

UDC 519.67

INFORMATION CRITERION FOR THE CATEGORIZATION QUALITY EVALUATION

¹ Yulia E. Balykina² Michail V. Svirkin

¹ Saint-Petersburg State University
7/9, Universitetskaya nab., St.Petersburg, 199034
assistant

² Saint-Petersburg State University
7/9, Universitetskaya nab., St.Petersburg, 199034
PhD, associated professor
E-mail: julia.balykina@gmail.com

The paper considers the possibility of using the variation of information function as a quality criterion for categorizing a collection of documents. The performance of the variation of information function is being examined subject to the number of categories and the sample volume of the test document collection.

Keywords: texts categorization, quality criterion, variation of information function, sample size.

Введение. Кластеризация, или нахождение разбиения множества объектов на группы, является одной из основных задач в области анализа данных [1, 2]. При этом, кластеризация массивов текстовой информации является относительно новой сферой исследований, где еще нельзя говорить о существовании точных и полных методов и алгоритмов. Особенно это касается обоснования эмпирической оценки работы алгоритмов кластеризации путем сравнения результатов с заданным «эталонным» разбиением, когда необходимо определить «расстояние» в пространстве разбиений данных.

Проблема качества кластеризации начала обсуждаться в 1950-х годах прошлого века. В данной работе рассматривается возможность использования функции вариации информации в качестве критерия качества категоризации коллекции текстов. Исследуется поведение функции вариации информации в зависимости от объема обучающей выборки и количества категорий. Проверяется гипотеза о существовании количественной оценки возможного минимального объема репрезентативной выборки текстов для качественной категоризации с точки зрения информационного критерия качества разбиения.

Среди наиболее часто используемых классических критериев для сравнения разбиений отметим следующие: Rand index, Jaccard index, Folwkes and Mallows index, Mirkin metric, Van Dongen metric и т.д. Однако, все эти критерии учитывают только формальную разницу в разбиениях, основываясь на статистических численных критериях. Например, при использовании критерия Миркина оценивается число пар объектов, попавших в один и тот же кластер при разбиении C , но в разные кластеры – при разбиении C' .

Функция вариации информации. В работе [3] была предложена функция, названная вариацией информации (variation of information, VI). Была показана эффективность ее использования при сравнении различных разбиений одного множества. В основе функции VI лежат понятия энтропии и информации. Данный

критерий определяет изменение количества информации при переходе от разбиения C к разбиению C' .

Рассмотрим множество текстов D и разбиение этого множества на K непересекающихся подмножеств, называемых кластерами:

$$\{C_1, C_2, \dots, C_K\}, \text{ при этом } C_k \cap C_l = \emptyset \text{ и } \bigcup_{k=1}^K C_k = D.$$

Пусть имеется $C' = \{C'_1, C'_2, \dots, C'_{K'}\}$ - другое разбиение множества D . Отметим, что два разбиения могут иметь различное число кластеров.

При рассмотрении вариации информации мы основываемся на том, насколько информативно каждое из разбиений, и сколько информации об одном разбиении содержится в другом. Тогда в качестве критерия для сравнения двух разбиений C и C' воспользуемся функцией вариации информации

$$V(C, C') = H(C) + H(C') - 2I(C, C'),$$

где $H(C)$, $H(C')$ - энтропии на множествах разбиений C и C' соответственно, $I(C, C')$ - взаимная информация между разбиениями C и C' .

Ниже мы проанализируем, как изменяются результаты кластеризации текстовой коллекции документов в зависимости от объема репрезентативной выборки при использовании в качестве критерия качества функцию вариации информации. В качестве базового алгоритма кластеризации будем рассматривать алгоритм k -средних.

Вычислительный эксперимент. Для решения поставленных задач был проведен ряд экспериментов по категоризации коллекции текстовых документов на основе частоты вхождений в текст различных термов, взятых из общего словаря текстовой коллекции. В качестве тестовой коллекции документов для проведения исследования взята коллекция 20NewsGroups [4].

Номера категорий соответствуют следующим папкам из каталога 20 Newsgroups:

- 1 - alt.atheism (атеизм);
- 2 - comp.graphics (компьютерная графика);
- ...
- 8 - rec.autos (автомобили);
- 9 - rec.motorcycles (мотоциклы);
- 10 - rec.sport.baseball (баскетбол);
- 11 - rec.sport.hockey (хоккей);
- 12 - sci.crypt (криптография).

