ОЦЕНИВАНИЕ КАЧЕСТВА КЛАСТЕРИЗАЦИИ ДАННЫХ ДО И ПОСЛЕ УСТРАНЕНИЯ АНОМАЛИЙ

С.В. Аксёнов, к.т.н., доц. ОИТ ИШИТР,
Е.В. Сафронова, аспирант гр. А0-36,
В.С. Сафронов, магистрант гр. 8ПМ0И1,
Томский политехнический университет
E-mail: ev.kashcheeva@mail.ru

Введение

Выделение групп из множества данных является одним из инструментов выявления зависимостей между теми или иными параметрами. Эффективность работы алгоритма кластеризации зависит, как от настройки гиперпараметров, так и от качественной подготовки поступающих на вход алгоритма данных [1]. В зависимости от предметной области, данные имеют свою специфику, сфера медицины не является исключением.

Отделением инфекционных заболеваний Сибирского государственного медицинского университета были предоставлены результаты анализов пациентов, страдающих рожистым воспалением, малярией и клещевым энцефалитом. Предоставленные данные содержали такие аномалии, как пропущенные значения и выбросы. Было решено определить, насколько качество кластеризации результатов анализа крови пациентов с разными заболеваниями зависит от наличия аномалий в данных.

Целью данного исследования является оценивание качества кластеризации результатов общего анализа крови до и после устранения аномалий.

Описание работы

Для достижения поставленной цели был написан программный код на языке программирования Python. Для начала была осуществлена загрузка данных результатов общего анализа крови из файлов формата «xls». В результате объединения данных из трех файлов был сформирован набор данных, состоящий из 341 записи и 45 столбцов. Большинство параметров содержало большое количество пропущенных значений, данные параметры были исключены из дальнейшей работы. Также были исключены показатели, рассчитывающиеся на основе других. В итоге из 45 было выбрано 5 основных показателей, на основе которых затем была проведена кластеризация.

Для кластеризации использовался алгоритм k-средних. Все параметры алгоритма кластеризации, кроме количества кластеров, были установлены по умолчанию. На вход алгоритма подавались такие показатели, как содержание эритроцитов (RBC), лейкоцитов (WBC), гемоглобина (HGB), тромбоцитов (PLT) и гематокрита (HCT). Для оценки качества кластеризации в качестве метки кластера выступало заболевание, которым страдает пациент.

Было подготовлено три набора данных: набор, в котором были исключены записи, содержащие пропуски и выбросы; набор, в котором пропуски были заменены на медиану по столбцу, а выбросы остались без изменения; набор, в котором пропуски и выбросы были заменены на медиану по столбцу. В таблице 1 представлены значения метрик оценки эффективности кластеризации [2] в зависимости от наличия аномалий данных.

Таблица 1. Метрики оценки качества кластеризации

Набор данных	Индекс Рэнда	Индекс Фаулкса–Маллоуз	Силуэтный коэффициент
Пропуски и выбросы удалены	0.581	0.397	0.458
Пропуски импутированы, выбросы без изменений	0.574	0.439	0.473
Пропуски и выбросы импутированы	0.576	0.441	0.472

Индекс Рэнда представляет собой меру сходства двух кластеров. Для расчета данного индекса формируются пары элементов. Подсчитывается количество пар, находящихся в одном кластере согласно реальным меткам кластера и меткам, полученным в результате работы алгоритма кластеризации. Также подсчитывается количество пар, находящихся в разных кластерах. Две эти величины суммируются, полученная сумма делится на общее количество возможных пар в наборе

данных. Если кластеры не согласуются не по одной паре элементов, то индекс принимает значение 0, если же кластеры согласуются по всем парам, то индекс принимает значение 1 [3]. Для всех трех наборов данных значение индекса превышает 0,5, однако, в первом случае, когда аномальные значения данных не участвовали в кластеризации, индекс Рэнда оказался выше.

Индекс Фаулкса-Маллоуз определяет сходство между двумя кластерами на основе числа истинно-положительных, ложно-положительных и ложно-отрицательных результатов. Количество истинно-положительных результатов – это количество пар элементов, которые принадлежат к одним и тем же кластерам, как в истинных метках, так и в предсказанных алгоритмом кластеризации. Количество ложно-положительных результатов – это количество пар элементов, принадлежащих к одинаковым кластерам в истинных метках, но не принадлежащих к одинаковым кластерам в предсказанных метках. Количество ложно-отрицательных результатом – это количество пар элементов, принадлежащих к одинаковым кластерам в предсказанных, а не в истинных метках. Индекс Фаулкса-Маллоуз рассчитывается как среднее геометрическое точности и полноты. Чем значение индекса выше, тем больше сходство между двумя кластерами [4]. В данном случае наилучший результат показала кластеризация данных, аномалии которых были импутированы значениями медианы по столбцу.

Также для подготовленных наборов данных была рассчитана такая метрика, как силуэтный коэффициент. По сравнению с рассмотренными выше метриками оценки эффективности кластеризации, данный коэффициент может использоваться, когда набор данных не размечен, т.е. истинные значения меток кластеров не известны. Коэффициент силуэта рассчитывается для каждого элемента набора данных. Рассчитываются среднее расстояние между элементом и другими элементами того же кластера, среднее расстояние между элементом и другими элементами в другом ближайшем кластере. Первая величина вычитается из второй, и их разность делится на максимальное значение из этих двух величин. Для набора данных коэффициент силуэта равен среднему значению коэффициентов для каждого элемента. Значения коэффициента находятся в диапазоне от минус 1 до 1. Высокое значение указывает на то, что элемент хорошо соответствует своему кластеру и плохо соответствует другим кластерам [5]. Значение данного показателя оказалось выше у данных, пропущенные значения которых были импутированы медианой, а выбросы остались без изменения.

Заключение

В результате проведенного исследования можно сделать вывод, что оценка эффективности кластеризации данных, подвергшихся различным мерам по устранению аномалий, дала неоднозначные результаты.

В случае, когда аномальные значения были удалены из набора данных, размер набора данных сократился, а значит и число возможных сочетаний пар элементов также сократилось. Знаменатель индекса Рэнда оказался меньше, чем у двух других наборов данных, поэтому, вероятнее всего, данный набор получил наивысшую оценку.

Что касается силуэтного коэффициента и Индекса Фаулкса-Маллоуз, то по сравнению с первым набором данных, второй и третий находятся в лидерах. Данный факт говорит о том, что такая мера по устранению аномалий, как импутация, повышает качество кластеризации.

Список использованных источников

1. Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2017. – 418 с.
2. Clustering. Clustering performance evaluation // Scikit-Learn. Machine Learning in Python [Электронный ресурс]. – URL: https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation (дата обращения 05.03.2021).
3. Rand W. M. Objective Criteria for the Evaluation of Clustering Methods // Journal of the American Statistical Association. - 1971. - №66. - С. 846-850.
4. Fowkles E. B., Mallows C. L. A Method for Comparing two hierarchical clusterings // Journal of the American Statistical Association. - 1983. - №78. - С. 553-569.
5. Rousseeuw P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. - 1987. - №20. - С. 53-65.