МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

ПРИЛОЖЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ К АНАЛИЗУ МЕДИЦИНСКИХ ДАННЫХ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика

механико-математического факультета
Борисовой Юлии Сергеевны

Научный руководитель
доцент, к. ф.-м. н., доцент _____ С. С. Волосивец

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

Саратов 2024

ВВЕДЕНИЕ

Актуальность темы исследования

С появлением и последующим развитием таких технологий как Big Data, бизнес-аналитика, а также приложений, требующий 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, метод K-средних, DBSCAN и сдвиг среднего значения, подходящие для обработки разнообразных данных, в том числе и медицинских.

Актуальность определила выбор темы данной работы: «Приложение методов кластеризации к анализу медицинских данных».

Цель работы — проведение анализа исследуемых методов кластеризации для определения наиболее эффективного алгоритма на примере обработки медицинских данных.

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, K-средних, DBSCAN и Mean Shift;
— Сравнить четыре алгоритма на основе показателей качества кластеризации;
— Разработать программный код на языке Python для воспроизведения расчетов.

Практическая значимость работы заключается в оперативной постановке диагноза онкологического заболевания на ранней стадии прогрессирования опухоли.

Научная новизна работы состоит в использовании методов кластеризации на разных типах данных: категориальных, числовых и графических, с целью выявления алгоритма наиболее стабильного к особенностям типов данных.

Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции «Актуальные проблемы математики и механики», которую проводил механико-математический факультет СГУ в

апреле 2024 года, в секции «Анализ данных», в XII Международной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2023 года.

Результаты работы были опубликованы в статье «Поиск эффективного алгоритма кластеризации для диагностики протоковой аденокарциномы поджелудочной железы на ранних стадиях».

Основное содержание работы

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

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

Первый раздел «Понятие кластеризации и ее методы» 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 и Mean Shift — 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.

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.

DBSCAN

DBSCAN — основанная на плотности пространственная кластеризация для приложений с шумами. Основная идея DBSCAN заключается в том, что точка принадлежит к кластеру, если она близка ко многим точкам из этого кластера.

Существует два ключевых параметра DBSCAN:

- ε — расстояние, определяющее окрестности. Две точки считаются соседними, если расстояние между ними меньше или равно ε.
- $minPts$ — минимальное количество точек данных для определения кластера.

Mean Shift

Сдвиг среднего значения — это алгоритм кластеризации на основе плотности, который определяет моды функции плотности, представляющих собой кластеры. Основная идея алгоритма заключается в смещении каждой точки данных в сторону моды (т.е. наибольшей плотности) распределения точек в пределах определенного радиуса. Алгоритм итеративно выполняет эти сдвиги до тех пор, пока точки не сходятся к локальному максимуму функции плотности, представляющему собой кластеры данных.

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.

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, то есть коэффициент 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.

Помимо рассмотренного ранее коэффициента Силуэта к внутрен-

ним критериям валидации относят Индекс Калинского-Харабаша (Calinski-Harabasz) и Индекс Дэвиса-Болдина.

Индекс Калинского-Харабаша имеет вид (a . разделение)/(b . сплоченность), где a и b — веса.

Более высокое значение индекса означает, что кластеры плотные и хорошо разделены, хотя «приемлемого» порогового значения не существует.

Индекс Дэвиса-Болдина рассчитывается как среднее значение меры сходства каждого кластера с наиболее похожим на него кластером. В данном контексте сходство определяется как отношение межкластерного и внутрикластерного расстояний. Таким образом, по индексу Дэвиса-Болдина хорошо разделенные кластеры с меньшим разбросом имеют более высокую оценку.

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

Кластеризация данных биохимических показателей как предикторов злокачественных новообразований поджелудочной железы

В этой части работы качество кластеризации определялось на основе числовых данных протоковой аденокарциномы поджелудочной железы (PDAC). Клинические данные были получены из нескольких центров: Банка тканей поджелудочной железы Barts, Университетского колледжа Лондона, Университета Ливерпуля, Испанского национального центра исследований рака, больницы Кембриджского университета и Белградского университета.

