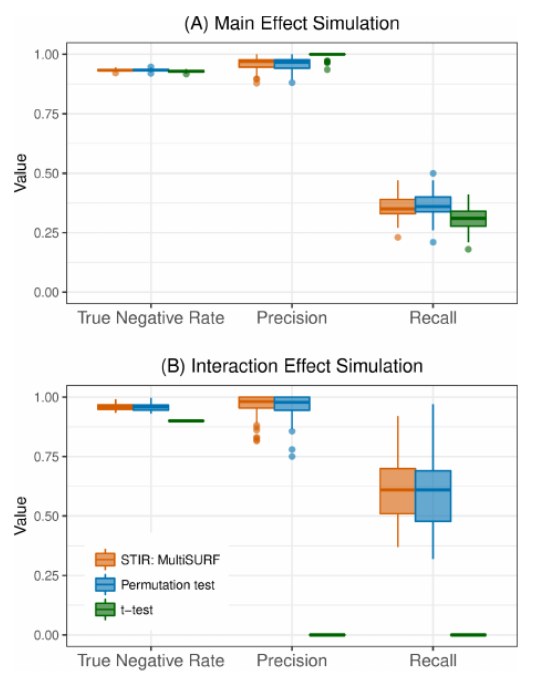
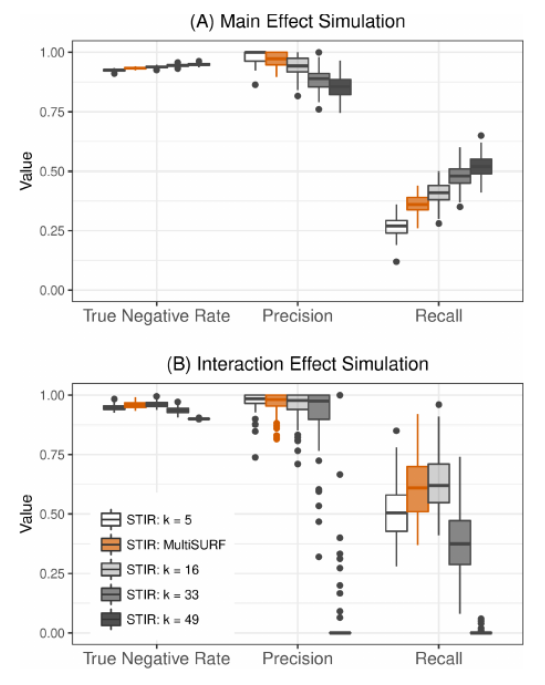
1. Statistical Inference Relief (STIR) feature selection

*STIR reformulates and slightly adjusts the original Relief formula by incorporating sample variance of the nearest neighbor distances into the attribute importance estimation. Currently, STIR supports binary outcome variable but will soon be extended to multi-state and continuous outcome.*

