Сравнительный анализ методов кластеризации текстовой информации

П. В. Соколов[1], Е. Н. Каруна[2]
Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И. Ульянова (Ленина)
[1]pvsokolov@etu.ru, [2]enkaruna@etu.ru

***Аннотация*. Работа посвящена сравнительному анализу методов кластеризации текстовой информации с использованием набора H-метрик. Исследованы: самоорганизующаяся сеть Кохонена, методы K-средних, спектральной кластеризации и агломеративной кластеризации. Рассмотрены особенности реализации алгоритмов и вопросы их оптимизации. Для наглядной визуализации результатов, применен метод снижения размерности, позволяющий на двумерной плоскости отобразить многомерное пространство данных. Сделаны выводы об эффективности исследованных алгоритмов.**

Ключевые слова: кластеризация; нейронные сети; машинное обучение; тематический анализ

I. Введение

Одним из актуальных направлений в области обработки естественного языка является кластерный анализ текстовой информации. Эта задача не теряет своей актуальности по ряду причин:

- непрерывно возрастающий объем данных в сети интернет и различных закрытых базах данных вызывает трудности в анализе и необходимость постоянно улучшать и оптимизировать текущие системы кластеризации;
- рост вычислительных мощностей, позволяющий применять новые алгоритмы и ускоряющий процесс исследований;
- появление новых онлайн-сервисов, требующих выполнять анализ большого количества текстовых данных для работы с живыми людьми;
- существующая нехватка точности работы систем кластерного анализа текстовых данных и способов оценки результатов работы систем;
- многообразие имеющихся алгоритмов и отсутствие универсального решения.

Алгоритмы кластерного анализа находят своё применение в различных сферах. Так, они применяются в службах поддержки клиентов для структурирования и группировки их заявок и жалоб. Кластеризация новостного потока выполняется с целью разбиения данных на схожие подгруппы новостей и дальнейшем предоставлении более релевантных данных. Также, одной из крупных областей применения являются сервисы поиска текстовой информации. Для них необходимо огромный объем данных, хранящийся в открытом доступе в сети или закрытых текстовых базах данных, разбить на категории, помогающие пользователю отбросить наименее релевантные группы данных. Это позволяет значительно ускорить процесс поиска необходимой информации.

Одной из проблем при выполнении исследований в области кластеризации текстовых данных, является наличие слишком большого количества алгоритмов машинного обучения и сложность в настройке параметров этих алгоритмов. В данной работе рассматривается задача выделения группы наиболее подходящих алгоритмов машинного обучения и поиска оптимальных параметров алгоритмов машинного обучения для выполнения кластеризации текстовой информации на основе русскоязычного корпуса данных. Производится сравнительный анализ выбранных алгоритмов.

II. Этап предварительной обработки исходных данных

Решение многих задач в области обработки естественного языка сводится к выполнению ряда стандартных процедур над текстом. Эти процедуры позволяют преобразовать текст к виду, удобному для дальнейшего анализа, одним из множества алгоритмов машинного обучения. Существующие подходы к решению этих задач подробно описаны в статье [1].

Один из важнейших подходов к анализу текстовой информации, относящейся к задаче машинного обучения без учителя, связан с кластеризацией текстовых данных. Его идея состоит в том, что определённую выборку данных, в которой каждый объект описывается набором признаков, необходимо разбить на несколько групп объектов. Все объекты внутри одной группы должны обладать некоторым набором общих свойств или закономерностей.

По ряду признаков задача кластеризации имеет много общего с задачей классификации текстовой информации. Основное сходство заключаются в итогах работы алгоритмов машинного обучения. В обоих случаях, результатом является разбиение исходной выборки текстовых данных на подгруппы. Главное различие состоит в применении обучающей выборки данных в алгоритме классификации.

Этапы предварительной обработки исходных текстовых данных для алгоритмов классификации и

кластеризации совпадают между собой. Главной целью является выполнение преобразования данных в набор числовых векторов, которые можно подавать на вход алгоритмам машинного обучения. Некоторые подробности этапов предварительной обработки текстовых данных описаны в [1][2].

