Санкт-Петербургский государственный университет

Кафедра информационно-аналитических систем

Группа 21.Б08-мм

Реализация модуля наивного профилирования в рамках профайлера Desbordante

ГОНЧАРОВ Даниил Юрьевич

Отчёт по учебной практике
в форме «Производственное задание»

Научный руководитель:
ассистент кафедры ИАС Г. А. Чернышев

Санкт-Петербург
2024

Оглавление

Введение

Для анализа и сравнения данных статистики используют различные меры, как-либо характеризующие конкретный набор данных. Самые простые из них: средняя величина (в математической статистике соответствует выборочному математическому ожиданию), максимальное и минимальное значение. Однако, для хоть какого-то серьёзного анализа, необходимо прибегнуть к более показательным величинам, таким как квартили распределения, дисперсия, стандартное отклонение и прочее.

На данный момент в индустрии существует немало инструментов, предоставляющих возможность произвести статистический анализ пользовательских данных. Это говорит о том, что данная сфера не обделена пользовательским вниманием. Но, несмотря на широкий спектр возможностей, статистический анализ — единственное, что может предложить большинство платформ. В то же время существуют профилировщики, позволяющие произвести глубокий анализ данных посредством различных инструментов, например, таких как функциональные зависимости, но при этом не имеют функционала для проведения обычного статистического анализа, одним из которых является Desbordante [2]. Поэтому было принято решение добавить модуль для статистического анализа в эту платформу. Таким образом, будет получен прототип, позволяющий произвести высококачественное исследование данных при помощи статистического анализа, поиска функциональных зависимостей и прочих наукоемких инструментов.

1. Постановка задачи

Целью работы является проектирование и реализация расширяемого класса для подсчёта статистических мер, а также написание методов для вычисления значений средней величины, среднеквадратичного отклонения, коэффициента асимметрии и коэффициента эксцесса. Для её выполнения были поставлены следующие задачи:

1. Разобрать предметную область, написать обзор основных понятий области, а также существующих решений.

2. Спроектировать и реализовать расширяемый класс для сбора статистических мер на основе платформы Desbordante.

3. Реализовать методы для вычисления вышеуказанных статистических мер.

4. Провести экспериментальное исследование реализованной функциональности.

2. Обзор

2.1. Основные понятия

Все нижеследующие определения взяты из книги [9].

Определение 1. Средняя арифметическая величина показывает среднее значение элементов в столбце, однако, подвержена сильному влиянию больших отклонений в данных. Вычисляется по следующей формуле:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Определение 2. Стандартное отклонение характеризует размер вариации: на сколько в среднем каждое значение отличается от средней величины. Для смягчения эффекта смещения при подсчёте используется поправка Бесселя [4]. Вычисляется по следующей формуле:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}.$$

Определение 3. Центральные моменты распределения помогают оценить ряд характеристик распределения случайной величины. Определяются по формуле:

$$\mu_k = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^k}{n}.$$

Определение 4. Коэффициент асимметрии характеризует асимметрию распределения случайной величины. Вычисляется по следующей формуле:

$$skewness = \frac{\mu_3}{\sigma^3}.$$

Определение 5. Коэффициент эксцесса — мера остроты пика распределения случайной величины. Вычисляется по следующей формуле:

$$kurtosis = \frac{\mu_4}{\sigma^4} - 3.$$

2.2. Существующие решения

На данный момент существует большое количество инструментов и библиотек, позволяющих произвести статистический анализ данных тем или иным образом.

Функционал, позволяющий произвести анализ данных, предоставляется индустриальным программным обеспечением, предназначенным для управления данными. Ниже приведены некоторые примеры с описанием возможностей:

- Informatica Data Quality and Profiling позволяет обнаружить возможные ошибки и аномалии в данных пользователя [3]. Также данный инструмент может вычислять процент пустых значений и находить заранее заданные шаблоны.

- Microsoft SQL Server позволяет запустить процесс Microsoft SQL Server Data Profiling Task, который, помимо прочего, найдёт минимальное и максимальное значение, среднюю арифметическую величину и стандартное отклонение для столбцов с числовыми данными [6].

- Talend Open Studio может посчитать количество значений, количество различных и количество повторяющихся значений для столбцов [8]. А для столбцов с числовыми данными позволяет найти среднее значение.

Из приведенной выше информации можно сделать вывод, что профилирование данных в подобных инструментах предоставляет довольно скудный набор статистических величин. Следовательно, данное программное обеспечение не подходит для основательного статистического анализа данных.

Кроме того, существуют решения предназначенные специально для профилирования данных. Тремя популярными и интересными решениями с открытым исходным кодом являются библиотеки для языка программирования Python: Pandas-Profiling, Lux и SweetViz. Ниже перечислены их основные особенности:

- Pandas-Profiling позволяет автоматизировать первичный анализ данных [1]. Пользователю предоставляется необходимая информация для каждого столбца: тип, уникальные значения, количество нулевых и пропущенных значений, квантили и описательная статистика (среднее значение, мода, стандартное отклонение, коэффициенты асимметрии и прочее).

