МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

Анализ работы алгоритмов кластеризации на медицинских данных

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Борисовой Юлии Сергеевны

Научный руководитель
д. ф.-м. н., доцент ______________ С. П. Сидоров

Заведующий кафедрой
д. ф.-м. н., доцент ______________ С. П. Сидоров

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы исследования

С появлением и последующим развитием таких технологий как Big Data, бизнес-аналитика, а также приложений, требующих автоматизации, возник и спрос на продвинутую аналитику данных, возможную только при использовании машинного обучения. В такой ситуации методы интеллектуального анализа данных приобретают особую актуальность. Их основная особенность заключается в установлении наличия и характера скрытых закономерностей, тогда как традиционные методы занимаются главным образом параметрической оценкой уже установленных правил.

Машинное обучение — это попытка научить компьютеры самостоятельно обучаться на большом количестве данных вместо жестко постулированных правил.

Алгоритмы машинного обучения могут быть либо контролируемыми, либо неконтролируемыми, хотя некоторые авторы также классифицируют такие алгоритмы как обучение с подкреплением, которое направленно на изучение данных и идентификации реагирующих на окружающую среду паттернов поведения. Контролируемое машинное обучение помимо использования входных атрибутов опирается на заранее определенный выходной атрибут.

Главная задача неконтролируемого машинного обучения заключается в попытке найти некую скрытую структуру в немаркированных данных. Поскольку примеры, приведенные обучаемому, не помечены, то сигнал ошибки или поощрения, позволяющий оценить правильность возможного решения, не возникает. Недостаток неконтролируемого машинного обучения состоит в том, чтобы определить, правильно ли работает программа, поскольку метка вывода неизвестна.

Бесконтрольное обучение включает в себя множество методов, направленных на обобщение и объяснение ключевых особенностей или структур данных. Одним из этих методов является кластеризация.

Кластеризацию используют и как самостоятельный инструмент анализа данных, и как предварительный этап для других методов анализа, таких как, например, классификация или деревья решений. Однако при всем этом оценка качества кластеризации является мало разработанной областью, и зачастую вопрос о том, насколько хороша или плоха структура кластеров, приходится решать «вручную».

На сегодняшний день существует различное множество алгоритмов кластеризации, самыми распространенными из которых являются иерархическая кластеризация и метод К-средних, подходящие для обработки разнообразных данных, в том числе и медицинских.

Актуальность определила выбор темы данной работы: «Анализ работы алгоритмов кластеризации на медицинских данных».

Цель работы — проведение анализа выбранных методов кластеризации, иерархической кластеризации и метода К-средних, для выявления наиболее эффективного способа кластеризации на примере медицинских данных с использованием языка Python.

Для достижения поставленных целей в работе необходимо решить следующие задачи:

— Рассмотреть современные методы неконтролируемого обучения, используемые для обработки данных;
— Провести кластеризацию медицинских данных методами иерархическим и К-средних;
— Сравнить два алгоритма на основе показателей качества кластеризации;
— Разработать программный код на языке Python для воспроизведения расчетов.

Практическая значимость работы заключается в анализе и сравнении алгоритмов кластеризации для выявления качественного и эффективного метода кластеризации на примерах медицинских данных.

Основное содержание работы

Бакалаврская работа состоит из: введения, теоретического и практического разделов, заключения, списка использованных источников, четырех приложений.

Введение содержит основные положения: актуальность темы исследования (цель, объект, предмет, задачи исследования); практическую значимость исследования.

Первый раздел «Методы кластеризации» описывает теоретические основы проведения и современные методы кластеризации.

Постановка задачи кластеризации

По определению кластеризация — это процесс группировки похожих объектов, то есть разбиения немаркированных данных на непересекающиеся подмножества кластеров таким образом, чтобы:

- Данные внутри кластера были идентичны (в этом случае говорится о высоком внутриклассовом сходстве);
- Данные в разных кластерах были различны (в этом случае говорится о низком межклассовом сходстве).

Формальная постановка задачи кластеризации:

Пусть X – множество объектов, Y – множество номеров (имён, меток) кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, \ldots, x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i.