Ключевыми характеристиками являются четыре биомаркера анализа мочи: креатинин, LYVE1, REG1B и TFF1.

Креатинин — это белок, часто используемый в качестве показателя функции почек. LYVE1 (рецептор 1 гиалуроновой кислоты эндотелия лимфатических сосудов) – белок, который может играть роль в метастазировании опухоли. REG1B — белок, связанный с регенерацией поджелудочной железы. TFF1 — фактор трилистника 1, связанный с регенерацией и восстановлением мочевыводящих путей.

Так как исходные значения признаков принадлежат большому диапазону и отличаются друг от друга на несколько порядков, необходимо их нор-

мализовать. Главная задача нормализации — преобразовать объекты таким образом, чтобы они были в аналогичном масштабе.

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, представляющий сумму квадратов расстояния между каждой точкой и центроидом в кластере. Значение точки, в которой кривая меняет свое направление, напоминая форму локтя, и будет оптимальным количеством кластеров, в данном случае 3.

Таким образом, кластеризацию с помощью K-средних проводили с уже с известным количеством кластеров с использованием библиотеки Scikit-learn.

Далее строились графики распределения объектов на кластеры и проводился анализ их наполнения.

Затем проводили анализ 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 полученных иерархической кластеризацией кластеров.

Третий метод кластеризации, который использовали в работе, это метод на основе плотности — DBSCAN, отличающий от K-средних и агломеративной кластеризации самостоятельным созданием кластеров на основе двух гиперпараметров: minPts — минимальное количество точек данных, которые должны присутствовать в области. ε — радиус площади центральной точки.

Значения гиперпараметров рассчитывали с помощью матрицы исследуемых комбинаций и метода k-ближайших соседей.

Был также проведен анализ кластеров, чтобы сделать предположения относительно их наполненности.

На этапе исследования кластеризации с помощью метода Mean Shift

установлено, что сдвиг среднего значения сам находит количество кластеров на основе пропускной способности. Ее можно выбрать вручную или воспользоваться встроенной функцией библиотеки scikit-learn, которая рассчитывается на основе выборки и квантиля, чьи значения находятся в промежутке от 0 до 1. С помощью пропускной способности было выбрано оптимальное количество кластеров — 4.

Для полной оценки качества кластеризации сравнивали алгоритмы на основе индекса Рэнда, коэффициента Силуэта, индекса Калинского-Харабаша и индекса Дэвиса-Болдина. Сравнительный анализ выше упомянутых показателей данных скрининга протоковой аденокарциномы поджелудочной железы выявил наиболее эффективный алгоритм кластеризации — K-средних.

Кластеризация данных показателей риска возникновения злокачественных новообразований легочной ткани

В данной части работе рассматривались данные, предоставленные Национальным институтом онкологии, США, которые содержат симптомы и факторы риска возникновения рака легких, а также демографические показатели, такие как возраст и пол. К факторам риска относятся курение (Smoking), наличие тревожности (Anxiety), давление со стороны окружающих (Peer pressure), наличие хронических заболеваний (Chronic disease), повышенная утомляемость (Fatigue), аллергии (Allergy) и злоупотребление алкоголем (Alcohol). К симптомам — наличие желтых пальцев (Yellow fingers), хрипов (Wheezing), кашля (Coughing), отдышки (Shortness of breath), затрудненного глотания (Swallowing difficulty) и боли в области грудины (Chest pain).

Все характеристики относятся к категориальным со значениями 1 — «Нет», 2 — «Да», за исключением возраста, который является числовой переменной.

В работе исследуется возможность проведения кластеризации и разбиение данных на кластеры на основе всех параметров. Поэтому чтобы уменьшить размерность, но при этом не потерять информативность был использован метод главных компонент (PCA).

