Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.021

Чочиева
Анна Сергеевна

СИСТЕМА РЕКОМЕНДАЦИИ АЛГОРИТМОВ

АВТОРЕФЕРАТ
диссертации на соискание степени
магистра информатики и вычислительной техники

по специальности
1-40 81 01 – Информатика и технологии разработки программного обеспечения

Научный руководитель
Пилецкий И.И.
к.ф.-м.н, доцент

Минск 2020

ВВЕДЕНИЕ

Алгоритмы кластерного анализа позволяют определить группы (кластеры) данных более схожих друг с другом, чем с остальными данными и выявить ранее незамеченные закономерности. Эти алгоритмы относятся к классу задач обучения без учителя. Метки классов в этих алгоритмах заранее неизвестны, в отличии от алгоритмов классификации, которым они иногда предшествуют.

Существует множество разных методов кластеризации данных, каждый со своими преимуществами и недостатками. Целью данной работы является изучение и сравнение нескольких из них и разработка системы, способной давать рекомендацию алгоритма по запросу пользователя.

В данной работе для рассмотрения были взяты методы разбиения (K-means), иерархические методы (агломерационные) и плотностные методы (DBSCAN, OPTICS). Агломерационные методы тестировались с различными методами связи: одиночной, средней, полной и метод минимальной дисперсии Уорда.

Для оценки качества кластеризации использовался силуэтный коэффициент и точность предсказания меток (при наличии меток у набора данных).

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

На данный момент многие компании стремятся внедрить алгоритмы машинного обучения для решения имеющихся проблем. Но выбрать подходящий алгоритм - также является само по себе задачей. Наличие рекомендующей программы сделало бы внедрение алгоритмов машинного обучения более быстрым и, одновременно, эффективным.

Цель и задачи исследования

Цель. Изучение и сравнение нескольких методов кластеризации и разработка системы, способной давать рекомендацию алгоритма по запросу пользователя.

Задача. Создание программного обеспечения, способного рекомендовать пользователям алгоритмы машинного обучения в соответствии с требуемыми параметрами (на пример: планируемый объём данных, требуемая точность и скорость).

В данной работе основные усилия направлены на исследования алгоритмов кластеризации, а именно алгоритмов K-means, OPTICS, Birch и более простые агломерационные алгоритмы.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трёх глав, заключения, списка использованных источников, списка публикаций автора. В первой главе представлен анализ предметной области, рассмотрены различные виды методов кластеризации, сделан кратуий обзор источников литературы. Вторая глава посвящена рассмотрению стека использованных технологий, использованных метрик и общему описанию рабочего процесса разработки рекомендующей программы. В третьей главе описана реализация этапов этого рабочего процесса (сбор данных, тестирование, сбор результатов, моделирование, рекомендация), с рассмотрением фрагментов кода. В конце третьей главы рассматриваются полученные от моделей точности (т.е. качество моделей) и как их можно было бы в перспективе улучшить, и анализируются трёхмерные графики тестовых данных.

Общий объем работы составляет 58 страницы, из которых основного текста – 44 страниц, 9 рисунков на 7 страницах, список использованных источников из 30 и 4 приложения на 13 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе был проведён обзор предметной области и постановка задачи. В пункте 1.1 была поставлена проблема и задача и указано, какие алгоритмы будут рассматриваться. В данной работе основные усилия направлены на исследования алгоритмов кластеризации, а именно алгоритмов K-means, OPTICS и агломерационные алгоритмы с различными методами связи.

В пункте 1.2 был сделан обзор алгоритмов кластеризации. В начале представлено их перечисление и краткое описание, затем более подробно рассмотрен каждый из них. Были рассмотрены:

- Методы разбиения;
- Иерархические методы;
- Плотностные методы;
- Сетевые методы;
- Модельные методы;
- Концептуальная кластеризация.

Наиболее подробно были рассмотрены первые три разновидности методов кластеризации, так как над ними проводились дальнейшие тесты. Для иерархических методов были рассмотрены различные методы связи (одиночная, полная, средняя связь, центроидный метод и метод дисперсии Уорда) и различные меры расстояния.

В остальных пунктах главы были рассмотрены перспективы развития и сделан краткий обзор источников литературы.

Во второй главе были рассмотрены использованные технологии, метрики, и поэтапно описан общий вид рабочего процесса разработки рекомен-

дующей программы. Был рассмотрен язык Python и библиотека scikit-learn, содержащая в себе широкий спектр различных алгоритмов машинного обучения. Также был сделан краткий обзор других использованных Python-пакетов. Работа с кодом велась через IDE PyCharm

Для оценки качества кластеризации были рассмотрены различные метрики:

- инерция (оценка внутри-кластерного расстояния);

- индекс Дюнна (минимум внутри-кластерного и максимум меж-кластерного расстояния);

- силуэтный коэффициент (среднее внутри-кластерное и минимальное среднее меж-кластерное расстояние).

Из них, при тестировании, был выбран силуэтный коэффициент за его наличие в пакете scikit-learn. Индекс Дюнна не включён в пакет scikit-learn, возможно в силу своей медленности и вычислительной сложности.