Обзор методов

В рамках различных алгоритмов кластеризации существуют алгоритмы мягкого разбиения, которые присваивают вероятность принадлежности данных к каждому кластеру, а также алгоритмы жесткого разбиения, где каждой точке данных присваивается определенная принадлежность к одному кластеру. Ярким примером алгоритма мягкого разбиения является EM кластеризация, основанная на смешанной Гауссовой модели (GMM), вероятностной модели, которая предполагает, что все точки данных генерируются из комбинаций конечного числа гауссовых распределений с неизвестными параметрами.

Альтернативой алгоритмам мягкого разбиения являются алгоритмы жесткого разбиения, которые присваивают каждому элементу пространства признаков уникальное кластерное значение. В соответствии с процессом группировки алгоритма жесткого разбиения существует три группы методов кластеризации:

- Алгоритмы разбиения: начинаются со случайного разбиения с последующим его совершенствованием. Иногда такие алгоритмы называют «плоской» кластеризацией. Примеры алгоритмов разбиения: K-средние и спектральная кластеризация;
- Иерархические алгоритмы: организуют данные в иерархические структуры, где данные могут быть объединены в направлении снизу вверх или разделены сверху вниз. Примеров иерархических алгоритмов является агломеративная кластеризация.
- Алгоритмы кластеризации на основе плотности: идентифицируют отличительные группы/кластеры в данных, основываясь на идее, что кластер в пространстве данных представляет собой непрерывную область с высокой плотностью точек, отделенную от других таких кластеров смежными областями с низкой плотностью. DBSCAN — один из примеров таких алгоритмов.

Метод K-средних

Кластеризация с использованием K-средних является распространенным примером эксклюзивного метода кластеризации, в котором точки данных назначаются в K групп, где K представляет количество кластеров на основе расстояния от центроида каждой группы. Точки данных, наиболее близкие к заданному центроиду, будут сгруппированы по одной и той же категории. Большее значение K будет указывать на меньшие группы с большей степенью детализации, тогда как меньшее значение K будет иметь более крупные группировки и меньшую степень детализации. Кластеризация K-средних обычно используется при сегментации рынка, кластеризации документов, сегментации изображений и сжатии изображений.

K-средних делит набор из n выборок X на k непересекающихся кластеров c_i, где $i = 1, ..., k$, каждый из которых описывается средним значением μ_i выборок в кластере. Эти средние значения обычно называют центроидами кластеров. Алгоритм K - средних предполагает, что все k групп имеют одинаковую дисперсию.

Несмотря на то, что метод кластеризации K-средних имеет преимущества, позволяющие легко использовать эвристику для выбора хороших начальных значений; инициализировать начальные значения другими методами; исследовать множество точек, которые еще не изучены. Для него характерны следующие недостатки: алгоритм не может гарантировать преодоление проблемы локальных минимумов; он является итеративным и, следовательно, медленным при значительном количестве образцов большой размерности; и стремится искать сферические кластеры.

Иерархическая кластеризация

Другим широко известным методом кластеризации, представляющим особый интерес, является иерархическая кластеризация. Иерархическая кластеризация состоит из общей группы алгоритмов кластеризации, которые создают вложенные кластеры путем последовательного слияния или разделения данных. Иерархия кластеров представлена в виде дерева. Дерево часто называют дендрограммой. Корнем дендрограммы является единственный кластер, содержащий все образцы; листья — это кластеры, каждый из которых содержит только один образец.

В целом, существует два типа иерархической кластеризации:

- Нисходящая (дивизивная) кластеризация;
- Восходящая (агломеративная) кластеризация.

Для того чтобы решить, какие кластеры следует объединить (для агломеративной) или как кластер должен быть разделен (для дивизивной), необходима мера несходства между наборами наблюдений. Критерий связывания определяет метрику, используемую для стратегии объединения кластеров:

- Максимальная или полная связь сводит к минимуму максимальное расстояние между наблюдениями пар кластеров. Формула: $max\{d(a,b) : a \in A, b \in B\}$;
- Метод одиночной (минимальной) связи также известен, как «метод ближайшего соседа». Формула: $min\{d(a,b) : a \in A, b \in B\}$;
- Средняя связь усредняет сходство между членами, то есть минимизирует среднее значение расстояний между всеми наблюдениями пар кластеров. Формула: $\frac{1}{|A||B|}\sum\limits_{a\in A}\sum\limits_{b\in B} d(a,b)$;
- Связь Уорда сводит к минимуму сумму квад-

ратов разностей внутри всех кластеров. Формула: $\Delta = \sum_{i} (x_i - \overline{x})^2 - \sum_{x_i \in A} (x_i - \overline{a})^2 - \sum_{x_i \in B} (x_i - \overline{b})^2$.