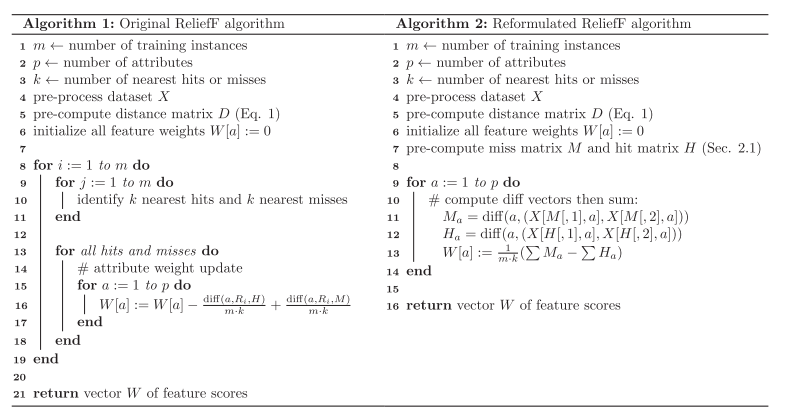
2. Graphical part



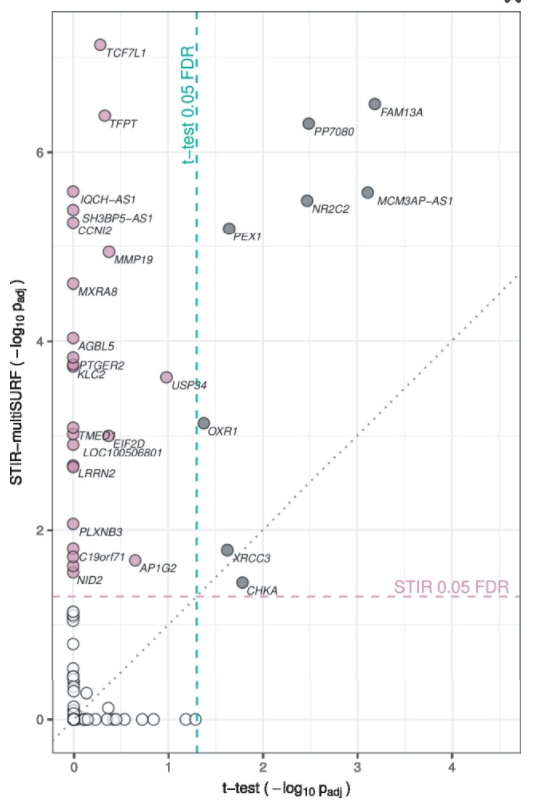
STIR versus Relief-based permutation and univariate t-test. Comparison of theperformance (True Negative Rate, Precision, and Recall) of STIR (with multiSURFneighborhood, orange), Relief-based permutation (blue), and univariate t-test (green) todetect functional attributes. Each simulation is replicated 100 times with m= 100 samplesand p= 1000 attributes with 100 functional main effects (A) and interaction networkeffects (B). All methods determine positives by 0.05 FDR adjusted p-value threshold. Each simulation is replicated 100 times with m = 100 samples and p = 1000 attributes with 100 functional (A) main effects (bias = 0.8) and (B) interaction network effects ()



The effect of kon the performance of STIR to detect functional attributes withmain effects (A) and interaction effects (B). Comparison of the performance (TrueNegativeRate, Precision, and Recall) of STIR-ReliefF for multiple values of nearest neighbors k(k= 5,16,33,49,gray scale) and STIR-multiSURF (adaptive radius, orange). Eachsimulation is replicated 100 times with m= 100 samples and p= 1000 attributeswith 100 functional. All methods determine positives using a 0.05 FDR adjusted p-valuethreshold

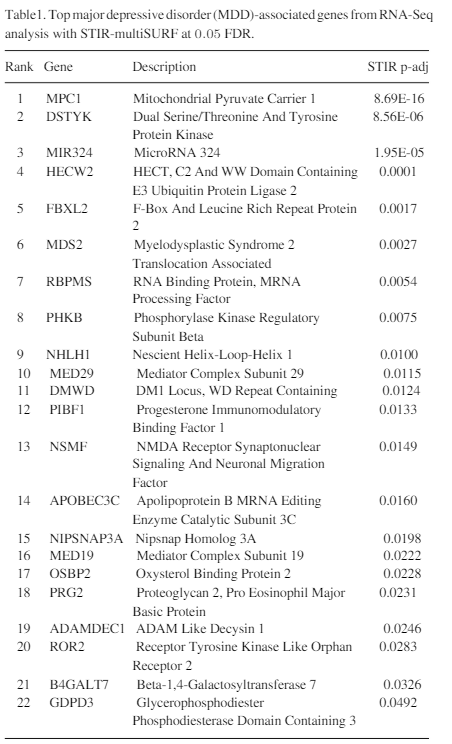


Comparison of the pseudo-code of the original ReliefF algorithm as implemented in ReBATE (Urbanowicz et al., 2018a) (Algorithm 1, left) versus the reformulated version of ReliefF (Algorithm 2, right). The reformulated version allows for algorithm optimization by precomputing miss and hit matrices (Algorithm 2, line 7) and using a vectorized diff function (Algorithm 2, lines 11 and 12). The sums in line 13 are over all elements of Ha and Ma (all pairs of neighbors for all instances). The pseudo-code for STIR works similarly.



Major depressive disorder gene scatter plot of adjusted significance for STIR-multiSURF and standard t-test for RNA-Seq differential expression. STIR-multiSURF finds 32 genes that are significant at the FDR-adjusted 0.05 level (above horizontal dashed line). Standard t-test finds 8 genes that are significant at the FDR-adjusted 0.05 level (to right of vertical dashed line). STIR identifies all eight significant main effects from the t-test (gray) and additional candidate genes (mauve) that may involve interactions. Due to overlap of plot points, not all significant genes are labeled.