Для оценки качества кластеризации также использовалась точность назначения меток, при наличии меток у тестируемого набора данных. При этом номер меток не обязательно должен совпадать, а учитывается только принадлежность точек к одному и тому же кластеру.

Далее было проведено поэтапное общее описание рабочего процесса.

Подход к решению. Для решения проблемы подбора оптимального алгоритма необходимо было сделать следующее:

- Провести сбор или генерирование наборов данных для кластеризации;

- Провести тестирование интересующих алгоритмов;

- Создать набор данных из результатов тестирования;

- Провести моделирование (интерполяцию) результатов при других начальных условиях методом регрессии;

- Разработать рекомендательную систему, которая по заданным начальным условиям будет рекомендовать наиболее подходящий алгоритм.

Далее было приведено более подробное описание каждого из этапов.

В третьей главе описывалась реализация вышеупомянутых этапов и рассматривались и анализировались полученные результаты. В пункте 3.1 рассматриваются некоторые из написанных функций генерации данных. В пункте 3.2 рассматриваются сами тесты. В пункте 3.3 рассматривается реализация рекомендующей программы, включая построение моделей для каждого параметра и каждого алгоритма, с подбором регрессоров дающих максимальную точность.

В пункте 3.4 приведены результаты. Выведены оценки точности построенных моделей по коэффициенту детерминации (R-квадрат).

Для некоторых алгоритмов и некоторых их параметров точность прогнозирования весьма низкая. Это может быть объяснено зависимостью значения от некоторого другого признака данных, кроме количества признаков и наблюдений, такого как, например, форма данных. Возможными решениями могут быть:

- добавление нового признака данных, описывающего их форму (нормальность, выпуклость);

- продолжение поиска более подходящего регрессора;

- проведение большего количества тестов (точность модели возрастёт при большем количестве данных).

Далее приводятся графики тестовых данных по каждому параметру (время выполнения, точность предсказания меток, силуэтный коэффициент, использование процессорного времени, использование оперативной памяти с их анализом. По нему можно сделать следующие выводы.

По времени выполнения, при малых и средних (1-1000) количествах признаков и шаровидной форме данных рекомендуется использовать K-means.

При доступе к большому количеству оперативной памяти предпочтительно использовать один из агломерационных методов, наименее точным из которых является метод с одиночной связью, но он также использует несколько меньше оперативной памяти. Они также могут быть несколько быстрее K-means при малых количествах признаков и наблюдений.

Время выполнения алгоритма OPTICS резко растёт по мере роста количества наблюдений (некоторые тесты длились до 27 часов), при этом не потребляя большого количества ресурсов (по наблюдениям во время тестов) и добиваясь хорошего качества кластеризации.

ЗАКЛЮЧЕНИЕ

Для реализации программного обеспечения, способного рекомендовать алгоритмы кластерного анализа, были собраны данные об эффективности выбранных алгоритмов на собранных наборах данных. Для них были написаны тесты производительности и подобраны или сгенерированны различные наборы данных.

Был реализован генератор, способный создавать наборы данных с нормально распределёнными n-мерными кластерами, между которыми регулируется минимальное и максимальное расстояние.

В ходе работы была разработана рекомендующая программа, которая использует собранные от тестов данные для обучения моделей, и с помощью обученных моделей даёт рекомендацию пользователю в соответствии с установленными им приоритетами (скорость, точность и потребление ресурсов).

Для получения более хороших результатов от прогнозов был произведён перебор различных регрессоров и выбраны те, которые давали наиболее высокую точность.

В ходе работы для некоторых алгоритмов с определёнными параметрами была обнаружена весьма низкая точность прогнозирования. Это может быть объяснено зависимостью значения параметра от другого признака данных, кроме количества признаков и наблюдений. Например, такого как форма данных. Возможными решениями могут быть:

добавление нового признака данных; продолжение поиска более подходящего регрессора; проведение большего количества тестов; рассмотрение других методов измерения ресурсов.

Исходя из полученных в процессе работы результатов, при малых и средних (1-1000) количествах признаков и шаровидной форме данных по времени выполнения рекомендуется использовать метод K-means.

При доступе к большому количеству оперативной памяти, исходя из качества кластеризации, предпочтительно использовать один из агломерационных методов. Несколько меньше оперативной памяти необходимо методу с одиночной связью, но он является наименее точным из агломерационных методов. Эти агломерационные методы при малых количествах признаков и наблюдений могут выполняться несколько быстрее метода K-means.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1. Чочиева, А.С. Критерии выбора методов кластеризации / А.С. Чочиева, И.И. Пилецкий // Передовые инновационные разработки. Перспективы и опыт использования, проблемы внедрения в производство. Сборник научных статей по итогам десятой международной научной конференции. – Казань, 2019 – С. 49 – 52.

2. Чочиева, А.С. Выбор алгоритмов кластеризации / А.С. Чочиева, И.И. Пилецкий // Big Data и анализ высокого уровня. Сборник материалов VI международной научно-практической конференции. – Минск, 2020 – С. 281–294.

.