Оценка качества кластеризации

Одним из наиболее известных способов сравнения результатов методов кластеризации в статистике является индекс Рэнда или мера Рэнда (названный в честь Уильяма М. Рэнда). Индекс Рэнда оценивает сходство между двумя результатами кластеризации данных. Дано множество из n элементов $S = \{o_1, ..., o_n\}$ и два разбиения S: $X = \{X_1, ..., X_r\}$, разбиение S на r подмножеств; и $Y = \{Y_1, ..., Y_s\}$, разбиение S на s подмножеств, определяющих следующее:

- a — количество пар элементов в S, находящихся в одном и том же подмножестве как в X, так и в Y;
- b — количество пар элементов в S, находящихся в разных подмножествах как в X, так и в Y;
- c — количество пар элементов в S, находящихся в одном и том же подмножестве в X , но в разных подмножествах в Y; и
- d — количество пар элементов в S , находящихся в разных подмножествах в X, но в одном и том же подмножестве в Y.

Индекс Рэнда, R, определяется формулой:

$$R = \frac{a+b}{a+b+c+d},$$

где его значение гарантированно находится между 0 и 1.

Одна из проблем индекса Рэнда заключается в том, что при наличии двух наборов данных со случайными метками он не принимает постоянного значения (например, нуля), как ожидалось. Более того, при увеличении числа кластеров желательно, чтобы верхний предел стремился к единице. Для решения этой проблемы используется форма индекса Рэнда, называемая скорректированным индексом Рэнда, которая изменяет его относительно случайной группировки элементов. Его результат принадлежит промежутку [-1, 1].

Альтернативой предыдущей оценке служит анализ окончательной «формы» результата кластеризации. Что в свою очередь является главной идеей, лежащей в основе коэффициента Силуэта. Он определяется как функция внутрикластерного расстояния выборки в наборе данных a и ближайшего кластерного расстояния b для каждой выборки. Коэффициент Силуэта для выборки i можно записать следующим образом:

$$Silhouette(i) = \frac{b - a}{\max(a, b)}$$

Следовательно, если коэффициент Силуэта $s(i)$ стремится к 0, то выборка находится на границе своего кластера и наиболее близка к остальным кластерам набора данных. Отрицательное значение говорит о том, что выборка находится ближе к соседнему кластеру.Высокое положительное значение, то есть близкое к 1, будет означать компактное скопление, и наоборот.

Второй раздел «Анализ данных на языке Python» описывает проведение сравнительного анализа алгоритмов кластеризации с использованием медицинских данных.

Кластеризация данных о риске возникновения ССЗ

В данной работе на основе данных оценки фактора риска развития сердечнососудистого заболевания определялось качество кластеризации. Медицинский центр V. A. в Лонг-Бич опубликовал анонимные данные своих пациентов граждан Калифорнии следующих показателей: пол, возраст, содержание холестерина в крови, артериальное давление в состоянии покоя, уровень сахара в крови натощак, результаты ЭКГ, максимальная частота сердечных сокращений, сопутствующие симптомы — тип боли в груди, стенокардия, вызванная физическими упражнениями и др., а также проявление заболеваемости среди пациентов сердечно-сосудистой системы.

На основе анализа медицинской литературы для кластеризации были отобраны следующие факторы: уровень холестерина в крови, результаты ЭКГ, максимальная частота сердечных сокращений и артериальное давление.

Количество кластеров выбиралось с помощью метода локтя. В методе локтя количество кластеров (K) изменялось от 1 до 11. Для каждого значения K вычисляется значение WCSS. WCSS — это сумма квадратов расстояния между каждой точкой и центроидом в кластере. Значение точки, в которой кривая меняет свое направление, напоминая тем самым форму локтя, и будет оптимальным количеством кластеров, то есть 2.

