КЛАСТЕРИЗАЦИЯ НЕСТРУКТУРИРОВАННОГО ТЕКСТА ПО ЗАРАНЕЕ ЗАДАННЫМ КАТЕГОРИЯМ

Раджабова Н.Ш.[1], Махмудов М.Р.[2] Email: Radzhabova678@scientifictext.ru

[1]Раджабова Наима Шамильевна – кандидат физико-математических наук, доцент;
[2]Махмудов Магомед Риадович – магистрант,
кафедра дискретной математики и информатики,
Дагестанский государственный университет,
г. Махачкала

Аннотация: *в данной работе исследуется актуальная проблема увеличения разметочных данных путём использования кластеризации. При широкой востребованности кластеризации возникает необходимость в программных системах, предоставляющих возможности как для анализа работы алгоритмов, так и данных, а также для удобного отображения результатов.*
Разметочные данные – конечное множество объектов в виде вектора параметров, описывающих объект. В контексте данной работы процесс векторизации производится на строковых объектах, необходимых для кластерного анализа.
Ключевые слова: *машинное обучение, кластеризация, метод к-средних, кластерный анализ.*

CLUSTERING OF UNSTRUCTURED TEXT BY ADVANCE OF PRESCRIBED CATEGORIES

Radzhabova N.Sh.[1], Makhmudov M.R.[2]

[1]Radzhabova Naima Shamilyevna – Candidate of Physical and Mathematical Science, Associate Professor;
[2]Makhmudov Magomed Riadovich – Undergraduate,
DEPARTMENT OF THE DISCRETE MATHEMATICS AND COMPUTER SCIENCE,
DAGESTAN STATE UNIVERSITY,
MAKHACHKALA

Abstract: *in this paper, we study the urgent problem of increasing markup data by using clustering. With the wide demand for clustering, there is a need for software systems that provide opportunities for both analyzing the operation of algorithms and data, as well as for convenient display of results.*
Markup data is a finite set of objects in the form of a vector of parameters describing the object. In the context of this work, the vectorization process is performed on string objects required for cluster analysis.
Keywords: *machine learning, clustering, k-means method, cluster analysis.*

УДК 004.89

Кластеризация – это процесс разбиения множества элементов на подгруппы в зависимости от их схожести. Элементами множества могут выступать разного рода объекты, в качестве описания объекта выступают вектор характеристик. Группы принято называть кластерами.

Кластер – это группа объектов, схожих между собой по определённым признакам – количественным характеристикам объектов, которые описываются векторами. В метрических пространствах степень схожести определяется путём использования нормы расстояния между векторами [1].

Кластерный анализ или кластеризация – это основная задача интеллектуального анализа данных и общий метод статистического анализа данных, используемый во многих областях, включая машинное обучение, распознавание образов, анализ изображений, поиск информации, биоинформатику, сжатие данных и компьютерную графику. Решение может быть достигнуто с помощью различных алгоритмов, которые существенно различаются в понимании того, что представляет собой кластер и как их эффективно находить.

При широкой востребованности кластеризации возникает необходимость в программных системах, предоставляющих возможности как для анализа работы алгоритмов, так и данных, а также для удобного отображения результатов. Существующие на текущий момент аналогичные программные системы [2], в основном, являются либо частными исследовательскими разработками, обычно использующими конечный набор алгоритмов, либо коммерческими приложениями, рассчитанными на корпоративных клиентов и зачастую недоступными для индивидуальных, в том числе студенческих, исследований. Особенно выделяется фактор недоступности реализации своих алгоритмов на подобных системах.

Разметочные данные – набор данных, элементами которого являются векторы, описывающие объекты и метки. Метка в контексте разметочных данных, – это вектор или скалярное значение, которое должно получиться на выходе алгоритма.

В данной работе исследуется **актуальная проблема** увеличения разметочных данных путём использования кластеризации.

Дан набор данных $X_m = \{x_1 \ \dots \ x_m\} \subset X\ (m>0)\dots$ и функция, определяющая степень сходства объектов, в большинстве случаев это функция от объектов $p(x_i, x_j)$.

Необходимо разделить последовательность *X* на однородные подмножества, называемые кластерами таким образом, чтобы каждый кластер был организован из объектов близких по метрике *p*, а объекты разных групп отличались по этой же метрике. Алгоритм кластерного анализа является функцией следующего отображения $f: X \rightarrow Y$, ставя объекту *x* из множества *X* в соответствие метку кластера *y* из множества *Y*.

Основная цель кластерного анализа – это получение сведений о структуре данных, что представляет собой первый этап детального анализа данных.

Целью данной работы является разработка веб-приложения, решающего проблему увеличения разметочных данных, функционал которого включает демонстрацию кластеризации и возможность увеличения размеченных строковых объектов.

Для достижения этой цели необходимо решить следующие **задачи**:

- реализовать платформу для взаимодействия пользователя с программным продуктом;
- предоставить пользователю удобное отображение кластеров;
- реализовать возможность увеличения разметочных данных.

Инструментарий**,** используемый в работе, включает в себя:

- Веб-фреймворк *Flask* для разработки серверной части веб-приложения.
- *Front-end* фреймворк *bootstrap* для разработки интерфейса веб-приложения.

Исходные данные – конечное множество объектов в виде вектора параметров, описывающих объект. Для каждого объекта в множестве собраны измерения определённого рода данных, обозначающих данный объект. Например, (x, y). Совокупность объектов в множестве называется выборкой. По этим данным необходимо выявить общие взаимосвязи, закономерности, относящиеся не только к данной выборке, но и ко всем данным, по которым система не обучалась.

Описание работы алгоритма

Предварительная обработка и очистка данных являются одними из важнейших шагов в обеспечении эффективного использования набора данных для алгоритмов машинного обучения. Векторизация – это процесс конвертирования данных в матричную или векторную форму удобную для матричных операций.

В контексте данной работы процесс векторизации производится на строковых объектах, необходимых для кластерного анализа.

Основным критерием для выбора алгоритма кластеризации служила скорость. При изучении и выявлении преимуществ и недостатков известных алгоритмов, был сделан выбор в пользу алгоритма k-means.

Алгоритм *k*-средних относится к классу итерационных алгоритмов кластерного анализа; поэтому нам заранее должно быть известно, на какое количество кластеров должно разбиться множество наблюдений.

Основная идея метода k-средних состоит в выделение определённых кластеров в данных с целью минимизировать среднеквадратическое отклонение расстояния каждого наблюдения от центра каждого кластера:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_j)^2 \rightarrow min$$

k – количество кластеров, S_i – кластеры (строковые объекты), μ – центры кластеров [2].

Программная система представляет собой веб-приложение, которое позволяет пользователю вводить запросы и просматривать результат кластерного анализа этих запросов, также на основе этих запросов можно провести процедуру увеличения разметочных данных. Данная возможность является демонстрацией решения проблемы увеличения разметочных данных.

Список литературы / References

1. Хранение данных: вендоры, объемы данных, прогнозы. [Электронный ресурс]. Режим доступа: https://www.crn.ru/news/detail.php?ID=124815/ (дата обращения: 25.11.2019).
2. Обзор алгоритмов кластеризации данных. [Электронный ресурс]. Режим доступа: https://habr.com/ru/post/101338/ (дата обращения: 1.12. 2019).
3.