1. Выбор функции статистического вывода (STIR)

STIR переформулирует и слегка корректирует исходную формулу рельефа, включив выборочную дисперсию расстояний до ближайших соседей в оценку важности атрибута. В настоящее время STIR поддерживает бинарную переменную результата, но вскоре будет расширен до многоуровневого и непрерывного результата.

2. Графическая часть

к1

STIR по сравнению с перестановкой на основе рельефа и одномерным t-тестом. Сравнение производительности (истинно отрицательный показатель, точность и полнота) STIR (с несколькими соседями SURF, оранжевый), перестановки на основе рельефа (синий) и одномерного t-критерия (зеленый) для обнаружения функциональных атрибутов. Каждое моделирование повторяется 100 раз с m = 100 образцов и p = 1000 атрибутов со 100 функциональными основными эффектами (A) и сетевыми эффектами взаимодействия (B). Все методы определяют положительные результаты по скорректированному порогу p-значения 0,05 FDR. Каждое моделирование повторяется 100 раз с m = 100 образцов и p = 1000 атрибутов со 100 функциональными (A) основными эффектами (смещение = 0,8) и (B) сетевыми эффектами взаимодействия ()

к2

Влияние на производительность STIR для обнаружения функциональных признаков с основными эффектами (A) и эффектами взаимодействия (B). Сравнение производительности (TrueNegativeRate, Precision и Recall) STIR-ReliefF для нескольких значений ближайших соседей k (k = 5,16,33,49, шкала серого) и STIR-multiSURF (адаптивный радиус, оранжевый). Каждое моделирование повторяется 100 раз с m = 100 образцами и p = 1000 атрибутами со 100 функциональными возможностями. Все методы определяют положительные результаты с использованием скорректированного порога p-значения 0,05 FDR.

псевл

Сравнение псевдокода исходного алгоритма ReliefF, реализованного в ReBATE (Urbanowicz et al., 2018a) (алгоритм 1, слева), и переформулированной версии ReliefF (алгоритм 2, справа). Переформулированная версия позволяет оптимизировать алгоритм путем предварительного вычисления матриц промахов и попаданий (алгоритм 2, строка 7) и использования векторизованной функции сравнения (алгоритм 2, строки 11 и 12). Суммы в строке 13 относятся ко всем элементам Ha и Ma (все пары соседей для всех экземпляров). Псевдокод для STIR работает аналогично.

картинка

Диаграмма рассеяния генов большого депрессивного расстройства с поправкой на значимость для STIR-multiSURF и стандартного t-теста для дифференциальной экспрессии RNA-Seq. STIR-multiSURF находит 32 гена, которые являются значимыми на уровне 0,05 с поправкой на FDR (над горизонтальной пунктирной линией). Стандартный t-критерий находит 8 генов, значимых на уровне 0,05 с поправкой на FDR (справа от вертикальной пунктирной линии). STIR идентифицирует все восемь значимых основных эффектов t-критерия (серый) и дополнительные гены-кандидаты (лиловый), которые могут включать взаимодействия. Из-за перекрытия точек графика помечены не все значимые гены. См. Дополнительный рисунок S2 для подробных этикеток.

примееерл

Мы применяем STIR к исследованию большого депрессивного расстройства (MDD) методом RNA-Seq в Ref. Мостафави и др. (2014). Используя порог FDR 0,05, STIR обнаружил 22 статистически значимые ассоциации из 15 231 гена (таблица 1). Между тем, при том же пороге FDR, равном 0,05, t-критерий Welcht с двумя выборками не выявил каких-либо значимых ассоциаций между уровнями экспрессии генов и MDD (исходное исследование с дополнительными выборками увеличило FDR до 0,25 для обнаружения ассоциаций). Воспроизведение ассоциаций из исходного исследования невозможно, потому что мы сосредоточились на субъектах женского пола, чтобы избежать путаницы (используя 360 MDD и 282 контроля), а методы, основанные на рельефе, в настоящее время не подходят для корректировки ковариат.