Также, как и в первом случае для выявления эффективного метода

кластеризации были выбраны четыре алгоритма: K-средних, иерархическая агломеративная кластеризация, DBSCAN и Mean Shift. На первом этапе сравнивали значимые показатели при помощи метода кластеризации K-средних.

С использованием метода локтя было определено оптимальное количество кластеров — 2. 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 факторами риска злокачественных новообразований. На следующем этапе кластеризацию проводили с помощью DBSCAN, предварительно выявив подходящие гиперпараметры.

Далее для анализа данных показателей риска возникновения злокачественных новообразований легочной ткани применяли кластеризацию с помощью сдвига среднего значения. Как уже упоминалось раннее алгоритм сам рассчитывает количество кластеров на основе пропускной способности.

Установленные предположения на основе анализа кластеризации и качество ее проведения доказывались при помощи математических расчётов: скорректированного индекса Рэнда, коэффициента Силуэта, индекса Калинского-Харабаша и индекса Дэвиса-Болдина. Из полученных результатов можно сделать вывод, что наиболее эффективным алгоритмом кластеризации, на данном этапе исследований является также алгоритм K-средних.

Кластеризация результатов гистопатологических исследований колоректального рака

В последнем подразделе работы исследовались данные, предоставленные Корнеллским университетом и содержащие гистопатологические изображения микросреза ткани прямой кишки, среди которых имеются изображения как аденокарциномы, так и доброкачественных новообразований.

Вначале извлекались признаки изображения для получения его вычислительной информации с помощью гистограммы ориентированных градиентов (HOG). Чтобы сравнить полученные векторы друг с другом, вычислялось их сходство с помощью дивергенции Йенсена-Шеннона.

Так получился достаточно большой по объему набор данных, чей размер был уменьшен с помощью метода главных компонент (PCA). Данные

были предварительно нормализованы, чтобы убрать сильный разброс значений.

Первым из испытанных алгоритмов кластеризации стал метод K-средних. Анализируя данные методом локтя, установлено оптимальное количество кластеров — 2, в этой в точке можно наблюдать тенденцию к изменению направления кривой.

Кластеризация с помощью иерархического алгоритма, согласно дендрограммы, показала эффективное разбиение данных на 2 кластера, подтверждая этим результаты метода локтя.

Прежде чем провести DBSCAN были выбраны значения двух гиперпараметров: ε и minPts, на основе которых алгоритм разделил данные на 2 кластера и кластер выбросов.

Кластеризация при помощи сдвига среднего значения проводилась с предварительно рассчитанным значением пропускной способности.

Результаты кластеризации данных гистопатологических изображений путем иерархической кластеризации, методом K-средних, DBSCAN и Mean shift анализировались при помощи индексов Рэнда, Калинского-Харабаша, Дэвиса-Болдина и коэффициента силуэта. Установлено, что для диагностики онкологии кишечника можно сделать выбор в пользу двух наиболее производительных алгоритмов кластеризации данных гистопатологических снимков ткани прямой кишки: K-средних и Mean-Shift.

В заключении следует отметить, что сравнительный анализ результатов оценки показателей качества: индекс Рэнда, коэффициент Силуэта, индекс Калинского-Харабаша и индекс Дэвиса-Болдина, доказал эффективность проведенной кластеризации медицинских данных методом K-средних.

Кроме того, анализ методов кластеризации проводился на разных типах данных: категориальных, числовых и графических, с целью выявления алгоритма наиболее стабильного к особенностям типов данных.

Основные результаты

1. В работе были изучены методы неконтролируемого обучения, в частности, кластеризации, с целью обнаружения закономерностей и структуры в немаркированных данных.

2. Построены графики кластеризации медицинских данных иерархическим методом, методом K-средних, DBSCAN и сдвигом среднего значения.
3. Описаны результаты проведения кластеризации медицинских данных разных типов иерархическим методом, методом K-средних, DBSCAN и сдвигом среднего значения.
4. Проведен сравнительный анализ алгоритмов кластеризации на основе оценки качества кластеризации.

   Программный код приводится в приложениях А — М.