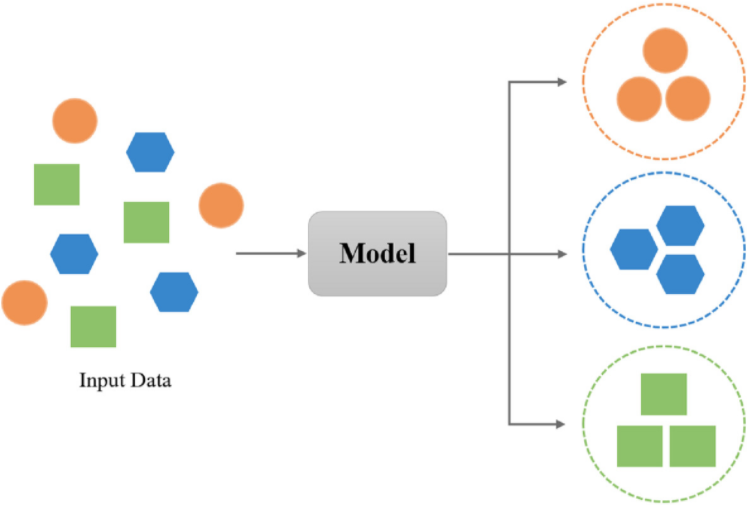
Example



We apply STIR to the RNA-Seq study of major depressivedisorder (MDD) in Ref. Mostafavi et al. (2014). Using an FDR thresholdof 0.05, STIR detected 22 statistically signiﬁcant associations out of 15,231genes (Table1). Meanwhile, with the same FDR threshold of 0.05, a Welchtwo-sample t-test did not identify any signiﬁcant associations between geneexpression levels and MDD (the original study with additional samplesincreased the FDR to 0.25 to detect associations).Reproducing associations from the original study is not feasiblebecause we focused on female subjects to avoid confounding (using360 MDD and 282 controls) and Relief-based methods are currently notadept at correcting for covariates.

3. Term part

– input



– multiSURF as the neighborhood algorithm in STIR

– FDR adjusted P-value threshold a ¼ = 0,05

– Distance between instances - the nearest neighbor distances

– space of all attributes

– Manhattan metric and Euclidean metric

– Attribute averaged over of all pairs of nearest-neighbor misses

– Performance metrics – True Negative Rate (TNR), Precision, and Recall of the statistical tests.

– STIR weight constructed from the Relief difference of means in the numerator and the standard error in the denominator

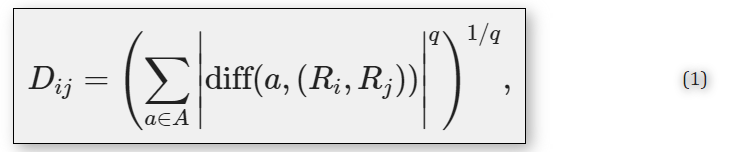
– total number of miss and hit neighbors across all instances

– binary outcome variable

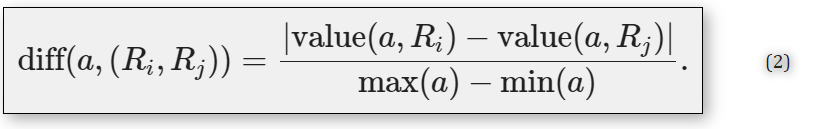
4. Math part

Diff function and nearest neighbor.

The distance between instances Ri and Rj is calculated in the space of all attributes a∈A, typically using a Manhattan (q = 1) metric but may also use a Euclidean (q = 2) metric:

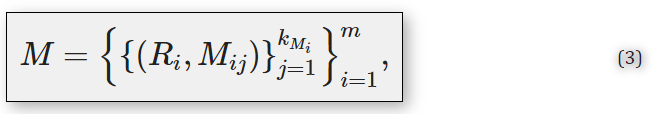


where the standard “diff” function between two instances Ri and Rj for a real-valued attribute a is:

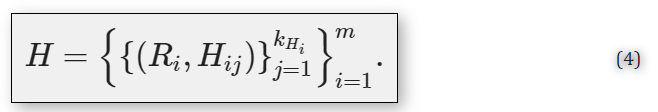


Hit and miss nearest-neighbor ordered pairs.

