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

Кафедра информатики и программирования

АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ ДАННЫХ И МЕТОДЫ АНАЛИЗА ИХ РЕЗУЛЬТАТОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 «Математическое обеспечение и администрирование информационных систем»
факультета компьютерных и информационных технологий

Агуровой Лидии Павловны

Научный руководитель
зав. кафедрой к.ф.-м.н, доцент _____ Огнева М.В.

Зав. кафедрой
к.ф.м.н., доцент _____ Огнева М.В.

Саратов, 2019

ВВЕДЕНИЕ

В современном мире информационные технологии присутствуют во всех областях жизни человека, что приводит к появлению огромного объема сопутствующих данных. И, соответственно, появлению нового направления – обработка и анализ этих данных. Например, резкий скачок роста объема персональной информации произошел с увеличением популярности социальных сетей – в социальной сети Facebook по данным 2017 года были зарегистрированы более 2 миллиардов пользователей и 200 миллиардов связей между ними, а в социальной сети Вконтакте – более 7 миллионов. человек. При соответствующей обработке эти данные представляют собой огромную ценность, например, для предоставления целевой рекламы.

Конечно, накопление данных касается не только социальных сетей, но и любого другого вида деятельности – в Интернет переходят магазины, кинотеатры, рестораны, биржи и т.д. Огромный объем структурированных (а часто и не структурированных) данных принято называть «большими данными».

В настоящий момент существует множество разных подходов для работы с «большими данными». Одна из важных задач данной области - это задача кластеризации.

Задача кластеризации сводится к распределению неких объектов на группы или классы по определенному признаку.

Кластеризация является актуальной темой для изучения. Механизм распределения объектов на группы используется при решении задач из разных областей: биологии, химии, социологии, экономики, менеджмента и т.д.

Одним из примеров может являться применение кластеризации в автостраховании. На основе информации из базы данных производится разбиение автомобилей и их владельцев на группы, каждый из которых соответствует конкретной рисковому классу. Объекты, которые оказались в

одном классе, будут иметь одинаковую вероятность наступления страхового случая, которая впоследствии и будет оцениваться страховщиком.[1]

Стоит отметить разницу между понятиями кластеризации и классификации. Кластеризация разбивает множество объектов на группы, которые определяются ее результатом. Классификация же относит каждый объект к одной из заранее определенных групп.

Например, биологическая систематика – дисциплина, занимающаяся созданием системы живых организмов, основоположником которой считается Карл Линней, является примером классификации. Живые организмы подразделяются на царства, типы, классы и так далее в зависимости от многих признаков.

Примером задачи кластеризации является, например, выделение сообществ при анализе социальных сетей. Такой анализ может проводиться с целью изучения связей между отдельными страницами или пользователями, определения степени популярности отдельных страниц.

Существует множество алгоритмов, решающих задачу кластеризации разными способами и с разной точностью, однако однозначных рекомендаций, для каких данных какой алгоритм является оптимальным не существует. Поэтому актуальной остается проблема анализа и оптимизации таких алгоритмов.

Целью данной работы является реализация, оценка качества и сравнительный анализ алгоритмов кластеризации k-means, k-means++, c-means, DBScan, Affinity propagation.

В рамках этой цели были поставлены следующие задачи:

- познакомиться с основными понятиями, связанными с проблемой кластеризации;
- рассмотреть алгоритмы k-means, k-means++, c-means, DBScan, Affinity propagation;
- произвести их сравнительный анализ, выявить преимущества, недостатки, определить для каких задач их применение наиболее целесообразно;

- рассмотреть оценки результатов работы алгоритмов кластеризации, отметить особенности их применения;
- реализовать рассмотренные алгоритмы;
- протестировать данные алгоритмы на различных входных данных;
- провести анализ результатов работы алгоритмов с помощью оценок качества кластеризации, сравнить фактическое время выполнения алгоритмов.

Методологические основы кластерного анализа представлены в работах С. Рашки, Ю. Лесковца, А. Н. Ткаченко, И. А. Щербатова, И. О. Беляева, А. Котова, Н. Красильникова, А.С. Герасимовой, Т. Кормена, Е. В. Сивоголовко.

