из причин является эвристический подход к заданию начальных параметров.

Заключение

В настоящей работы мы рассмотрели важную прикладную задачу классификации текстовых документов. С помощью автоматического снятия омонимии на основе кластеризации контекстов нам удалось повысить качество классификации. Однако перед нами стояла проблема в автоматическом определении количества кластеров. В ходе работы на основе плотностного алгоритма DBSCAN нам удалось вычислить количество кластеров, далее на основе агломеративной иерархической кластеризации разбить омонимичные контексты на кластеры и снять омонимию. После того как сняли омонимию, мы проверили качество классификации на основе алгоритма наивного байесовского классификатора. В результате среднее значение доли распознанных документов увеличилось на 7%. Также можно заметить следующие свойства:

- «бесконечное» увеличение контекстов не повышает качество классификации, а наоборот, приводит к его снижению;
- разбиение на небольшое количество кластеров даёт ощутимое повышение качества классификации.

Повышение качества классификации текстовых документов позволит эффективнее решать прикладные задачи компьютерной лингвистики, например, такие как: машинный перевод, информационный поиск, составление тематических каталогов и пр.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19–31–27001.

Acknowledgments: The reported study was funded by RFBR, project number 19–31–27001.

ЛИТЕРА**Т**УРА

1. Дударин П.В., Ярушкина Н. Г. Алгоритм построения иерархического классификатора коротких текстовых фрагментов на основе кластеризации нечеткого графа // Радиотехника. 2017. № 6.
2. Капитанов А. И. Проблема аддитивного сглаживания вероятностей наивного байесовского классификатора // Актуальные проблемы информатизации в науке и образовании — 2017. 10-я Всероссийская межвузовская научно-практическая конференция: тезисы докладов. М.: МИЭТ. 2017.
3. Кузнецов И. П. Организация семантико-ориентированных систем поиска и обработки информации // Системы и средства информ. 2006.
4. Ломакина Л.С., Родионов В. Б., Суркова А. С. Иерархическая кластеризация текстовых документов. // Системы управления и информационные технологии. 2012. № 2(48).
5. Марчук Ю. Н. Контекстное разрешение лексической многозначности. Вестник МГОУ. Серия: Лингвистика. 2016. № 1.
6. Яцкив И., Гусарова Л. Методы определения количества кластеров при классификации без обучения // Transport and Telecommunication. 2003. № 1.
7. A density-based algorithm for discovering clusters in large spatial database / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // Proc. 1996 Intern. Conf. on Knowledge Discovery and Data Mining. 1996.
8. A. Kapitanov, I. Kapitanova, V. Troyanovskiy, V. Ilyushechkin and E. Dorogova, "Clustering of Word Contexts as a Method of Eliminating Polysemy of Words," 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus). 2019.
9. Birant D., Kut A. ST-DBSCAN: an algorithm for clustering spatial-temporal data. Data & Knowledge Engineering. 2007. Vol. 60.

10. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN / E. Schubert, J. Sander, M. Ester, H. P. Kriegel, X. Xu // ACM Trans. Database Syst. 2017. Vol. 42.
11. Krovetz R. Homonymy and Polysemy in Information Retrieval. // In Proc. of EACL '97 Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics archive. 1997.
12. Kruzhilin, S., Baranova, T., Mishenina, M., & Zaitseva, M. (2018). Regional specificity creation of protective afforestations along highways. World Ecology Journal, 8(2). https://doi.org/https://doi.org/10.25726/NM.2018.2.2.003
13. Lewis D. D. Naive (Bayes) at forty: The independence assumption in information retrieval // Proceedings of 10th European Conference on Machine Learning. 1998.
14. Revelle, W. Hierarchical Cluster Analysis and the Internal Structure of Tests // Multivariate Behavioral Research. 1979. Vol. 14.
15. Webb G.I., Boughton J. R., Wang Z. Not So Naive Bayes: Aggregating One-Dependence Estimators // Machine Learning. Springer. 2005. № 58.
16. Zelenyak, A., & Kostyukov, S. (2018). Features of the development of architectonics of crowns of bushes as a criterion of decorativeness in green building. World Ecology Journal, 8(3). https://doi.org/https://doi.org/10.25726/NM.2019.99.51.001

МИЭТ