Векторы, соответствующие документам d_i ($i = \overline{1, N}$), представлены частотными характеристиками документов: каждому документу d_i , $i = \overline{1, N}$ соответствует вектор $n(w_j, d_i)$, $j = \overline{1, M}$, где $n(w_j, d_i)$ - частота встречаемости слова w_j в документе d_i . Все документы были разбиты на термы, при этом была произведена предобработка коллекции: убраны знаки препинания и повторяющиеся поля в начале каждого текстового файла. Также, из всей коллекции были изъяты так называемые стоп-слова, такие как неопределенные артикли, наречия, союзы, предлоги и т.д. После таких преобразований объем словаря коллекции составил 112966 слов.

Для решения поставленной задачи было разработано программное обеспечение на языке C# на основе базы данных MS SQL Server 2008. Для работы с большими массивами переменной длины были использованы хэш-структуры, что позволило избежать копирования элементов при увеличении длины массива и, как

следствие, существенно ускорить процесс вычислений. Этот метод весьма эффективен, поскольку как время размещения элемента в таблице, так и время его поиска определяются только временем, затрачиваемым на вычисление хэш-функции, которое в общем случае несопоставимо меньше времени, необходимого на многократные сравнения элементов таблицы.

Для оценки качества разбиения при выборках различного объема проводилось сравнение результатов с уже известным «эталонным» разбиением с использованием функции качества $VI(C, C')$. Кластеризация проводилась на выборках различного объема (50, 100, 200, 250, 300, 400, 500, 600 и 700 документов из каждой категории) и для разного числа кластеров (2, 3, 5, 7, 10). Для каждого случая проводилось пять экспериментов, в качестве результата бралось среднее значение функции вариации информации.

В качестве выбора начальных центров выбирались средние значения векторов, вычисленные по формуле $\mu_l = \frac{1}{K} \sum_{i=1}^K \mu_{li}$, $l=1, \dots, k$. Здесь μ_{\max} - вектор максимальных значений $n(w_j, d_i)$, $j=1, M$, а μ_{\min} - вектор минимальных значений $n(w_j, d_i)$, $j=1, M$, $i=1, N$.

Используя созданное программное обеспечение для анализа поведения функции $VI(C, C')$ в зависимости от объема выборки и количества категорий, были получены результаты, графическое отображение которых представлено на рис. 1-3.