Практическая значимость бакалаврской работы. В ходе выполнения практической части бакалаврской работы были реализованы алгоритмы кластеризации и алгоритмы вычисления оценок качества кластеров, а также проведен сравнительный анализ алгоритмов, после чего были сделаны выводы о преимуществах и недостатках рассматриваемых алгоритмов.

Структура и объём работы. Бакалаврская работа состоит из введения, четырех разделов, заключения, списка использованных источников и трех приложений. Общий объем работы – 62 страницы, из них 45 страниц – основное содержание, включая 32 рисунка и 4 таблицы, список использованных источников информации – 21 наименование.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Задача кластеризации» посвящен основным понятиям и определениям, связанным с кластерным анализом, а также областям применения кластеризации.

Кластеризация – это разбиение определенного множества объектов на непересекающиеся подмножества, называемые кластерами. При этом каждый кластер содержит похожие объекты, а объекты разных кластеров различаются между собой по некоторому признаку.

В задаче кластеризации метрика – это некоторая характеристика, по которой может быть определена схожесть или различие объектов.

Каждому объекту x из множества X можно поставить в соответствие вектор характеристик $x = (x_1, ..., x_m)$, где d – количество характеристик, а x_i – некоторая характеристика объекта.

Кластерный анализ применяется для различных задач из широкого диапазона областей.

Второй раздел «Алгоритмы кластеризации» посвящен описанию рассматриваемых в данной работе алгоритмам кластерного анализа: k-means, k-means++, c-means, DBScan и Affinity propagation.

Алгоритм k-means или алгоритм k-средних является одним из самых популярных алгоритмов кластеризации. Свою известность он получил во многом благодаря своей понятности, относительной простоте реализации и универсальности, то есть способности решать различные типы задач.

Существуют различные варианты модификаций алгоритма k-средних, направленных на улучшение результатов кластеризации. Одной из таких модификаций является алгоритм k-means. Его отличие состоит в выборе начальных центров с использованием вероятностного распределения. Это позволяет выбрать центра максимально далеко друг от друга и обеспечивает более точный результат.

Алгоритм c-means или алгоритм мягких k-средних считается первым созданным методом нечеткой кластеризации и является, пожалуй, самым известным алгоритмом этого класса.

Плотностный алгоритм пространственной кластеризации с присутствием шума или DBSCAN (Density-based spatial clustering of applications with noise) является алгоритмом четкой кластеризации и позволяет находить кластеры произвольной формы в метрическом пространстве. Особенность этого алгоритма состоит в том, что объекты, принадлежащие кластеру, представляются вершинами связного графа.

Метод распространения близости или Affinity propagation – это относительно молодой алгоритм кластерного анализа. Affinity propagation одновременно рассматривает все точки как потенциальные центры кластеров. Алгоритм подразумевает так называемый «обмен сообщениями» между потенциальными центрами и остальными точками.

Третий раздел «Оценки качества кластеризации» посвящен критериям качества результатов алгоритмов кластерного анализа.

Индекс Данна является отношением межкластерного расстояния к диаметру кластера. Следовательно, чем больше значение индекса Данна, тем лучше проведена кластеризация, так как диаметр кластера значительно меньше межкластерного расстояния.

К оценкам качества кластеров также относится так называемая функция кластера или внутрикластерная сумма квадратичных ошибок (SSE).

На основе вычисления функции SSE работает метод локтя или ElbowMethod, который позволяет определить оптимальное число кластеров для предоставленных входных данных.

Метод локтя предлагает построить график функции SSE для разного числа кластеров. Другими словами, на оси абсцисс будут заданы количества кластеров, на которые делятся входные данные, а на оси ординат будут откладываться значения функции SSE.

Таким образом, оптимальное число кластеров можно определить по так называемому «локтю» – точке резкого изменения значения внутрикластерной функции SSE.

Четвертый раздел «Алгоритмы кластеризации» посвящен реализации и тестированию рассмотренных алгоритмов.

Все алгоритмы кластеризации и алгоритмы вычисления оценок качества были реализованы на языке C#. Визуализация результатов работы алгоритмов, а также построение графиков осуществлялась на языке Python 3, в частности, использовалась библиоке matplotlib.