Наиболее распространённым методом получения векторов текстов является модель, именуемая «мешок слов» (от англ. bag-of-words). В этой модели весь текст представлен одним вектором, хранящим информацию о количественном составе каждого слова в тексте. Недостатком данного подхода является отсутствие семантических связей между словами в тексте. Частично эта проблема решается за счет использования *n*-грамм, которые добавляют в вектор информацию о количественном составе сочетаний по *n* слов.

Перед началом формирования векторов, как правило, применяется операция фильтрации текстов, необходимая для удаления различных небуквенных символов из текста и общеупотребительных слов. Удаляемые общеупотребительные слова обладают большой частотой появления в текстах, но несут слишком мало информации о принадлежности текста к одной из тематик.

III. Алгоритмы кластеризации и метрики оценки

A. *Алгоритмы кластерного анализа*

Формальная постановка задачи кластеризации выглядит следующим образом.

Введем следующие обозначения: $X = \{x_1, \dots, x_n\}$ – тестовая выборка данных, $Y = \{y_1, \dots, y_m\}$ – множество кластеров (меток), $p(x, x')$ – функция расстояния между примерами из тестовой выборки данных. Тогда алгоритм кластеризации – это отображение $a: X \to Y$. Иными словами, это функция, ставящая объект множества X в соответствие с меткой кластера из множества Y. Причём все объекты одного кластера должны быть близки в заданном смысле друг к другу, а различных кластеров должны существенно различаться.

Характерной особенностью задачи кластерного анализа текстовых данных, затрудняющей выполнение машинного обучения, является наличие данных очень высокой размерности. Этому способствует ряд причин:

- в таких данных разница в расстоянии между двумя любыми точками будет минимальна и многие алгоритмы, использующие различные метрики на основе расстояния между объектами, могут потерять свою эффективность;
- многие признаки могут быть незначительны;
- требуется больше ресурсов на выполнение вычислений и хранение промежуточных результатов.

Помимо перечисленных особенностей, векторы данных, полученные на основе частоты встречаемости слов, характеризуются высокой разреженностью матрицы признаков. Следовательно, возникает необходимость делать вектора большей размерности, чтобы получить достаточное количество признаков для выполнения задачи кластеризации. Более детально различные подходы к кластеризации раскрываются в источниках [3][4].

Наиболее широко применимый подход к кластерному анализу это алгоритм К-средних. Этот алгоритм выполняет случайную инициализацию центров кластеров и определяет ближайшие вектора данных к центру каждого кластера. Далее, выполняется итеративное смещение центров кластеров и пересчёт ближайших векторов данных. В конечном итоге алгоритм минимизирует суммарное квадратичное отклонение данных от центров полученных кластеров.

Помимо алгоритма К-средних, проведенное в данной работе исследование показало достаточно хорошие результаты для алгоритмов спектральной кластеризации, агломеративной кластеризации и самоорганизующаяся карта Кохонена. Результаты работы исследованных алгоритмов будут рассмотрены в следующих разделах.

B. *Метрики качества*

Одной из проблем кластерного анализа является отсутствие метода, позволяющего точно выполнять оценку качества работы алгоритма кластеризации. Есть две группы методов для оценки моделей: внешние, позволяющие выполнять сравнение результатов кластеризации с заранее промаркированными классами и внутренние, которые используют только ту информацию, которая хранится в самих данных. В данной работе используются внешние методы оценки.

Внешняя метрика качества должна оценить степень соответствия маркировки объектов, полученной заранее, той маркировке, которая была получена в результате работы алгоритма кластеризации.

В данной работе, при оценке качества алгоритмов кластерного анализа, используется два вида оценок: V-мера и скорректированный индекс Ранда.

V-мера является средним гармоническим значением между двумя другими метриками: однородностью и полнотой. Однородность характеризует степень соответствия тому, что каждый кластер содержит только члены одного класса. Полнота показывает, в какой мере все члены данного класса собраны в одном кластере. Эти величины изменяются в диапазоне от 0 до 1.