- Lux предоставляет пользователю информацию о базовом статистическом анализе данных в визуальном формате, а также даёт рекомендации по дальнейшему анализу [5].

- SweetViz, в свою очередь, имеет возможность представить статистическую информацию в виде удобной панели с помощью HTML файла [7]. Также благодаря SweetViz можно произвести сравнение разных наборов данных.

Описанные выше библиотеки предоставляют широкий спектр возможностей для проведения статистического анализа. Однако, они не предоставляют инструментов для более глубокого анализа, в частности, инструментов для поиска функциональных зависимостей.

3. Предложенное решение

Было решено сделать класс под названием CsvStats, который будет хранить в себе информацию о таблице при помощи типов данных проекта Desbordante. Данный класс будет содержать в себе методы, которые для указанного столбца могут вычислить следующие статистические меры:

- Сумма всех элементов столбца;
- Среднее значение в столбце;
- Среднеквадратичное отклонение;
- Центральный момент распределения;
- Стандартизированный центральный момент распределения;
- Коэффициент асимметрии;
- Коэффициент эксцесса;
- Количество непустых значений.

За счет того, что будет существовать отдельный метод для вычисления стандартизированного центрального момента распределения, методы по поиску коэффициентов асимметрии и эксцесса будут содержать только вызов одного метода с необходимыми параметрами, что сильно сократит код. Более того, с помощью методов поиска центрального момента распределения и стандартизированного центрального момента распределения, можно будет легко получить значение некоторых других статистических мер благодаря изменению входных данных. Например, второй центральный момент распределения — это не что иное, как дисперсия. Таким образом, на основе существующих методов можно будет реализовывать новые.

Подводя итоги, будет получен расширяемый класс, способный вычислять необходимые статистические меры. Данный класс полностью подходит под поставленную задачу.

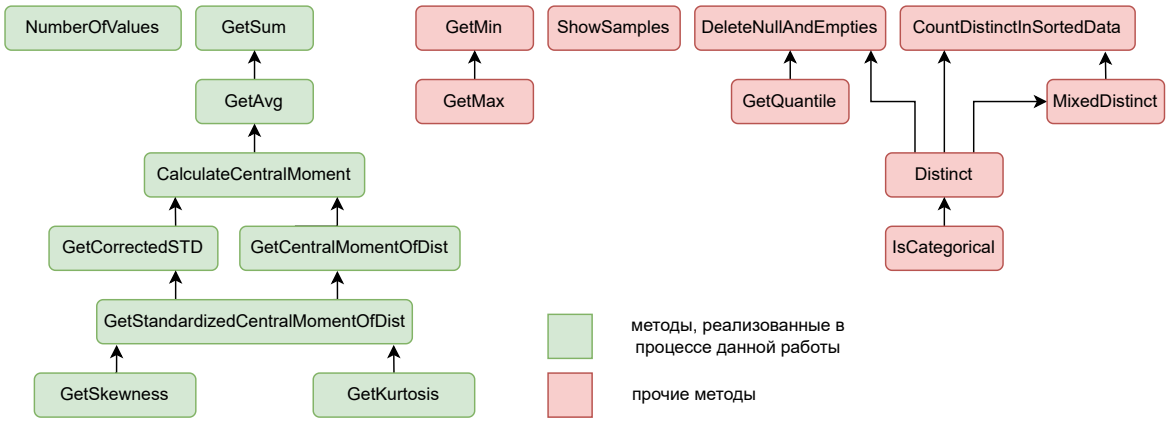
4. Реализация

Класс CsvStats реализовывался в рамках проекта Desbordante на языке C++, придерживаясь Google C++ Style Guide. Так как все необходимые структуры данных для работы с таблицами и типами уже были реализованы, в методах класса достаточно было правильно считать данные нужного столбца и аккуратно произвести математические расчёты по формулам, следя при этом за выделенной под переменные памятью. Таким образом были реализованы следующие методы:

- GetSum (size_t index) const;
- GetAvg (size_t index) const;
- CalculateCentralMoment (size_t index, int number, bool bessel_correction) const;
- GetCorrectedSTD (size_t index) const;
- GetCentralMomentOfDist (size_t index, int number) const;
- GetStandardizedCentralMomentOfDist (size_t index, int number) const;
- GetSkewness (size_t index) const;
- GetKurtosis (size_t index) const;
- NumberOfValues (size_t index) const.

В результате был получен следующий класс:

Ниже представлена схема, показывающая взаимодействие методов класса CsvStats:



На данной схеме отношение стрелка показывает вызов метода.

Для тестов были использованы существующие в проекте наборы данных, а также была создана своя таблица для проверки отдельных ситуаций. Сами тесты реализованы при помощи библиотеки Google

C++ Testing Framework. Благодаря ним можно убедиться, что методы класса CsvStats корректно вычисляют статистические меры.

Реализация данного класса была принята в основной репозиторий GitHub[1].