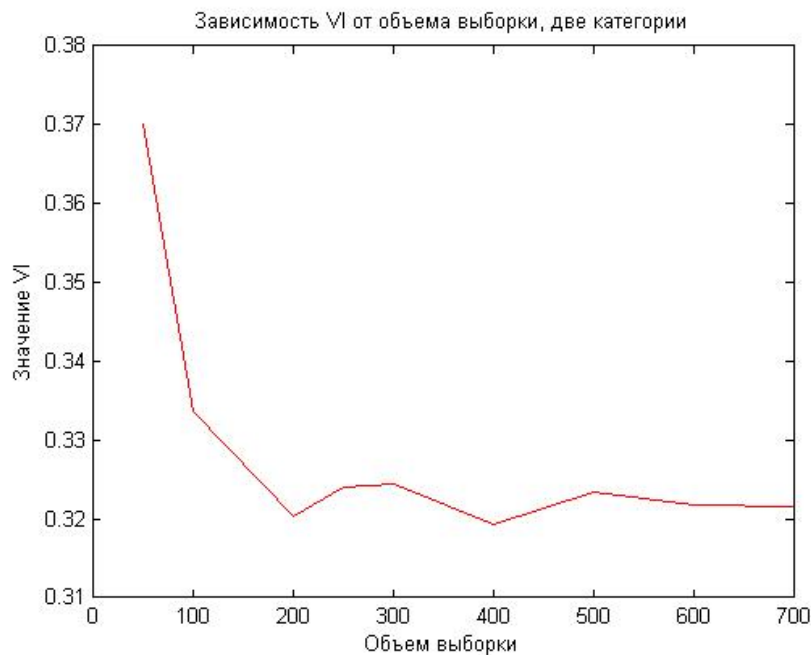


Рис. 1. Категории 1-2

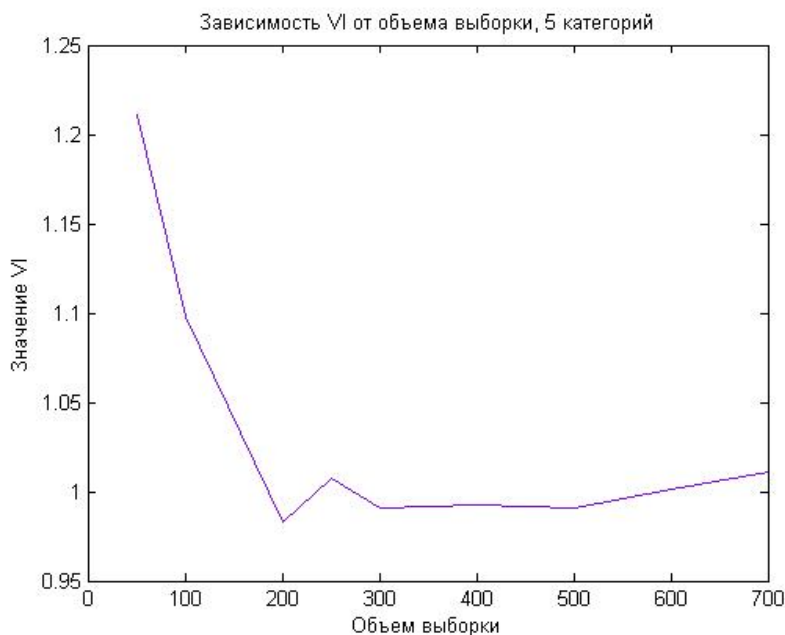


Рис. 2. Категории 8-12

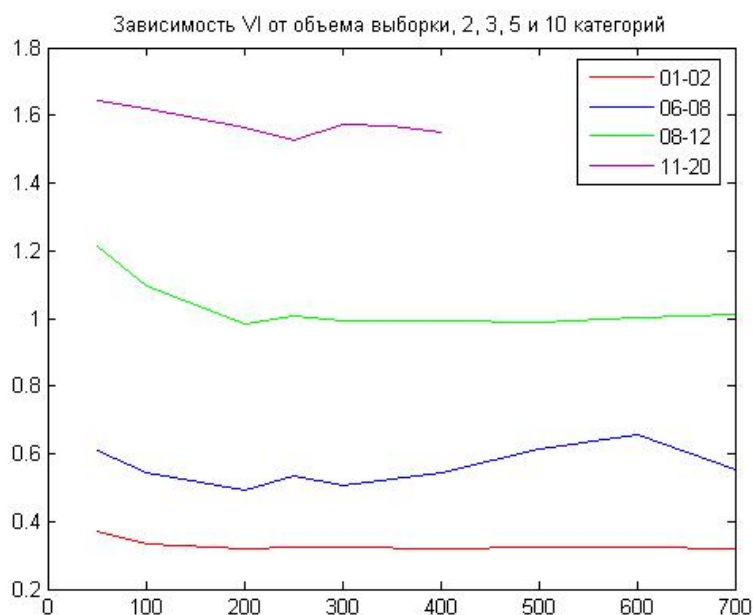


Рис. 3. Значение функции вариации информации в зависимости от объема выборки и количества категорий

Результаты обработки экспериментальных данных иллюстрируют связь между минимальным объемом репрезентативной выборки и значением функции вариации информации $VI(C, C')$, а также позволяют оценить возможность экстраполяции полученных результатов на коллекции документов большего объема. Исследование показало, что с ростом числа категорий в эталонном разбиении значение информационного расстояния увеличивается. При этом значение функции $VI(C, C')$ сначала уменьшается с ростом объема выборки, достигая минимума, после чего начинает снова возрастать. Минимальному значению $VI(C, C')$ соответствует

возможный минимальный объем выборки текстов, которая включает в себя весь набор термов, характерных для эталонного разбиения.

Примечания:

1. Айвазян С.А. Прикладная статистика в задачах и упражнениях / С.А. Айвазян, В.С. Мхитарян. М. : ЮНИТИ-ДАНА, 2001. 270 с.
2. Thomas M. Cover. Elements of Information Theory. / Thomas M. Cover, Joy A. Thomas. Willey, 2006. 776 p.
3. Marina Meilă. Comparing clusterings – an information based distance. / Journal of Multivariate Analysis, Vol. 98, Issue 5, May 2007, P. 873-895.
4. Homepage for 20 Newsgroups Data Set. URL: <http://people.csail.mit.edu/jrennie/20Newsgroups/>.

УДК 519.67

**ИНФОРМАЦИОННЫЙ КРИТЕРИЙ ОЦЕНКИ КАЧЕСТВА
КАТЕГОРИЗАЦИИ ТЕКСТОВОЙ КОЛЛЕКЦИИ ДОКУМЕНТОВ**

¹ Юлия Ефимовна Балыкина

² Михаил Владимирович Сvirкин

¹ Санкт-Петербургский государственный университет
199034, г. Санкт-Петербург, Университетская наб., 7/9
ассистент

² Санкт-Петербургский государственный университет
199034, г. Санкт-Петербург, Университетская наб., 7/9
кандидат физико-математических наук, доцент
E-mail: julia.balykina@gmail.com

В работе рассмотрена возможность использования функции вариации информации в качестве критерия качества категоризации коллекции текстов. Проанализировано поведение функции вариации информации в зависимости от объема обучающей выборки и количества категорий.

Ключевые слова: категоризация текстов, критерий качества, функция вариации информации, объем выборки.