Работа алгоритмов была рассмотрена на выборке из 100 точек и на открытых датасетах кластеризации размером 600, 800 и 3100 точек.

При сравнении времени выполнения алгоритмов было замечено следующее. Алгоритмы k-means, k-means++ и c-means обладают меньшим временем выполнения, в то время как более трудоемкие алгоритмы DBScan и Affinity Propagation требуют значительных временных затрат. Однако, при работе с большим объемом данных, полученный результат является приемлемым и применение алгоритмов DBScan и Affinity Propagation оправдано.

После получения результатов работы алгоритмов на 100 точках можно сделать следующие выводы. Результаты работы алгоритма k-means являются приемлемыми. Однако результат его модификации k-means++ можно визуально определить как более логически верный.

Кластеры, определенные методом c-means, практически не отличаются от полученных алгоритмом k-means++. Таким образом, преимущество нечеткой кластеризации заключается в получении более полной информации об отношении точек к кластерам.

Алгоритмы DBScan и Affinity Propagation самостоятельно определяют число итоговых кластеров. Можно заметить, что метод DBScan разделил исходные объекты на большое количество небольших групп. При этом в самый большой кластер попали отдельно стоящие друг от друга точки.

Метод Affinity Propagation разделил объекты только на два кластера. Так как в качестве центров кластеров алгоритм выбирает наиболее оптимальную для своего окружения точку, то выделение третьего кластера метод определил как нецелесообразное.

После получения результатов работы алгоритмов на датасетах можно сделать следующие выводы. Результаты работы алгоритмов k-means, k-means++ и c-means схожи и имеют примерно одинаковую степень погрешности. Учитывая возможные неточности, такое разбиение может считаться приемлемым решением задачи.

Алгоритм DBScan эффективно работает с кластерами необычной формы. Кроме этого, можно заметить выделение им отдельного кластера для шума, то есть для точек-выборосов. Это является серьезным преимуществом алгоритма, так как определение выбросов относится к возможным задачам кластерного анализа.

Алгоритм Affinity Propagation распределил точки кластеры, имеющие четкие границы и форму, однако степень качества полученного результата может меняться в зависимости от конкретной задачи.

С точки зрения качества кластеров, наиболее высокий результат показали алгоритмы k-means и k-means++. Это связано с тем, что данные алгоритмы выделяют группы правильной формы, которые имеют высокий критерий качества.

ЗАКЛЮЧЕНИЕ

В ходе данной работы были рассмотрены и реализованы алгоритмы кластеризации k-means, k-means++, c-means, DBScan, Affinity Propagation.

Помимо этого были реализованы два алгоритма оценок качества результатов кластеризации: вычисление индекса Данна и внутрикластерной функции квадратичных ошибок.

В результате стравнительного анализа данных алгоритмов были сделаны выводы об особенностях работы каждого метода.

Например, классический алгоритм k-means и его модификации будут эффективны на относительно небольших выборках данных, при условии, что возможно достаточно точно заранее определить количество будущих кластеров. К тому же для данных методов имеет огромное значение форма кластеров и отсутствие точек-выбросов.

Алгоритм DBScan обычно используется для выделения кластеров необычной формы, однако результат будет сильно зависеть от правильности выбора начальных параметров. Чем меньше среднее расстояние между точками, тем сильнее нужно уменьшать задаваемую величину окрестности, в которую будут попадать соседи каждой точки.

Метод Affinity Propagation является самым ресурсозатратным из рассмотренных алгоритмов, так как для его работы необходимо хранить достаточно большое количество информации о точках. Данный алгоритм подходит для задач, в которых необходимо разбить большое количество данных на небольшие, систематизированные группы.

Таким образом, каждый из рассмотренных алгоритмов имеет свой подходящий диапазон задач, для которых применение этих методов будет давать эффективный результат.

По тематике бакалаврской работы был представлен доклад: «Анализ эффективности и оптимизация алгоритма k-средних» на VIII международной конференции «Компьютерные науки и информационные технологии», Саратов, СГУ, июль 2018 г. Доклад опубликован в материалах конференции.