Скорректированный индекс Ранда – это метрика качества, построенная на основе метрики, вычисляющей сходство между двумя кластерами, подсчитывая пары элементов, которые находятся в одном классе и в разных классах, с поправкой на случайность. Более подробно существующие подходы описываются в работе [5].

IV. Разработка и выполнение исследований

A. *Используемые инструменты разработки*

В данной работе, в качестве корпуса текстов, использована коллекция новостных статей на русском языке, которые разделены на отдельные категории по тематикам. Коллекцией текстов является находящийся в

открытом доступе архив новостных статей Lenta.ru, содержащим более 700 тысяч текстов.

В качестве инструмента для фильтрации текстов был разработан парсер текстов, способный, помимо фильтрации текста от лишних символов, выполнять операцию нахождения основы слова. Парсер использует для этого два разных алгоритма: стемминга и лемматизации. Составление векторов текстов в виде модели «мешка слов» выполняется набором инструментов из библиотеки scikit-learn.

Построенные вектора текстов должны иметь большую размерность, т.к. длина векторов зависит от длины набора слов, которые будут выбраны в качестве основных признаков. Для визуализации полученного пространства признаков использовался алгоритм t-SNE, который позволяет выполнять визуализацию элементов многомерного пространства путём снижения размерности до двух. Полученное двумерное отображение пространства может использоваться для наглядной визуализации результатов исследований. На рис. 1 изображена визуализация двумерного отображения пространства данных с предопределенной принадлежностью текстов к классам (идеальный случай), где каждая точка характеризует расположение вектора текста в этом пространстве. Для визуальной наглядности, каждый объект был помечен цветом своего класса.

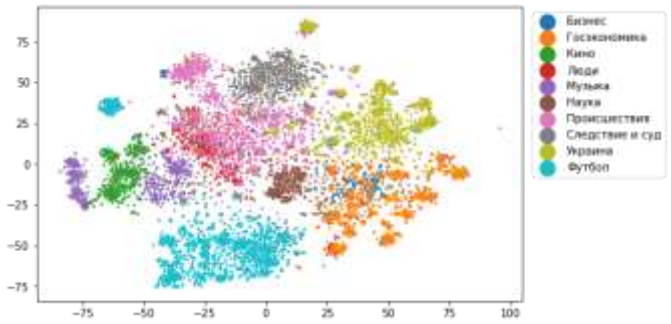

Рис. 1. Результат работы алгоритма снижения размерности t-SNE

Алгоритмы машинного обучения, использованные в работе, требуют точной настройки большого количества параметров. Далеко не всегда очевидно, как следует выполнять настройку параметров, поэтому целесообразно применить алгоритм оптимизации, перебирающий в многопоточном режиме большое количество параметров. Алгоритм оценивает результаты работы по каждому набору возможных параметров, внутри заданных диапазонов, с помощью одной из метрик качества, которые описывались выше.

Наиболее очевидным способом выполнения оптимизации, но не самым эффективным с точки зрения затраченных ресурсов, является простой перебор по сетке параметров. Он позволяет задавать диапазон и шаг изменения всех возможных параметров. Данный метод оптимизации дает возможность определить группу параметров, которая демонстрирует наилучший результат по выбранной метрике качества.

B. Полученные результаты исследований

В ходе исследований требовалось выполнить сравнение результатов работы алгоритмов машинного обучения и выявление оптимального набора параметров алгоритмов, показывающих наилучшую оценку по выбранной метрике данных.

Первым этапом были выполнены исследования по сравнению алгоритмов кластерного анализа при выборке данных в 10 тысяч текстов и размере словаря на 1000 элементов. Результаты сравнения точности кластеризации по двум видам метрик и при четырёх способах предобработки данных представлены в табл. 1.