Таким образом, кластеризацию с помощью K-средних проводили с уже с известным количеством кластеров с использованием библиотеки Scikit-learn, в которой реализована схема инициализации k-means ++, решающая проблему локальных минимумов в кластеризации K-средних.

Далее строились графики распределения объектов на кластеры. Установленные предположения на основе анализа графиков кластеризации K-средних и качество ее проведения доказываются при помощи математических расчётов: скорректированного индекса Рэнда и коэффициента Силуэта.

Следует отметить, что индекс Рэнда оценивает, насколько много из тех пар элементов, которые находились в одном кластере, и тех пар элементов, которые находились в разных кластерах, сохранили это состояние после кластеризации. В свою очередь коэффициент Силуэта показывает, насколько объект похож на свой кластер по сравнению с другими кластерами.

Таким образом, чем ближе значение коэффициента или индекса к 1, тем эффективней была проведена кластеризация.

Следующим этапом данной работы являлось проведение анализа данных при помощи иерархической агломеративной кластеризации. Прежде чем применять иерархическую кластеризацию, необходимо узнать количество кластеров с помощью дендрограммы.

Агломеративная кластеризация проводилась по методу Уорда, который применяется для задач с близко расположенными кластерами и аналогичен целевой функции K-средних, но решается с помощью агломеративного иерархического подхода.

Аналогично алгоритму проведения анализа кластеризации методом K-средних проводился анализ результатов иерархической кластеризации. Для полной оценки качества кластеризации сравнивали алгоритмы на основе индекса Рэнда и коэффициента Силуэта, полученные при помощи кластеризации данных ССЗ по методу K-средних и иерархической агломеративной кластеризации.

Кластеризация данных о заболеваемости гепатитом С

В работе рассматривались данные о заболеваемости гепатитом С, предоставленные Высшей медицинской школой Ганновера, Германия, которые содержат лабораторные показатели доноров крови и пациентов с гепатитом С, а также демографические показатели, такие как возраст и пол.

С целью повышения чистоты эксперимента проведения анализа данных были отобраны сразу несколько показателей, чтобы выявить такие биохимические показатели, которые помогут диагностировать вирусный гепатит С.

Также, как и в первом случае для выявления эффективного метода кластеризации были выбраны два алгоритма: K-средних и иерархическая агломеративная кластеризация. На первом этапе сравнивали значимые показатели при помощи метода кластеризации K-средних.

С использованием метода локтя было определено оптимальное количество кластеров — 3. После определения количества кластеров проводилась кластеризация методом K-средних, результаты которой были показаны на графиках.

Проверку гипотез относительной значимости показателей вирусного гепатита С, сделанных на основе графиков, проводили при помощи скорректированного индекса Рэнда и коэффициента Силуэта.

Затем для анализа данных использовали иерархическую агломеративную кластеризацию, перед этим подтвердив оптимальное количество кластеров построением дендрограммы. Воспользовавшись методом Уорда, агломеративную кластеризацию проводили на данных с предполагаемыми значимыми биохимическими показателями. Результаты распределения объектов на кластеры представлялись на графиках.

Установленные предположения на основе анализа графиков агломеративной кластеризации и качество ее проведения доказывались при помощи алгебраических расчётов: скорректированного индекса Рэнда и коэффициента Силуэта.

Затем метод K-средних и иерархическая кластеризация сравнивались на основе качества кластеризации.

Рассчитанные методы оценки качества кластеризации показали разный результат, поэтому нельзя однозначно сказать какой алгоритм будет универсальным и хорошо работать на любых данных. Таким образом следует разработать совершенный неизвестный на данный момент алгоритм, который позволит качественно проводить кластеризацию для разных типов данных.

Основные результаты

1. В работе была рассмотрена кластеризация как одна из задач неконтролируемого обучения, предназначенная для обнаружения закономерностей и структуры в немаркированных данных.
2. Построены графики кластеризации медицинских данных иерархическим методом и методом K-средних.
3. Описаны результаты проведения кластеризации медицинских данных иерархическим методом и методом K-средних.
4. Проведен сравнительный анализ двух алгоритмов кластеризации на основе оценки качества кластеризации.

   Программный код приводится в приложениях А, Б, В, Г.