ClRank 2.0. Руководство пользователя

Программа ClRank 2.0 состоит из двух файлов: `clrank.exe` и `clrank_gamma.exe`.

Программа `clrank.exe` предназначена для вычисления множества максимальных промежутков. Программа `clrank_gamma.exe` предназначена для получения дополнительной информации: значений параметра γ и значений величины $\delta_{\min}$ для нескольких интересующих промежутков.

Программа `clrank.exe` принимает на вход 2 файла: `input.csv` и `input.txt`. Входной файл `input.csv` содержит исходные данные в формате CSV. Входной файл `input.txt` содержит входные параметры для программы. Программа выводит результаты вычислений в файл `output.html` в формате HTML.

Программа `clrank_gamma.exe` принимает на вход 3 файла: `input.csv`, `input.txt` и `input_gamma.txt`. Входной файл `input_gamma.txt` содержит список интересующих промежутков. Программа выводит результаты вычислений в файл `output_gamma.html` в формате HTML.

Входной файл в формате CSV

Входной файл в формате CSV имеет имя `input.csv`.

Записи в файле разделяются символами конца строки (CRLF). После последней записи могут идти и могут не идти символы CRLF.

Все записи должны состоять из трёх полей. Поля записи разделяются символом «точка с запятой» (`;`).

В качестве первой записи должен быть заголовок, состоящий из полей: «Категория», «Частота», «Ранг». Регистр в заголовке важен, кодировка Windows-1251.

В следующих записях поле «Частота» должно содержать вещественное положительное число (разделитель целой и дробной части — запятая (`,`)). Поле «Ранг» должно содержать натуральное число.

Значения поля «Ранг» должны быть 1, 2, 3, …. Значения поля «Частота» должны идти по возрастанию.

Входной файл с входными параметрами

Входной файл с входными параметрами имеет имя `input.txt`.

Этот файл должен содержать три числа: δ_0 ν_0 x.

δ_0 — пороговое значение для половины минимальной высоты полосы, содержащей все точки промежутка — максимально допустимое отклонение точки от прямой (вещественное неотрицательное число).

ν_0 — допустимое количество аномальных точек (целое неотрицательное число).

x — код функциональной зависимости (1 или 2).

Используется функциональная зависимость

$$\ln w \cong -\gamma \ln R + c \equiv -\gamma \ln\left(\frac{N-r}{r}\right) + c\,.$$

Если $x = 1$, то применяется модифицированный В.П. Масловым закон Ципфа I: $N = n + 1$.

Если $x = 2$, то применяется модифицированный В.П. Масловым закон Ципфа II: $N = 2n + 1$.

Входной файл с интересующими промежутками

Входной файл с интересующими промежутками имеет имя `input_gamma.txt`.

Этот файл содержит список промежутков, для которых требуется вычислить значение параметра γ и величину $\delta_{\min}$, в произвольном порядке.

Промежуток — это пара a b, где a и b — это ранги точек. В промежутке должно быть не менее трёх точек.

Выходной файл с множеством максимальных промежутков в формате HTML

Выходной файл с множеством максимальных промежутков имеет имя `output.html`.

Выходной файл содержит таблицу с входными данными и входные параметры.

Далее следует множество максимальных промежутков.

Выходной файл со значениями параметра γ и значениями величины $\delta_{\min}$ в формате HTML

Выходной файл со значениями параметра γ и значениями величины $\delta_{\min}$ имеет имя `output_gamma.html`.

Выходной файл содержит таблицу с входными данными и входные параметры.

Далее следует таблица со значениями параметра γ для каждого из интересующих промежутков при 0, 1, …, ν_0 аномальных точек.

Далее следует таблица со значениями величины $\delta_{\min}$ для каждого из интересующих промежутков при 0, 1, …, ν_0 аномальных точек.

Рекомендации к использованию

Обратите внимание, что точка с рангом 1 находится на графике правее всех, то есть точки на графике идут справа налево.

Программа выдаёт множество максимальных промежутков, а не разбиение, потому что разбиение невозможно определить однозначно. Выбор конкретного разбиения остаётся за пользователем (здесь пользователь может использовать априорную информацию).

Рекомендуется вычислять множество максимальных промежутков при разных значениях параметров. Вначале положите ν_0 равным 0. Затем получите множество максимальных промежутков. Исключите не интересующие промежутки. Можете получить конкретное разбиение данных на кластеры. Далее возьмите ν_0, равное 1. Найдите множество максимальных промежутков. Сравните полученное ранее разбиение с множеством максимальных промежутков. Если кластеры значительно увеличились, то, значит, стоит пересмотреть разбиение, увеличив некоторые кластеры. Если промежутки увеличились на 1, то можно оставить промежутки прежними.

Рекомендуется вычислять значение параметра γ с некоторым количеством аномальных точек, так как небольшое количество точек может существенно повлиять на значение параметра γ.

Пример

Приведём пример применения программы. На рисунке цветом выделено разбиение данных на кластеры, приведённое в статье М.А. Гузева и Е.В. Черныш. Это разбиение было найдено вручную. Фигурными скобками выделено множество максимальных промежутков, выданное программой ($\delta_0 = 0{,}03$, $\nu_0 = 0$).

Отметим, что разбиение, приведённое в данной статье, отличается от множества максимальных промежутков. Таким образом, программа позволяет избавиться от субъективности при нахождении разбиения.