ТАБЛИЦА I РЕЗУЛЬТАТЫ КЛАСТЕРИЗАЦИИ

Название алгоритма	**Stem-(1-3)**		**Stem-(1-1)**		**Lemma-(1-3)**		**Lemma-(1-1)**	
	v	***ARI***	***v***	***ARI***	***v***	***ARI***	***v***	***ARI***
k-means	0.67	0.57	0.63	0.51	0.61	0.50	0.58	0.45
SOM	0.61	0.52	0.65	0.58	0.63	0.57	0.62	0.58
Spectral	0.65	0.53	0.62	0.51	0.62	0.53	0.61	0.51
Agglomerat	0.56	0.47	0.56	0.48	0.56	0.46	0.55	0.40

В табл. 1 слово Stem – означает выполнение операции стемминга для нахождения основы слова, а Lemma – операцию лемматизации. Значения в скобках означают максимальную и минимальную длину последовательности *n*-граммы в словаре. Столбец с полем *v* обозначает оценку по V-мере, а столбец с полем *ARI* обозначает оценку по скорректированному индексу Ранда.

Как можно определить из результатов, алгоритмы показывают достаточно близкую точность работы между собой, но, в среднем, самоорганизующиеся карты Кохонена оказываются лучше других.

На рис. 2 изображено отображение точек данных на двумерное пространство, где каждая точка выделена в соответствии с тем кластером, которое выбрал алгоритм кластеризации. На данном рисунке показан результат работы алгоритма самоорганизующихся карт Кохонена. Можно заметить, что в некоторых местах построенного изображения кластеры были выбраны достаточно точно.

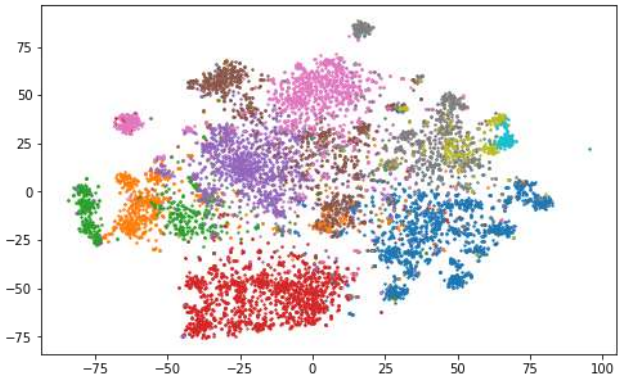

Рис. 2. Визуализация построения кластеров алгоритмом SOM

В результате работы алгоритма оптимизации были выявлены следующие оптимальные наборы параметров для некоторых алгоритмов:

- алгоритм К-средних: максимальное число итераций *max_iter* – (30-50); количество запусков алгоритма с обновлёнными значениями центров кластеров *n_init* – 5;
- самоорганизующиеся карты: скорость обучения *learning_rate* – 0.1; число итераций – 2000; функция соседства *h* – *Gaussian*;
- спектральная кластеризация: способ построения матрицы подобия – *rbf*; алгоритм для кластеризации данных по матрице подобия – *discretize*; коэффициент ядра *gamma* – 0.1.

Параметры, которые не были представлены в заданном наборе, вероятно, не удалось однозначно определить даже в результате множества запусков алгоритма оптимизации.

Следующим этапом были получены данные с теми же условиями, но с уменьшенным количеством объектов в выборке данных. Результаты для выборки размером 5000 текстов представлены в табл. 2, для выборки размером 1000 текстов в табл. 3.

ТАБЛИЦА II РЕЗУЛЬТАТЫ КЛАСТЕРИЗАЦИИ ДЛЯ ВЫБОРКИ РАЗМЕРОМ 5000 ТЕКСТОВ

Название алгоритма	**Stem-(1-3)**		**Stem-(1-1)**		**Lemma-(1-3)**		**Lemma-(1-1)**	
	v	***ARI***	***v***	***ARI***	***v***	***ARI***	***v***	***ARI***
k-means	0.67	0.55	0.64	0.55	0.64	0.55	0.64	0.52
SOM	0.62	0.52	0.63	0.52	0.63	0.52	0.62	0.50
Spectral	0.60	0.49	0.60	0.48	0.60	0.48	0.60	0.48
Agglomerat.	0.55	0.41	0.59	0.52	0.59	0.52	0.59	0.51