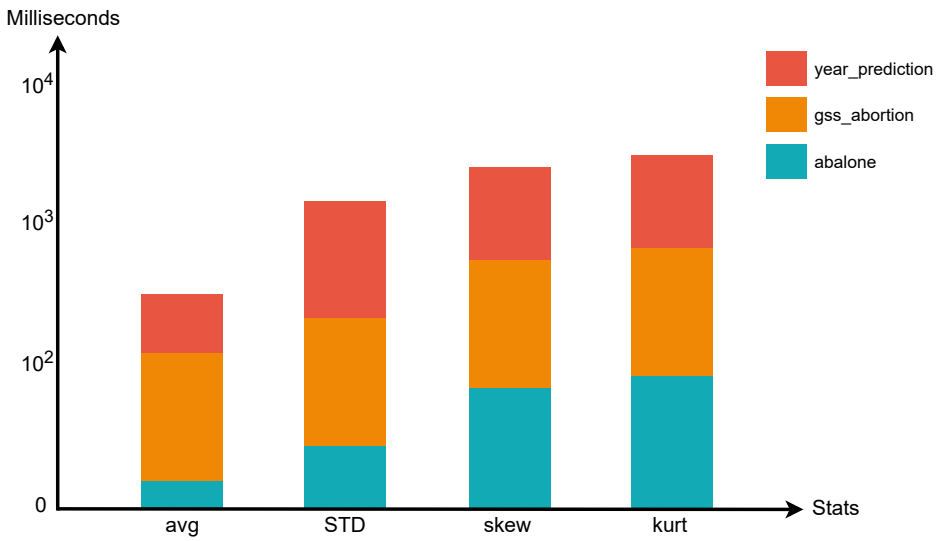
[1] https://github.com/Desbordante/desbordante-core/pull/120

5. Эксперимент

Эксперимент проводился с использованием программы для виртуализации Oracle VM VirtualBox 6.1.36. Характеристики виртуальной машины: AMD Ryzen 7 5800h, 8 GB DDR4 1600MHz RAM; ОС: Ubuntu 21.10, ядро: 5.13.0, версия gcc: 9.4.0. Был использован набор из трёх таблиц:

название	колонки	строки	размер в МБ
abalone	9	4178	0.2
gss_abortion	19	64814	15.4
year_prediction	91	515346	443.4

Был получен следующий результат:



В ходе выполнения эксперимента не было выявлено никаких аномалий: время вычисления зависит от количества данных в таблице, что является ожидаемым поведением.

Заключение

В рамках данной работы были достигнуты следующие результаты:

1. Выполнен обзор статистического анализа и инструментов для его проведения.

2. Спроектирован и реализован расширяемый класс CsvStats на основе проекта Desbordante, благодаря которому можно реализовать сбор статистических мер.

3. Реализованы методы класса CsvStats, вычисляющие некоторые статистические меры.

4. Проведено экспериментальное исследование реализованной функциональности.

Стоит отметить, что класс CsvStats ещё можно развивать. Например, реализовать больше методов для подсчёта различных статистических мер. Но уже в данном виде CsvStats способен выполнять базовый статистический анализ.

Список литературы

[1] Brugman Simon. pandas-profiling: Exploratory Data Analysis for Python. — https://github.com/pandas-profiling/pandas-profiling. — 2019.

[2] Desbordante: a Framework for Exploring Limits of Dependency Discovery Algorithms / Maxim Strutovskiy, Nikita Bobrov, Kirill Smirnov, George A. Chernishev // 29th Conference of Open Innovations Association, FRUCT 2021, Tampere, Finland, May 12-14, 2021. — IEEE, 2021. — P. 344–354. — URL: https://doi.org/10.23919/FRUCT52173.2021.9435469.

[3] Informatica Data Quality Data Discovery Guide. — https://docs.informatica.com/data-engineering/data-engineering-quality/10-2-2/getting-started-guide/getting-started-overview/informatica-developer-overview/data-quality-and-profiling.html. — 2022.

[4] Kenney J.F. Mathematics of Statistics. Mathematics of Statistics no. т. 2. — Van Nostrand, 1946. — P. 124–125. — URL: https://books.google.ru/books?id=UdlLAAAAMAAJ.

[5] Lux: A Python API for Intelligent Visual Discovery. — https://lux-api.readthedocs.io/en/latest/source/getting_started/overview.html. — 2022.

[6] SQL Server Functional Dependency Profile Request Options (Data Profiling Task). — https://learn.microsoft.com/en-us/sql/integration-services/control-flow/data-profiling-task?view=sql-server-ver16. — 2022.

[7] Sweetviz: an open source Python library that generates beautiful, high density visualizations to kickstart EDA (Exploratory Data Analysis) with a single line of code. — https://github.com/fbdesignpro/sweetviz. — 2020.

[8] Talend Open Studio for Data Quality User Guide. — https://help.talend.com/r/w67JnAl9QHSdc9CHFPVrHA/81OI10M6Kv~748vl2xnngw. — 2022.

[9] И. С. Шорохова Н. В. Кисляк О. С. Мариев. Статистические методы анализа : учебное пособие. — Издательство Уральского университета, 2015. — ISBN: 978-5-7996-1633-5. — URL: http://hdl.handle.net/10995/36122.