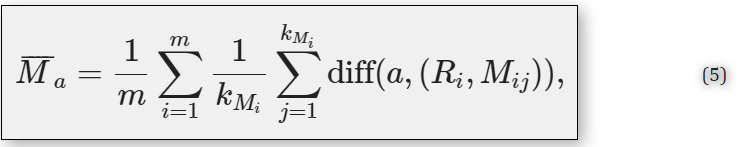
For a given instance Ri (i=1,…,m), a hit is defined as a neighbor instance that has the same class label as that of Ri, and a miss is a neighbor instance with a different class label from Ri. In general Relief-based algorithms, one may represent the set of ordered pairs (Ri,Mij(Ri)), or simply (Ri,Mij), of m instances Ri with their nearest kMi misses, Mij, as nested sets:



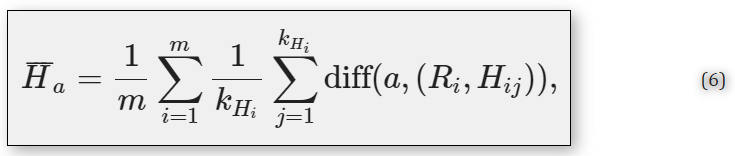
where the index j for the inner set ranges from 1 to kMi, which is the number of nearest miss neighbors for subject Ri. The outer set ranges over all m instances. Similarly for hits, the set of ordered pairs (Ri,Hij(Ri)) of m instances Ri (i=1,…,m) with their kHi nearest hits, Hij, may be written as



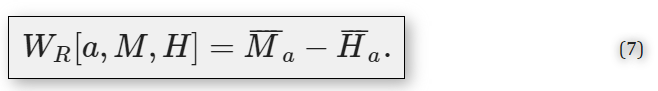
The mean diff for attribute a averaged over of all pairs of nearest-neighbor misses M (Eq. 3) can be expressed as



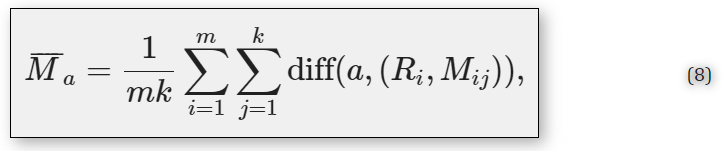
where Mij is the jth nearest neighbor from different classes of the ith instance, Ri, and kMi is the number of nearest miss neighbors of instance Ri. This scaling by 1/kMi inside the sum makes the neighborhood average weighting consistent with multiSURF and with uniform neighborhood methods like SURF and ReliefF. For nearest neighbor hits, the mean is

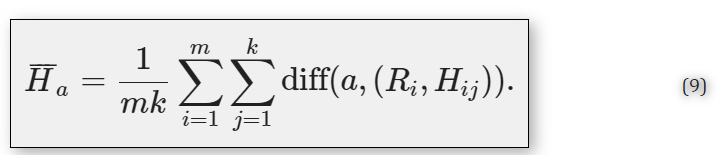


where, similarly, kHi is the number of nearest hit neighbors of instance Ri. The Relief-based importance score can then be expressed simply as

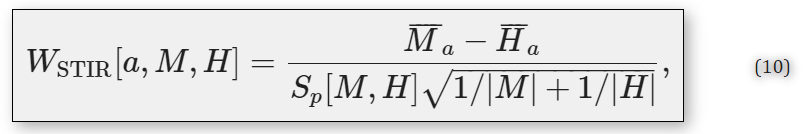


In the case of Relief-based methods with constant k (ReliefF), we have kMi=kHi=k ∀i, and Eqs. (5) and (6) become

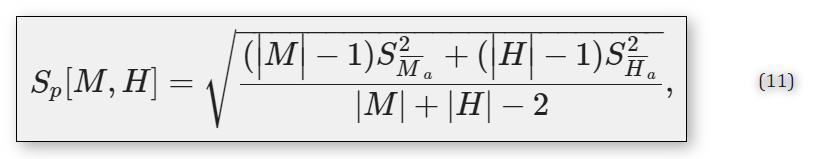


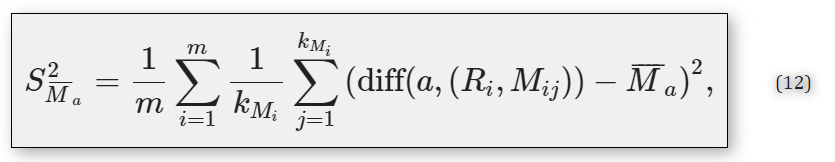


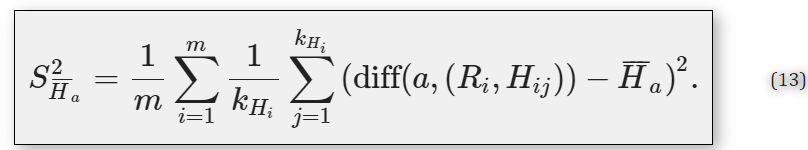
For attribute a, we construct the following STIR weight (or STIR score) from the Relief difference of means (WR in Eq. 7) in the numerator and the standard error in the denominator:



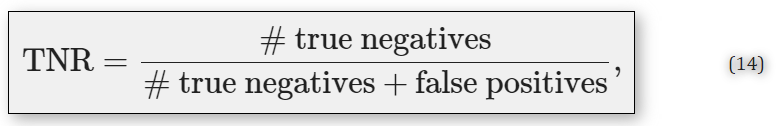
where |M|=∑mi=1kMi and |H|=∑mi=1kHi are the total number of miss and hit neighbors across all instances. The pooled standard deviation is

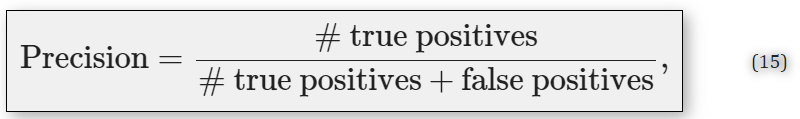


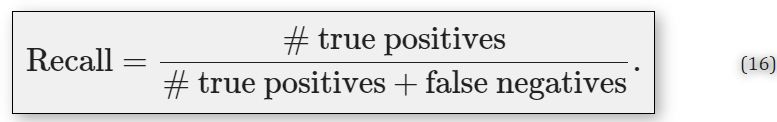




 Performance metrics – True Negative Rate (TNR), Precision, and Recall of the statistical tests.







**Джордж Буль**

****

**2 ноября 1815 – 8 декабря 1864**

– Буль работал помощником учителя в Донкастере, а также недолгое время преподавал в Ливерпуле. Некоторое время он был связан с открывшимся в 1833 году институтом механики Линкольна.

– В 1834 году он открыл свою школу в Линкольне.В течение этого времени он много времени уделял социальной работе и образованию взрослых.

– Основал «Приют раскаявшихся женщин». С целью образования неимущих, Буль также работал в институте механики. Через четыре года Буль стал владельцем «Hall’s Academy» в Уоддингтоне, под Линкольном.

– В 1839 году он представил несколько работ, среди которых были «Теория математических преобразований» для «Кембриджского математического журнала». В этих работах речь шла о дифференциальных уравнениях и алгебраической проблеме линейной трансформации путём выделения идеи инвариантной линейной трансформации через выделение идеи инвариантности.

– В 1840 году он вернулся в Линкольн для руководства закрытой школой.

– В 1841 году он открыл теорию инвариантов – новый раздел математики. Этот раздел математики впоследствии был источником вдохновения Эйнштейна.

– В 1844 году он анализировал комбинированные методы алгебры и исчислений в публикации с названием «Философские труды королевского общества».

– В 1847 году, совместно с Э. Р. Ларкеном, он основал жилищно-строительное общество. В том же году в памфлете «Математический анализ логики» он высказал мнение, что логика должна быть связана с математикой. Инновационный вклад Буля в математику был по-настоящему эффективен при создании цифрового компьютера и электронных схем. В работе «Математический анализ логики» изложил основы булевой алгебры. Разработал алфавит, орфографию и грамматику. Вычисление истинности или ложности рассуждений, записанных с помощью специальных знаков, – основная задача созданной Булем алгебры логики или, как её чаще называют булевой алгебры.

– Развитие идей Буля привело к созданию современной математической логики, которая включает в себя алгебру множеств, алгебру высказываний, алгебру релейных схем (реле – это переключатель в электрических схемах), без которых было бы невозможным проектирование и программирование вычислительных машин. Именно булева алгебра лежит в основе работы компьютера.

– В 1849 году он стал первым профессором математики в Королевском колледже в Корке, Ирландия.

– В 1854 году он занимался алгеброй и логикой, и его труды в этой области более известны как булева алгебра (алгебра логики). В том же году он ввёл понятие символический метод логического вывода в публикации «Законы мысли».

– Булева алгебра служит в качестве основ анализа обоснованности логических суждений, так как она носит бинарный характер утверждений, которые могут оказаться либо положительными, либо ложными. Метод бинарности и логические элементы булевой логики используются в телефонной коммутации и в электронных компьютерах во время их создания и работы. Во второй части «Закона мысли» Буль пытался открыть общий метод в вычислении вероятностей.

– В 1857 году Буль представил публикацию «О сравнении трансцендентных функций» с определёнными наложениями на теорию определённых интегралов. В публикации он изучает сумму остатков рациональной функции. А частью изучения стало доказательство булева тождества.

– В 1859 году Буль публикует «Трактат по дифференциальным уравнениям», в котором он сообщает об общем символическом методе;

– В 1860 году он публикует продолжение с названием «Трактат об исчислении конечных разностей».