ТАБЛИЦА III РЕЗУЛЬТАТЫ КЛАСТЕРИЗАЦИИ ДЛЯ ВЫБОРКИ РАЗМЕРОМ 1000 ТЕКСТОВ

Название алгоритма	**Stem-(1-3)**		**Stem-(1-1)**		**Lemma-(1-3)**		**Lemma-(1-1)**	
	v	***ARI***	***v***	***ARI***	***v***	***ARI***	***v***	***ARI***
k-means	0.63	0.47	0.63	0.48	0.65	0.49	0.62	0.47
SOM	0.66	0.51	0.67	0.56	0.67	0.52	0.67	0.53
Spectral	0.63	0.49	0.63	0.50	0.63	0.49	0.64	0.51
Agglomerat.	0.67	0.59	0.62	0.49	0.65	0.60	0.63	0.56

По полученным данным видно, что уменьшение количества данных не оказывает достаточно значительного влияния на качество кластеризации. Более того, агломеративная кластеризация с уменьшением числа данных достигает больших результатов.

Для выполнения дополнительного поиска наиболее подходящих параметров предобработки данных были выполнены дополнительные исследования, результаты которых приведены в табл. 4. Здесь «Stem-(2-3)-1000» означает, что выполняется операция стемминга для определения основы слова, длина последовательности *n*-граммы от 2 до 3 и размерность вектора признаков 1000. Обозначения для других столбцов сделаны аналогичным образом.

ТАБЛИЦА IV РЕЗУЛЬТАТЫ КЛАСТЕРИЗАЦИИ С НОВЫМИ ПАРАМЕТРАМИ ПРЕДОБРАБОТКИ ДАННЫХ

Название алгоритма	**Stem-(2-3)-1000**		**Stem-(2-2)-1000**		**Stem-(1-1)-300**	
	v	***ARI***	***v***	***ARI***	***v***	***ARI***
k-means	0.340	0.180	0.376	0.219	0.555	0.425
SOM	0.338	0.221	0.384	0.318	0.555	0.493
Spectral	0.367	0.273	0.413	0.203	0.518	0.382
Agglomerat.	0.369	0.238	0.363	0.275	0.479	0.310

Из табл. 4 видно, что качество кластеризации при использовании размера последовательности *n*-граммы (2-3) или (2-2) значительно ухудшается. Также, по третьему набору параметров предобработки можно заметить небольшое изменение качества кластеризации при уменьшении размера векторов признаков.

V. ЗАКЛЮЧЕНИЕ

По результатам выполненных исследований была разработана система, выполняющая полный цикл работ для выполнения кластерного анализа данных. Полученные результаты работы демонстрируют ряд закономерностей, которые могут помочь дальнейшим исследованиям систем кластерного анализа текстовой информации:

- с помощью алгоритма оптимизации по сетке параметров были определены диапазоны настройки некоторых параметров алгоритмов машинного обучения;
- по результатам исследований были определены некоторые основные параметры предобработки данных, которые позволяют выполнять алгоритмы кластеризации с большей точностью;
- подробно описан состав программных средств, который использовался при разработке алгоритма, позволяющий точнее понять условия, при которых выполнялись исследования и учитывать их при дальнейшей разработке.

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, no. 7, pp. 34-47.

[2] Каруна Е.Н., Соколов П.В. Нейросетевой классификатор текстовой информации // Известия СПбГЭТУ "ЛЭТИ", 2020, №8-9, сс. 66-71.

[3] Rodriguez M.Z., Comin C.H., Casanova D., Bruno O.M., Amancio D.R., Costa L., Rodrigues F.A. Clustering algorithms: A comparative approach. PLoS ONE, 2019, no. 14(1). Available at: https://doi.org/10.1371/journal.pone.0210236 (Accessed at 28 February 2021)

[4] Tomi Kinnunen, Ilja Sidoroff, Marko Tuononen. Comparison of Clustering Methods: a Case Study of Text-Independent Speaker Modeling. Pattern Recognition Letters, no. 32, pp. 1604-1617. DOI: 10.1016/j.patrec.2011.06.023.

[5] Julio-Omar Palacio-Nino, Fernando Berzal. Evaluation Metrics for Unsupervised Learning Algorithms. Cornell University, 2019. Available at: https://arxiv.org/abs/1905.05667 (Accessed at 28 February 2021)