Quick Permutation Test (QuiPT)

Piotr Sobczyk[1], Michał Burdukiewicz[2]

[1]Wrocław University of Technology, Institute of Mathematics and Computer Science, Poland

[2]University of Wrocław, Department of Genomics, Poland

Outline

Outline

n-grams (k-tuples) are sets of n characters derived from the input sequence(s). They may form continuous sub-sequences or be discontinuous.

	X1	X2	X3	X4	X5	X6
1	1	2	1	4	2	1
2	3	1	4	3	3	2
3	4	1	3	3	4	2

Sample sequences.

1	2	3	4
3	2	0	1
1	1	3	1
1	1	2	2

Unigrams.

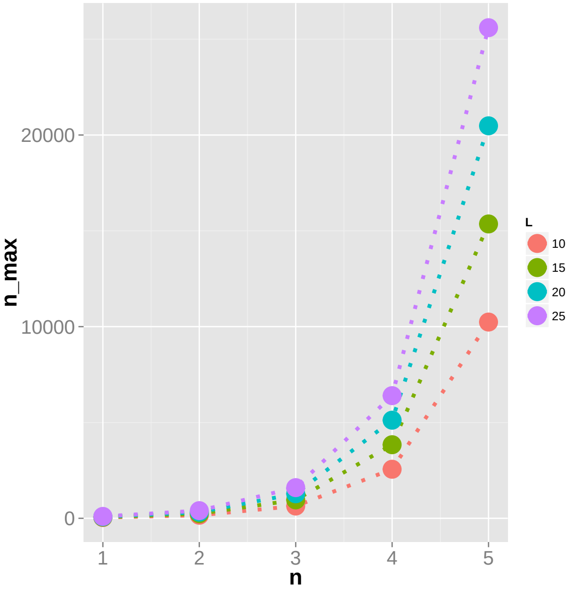
X1_1_0	X2_1_0	X3_1_0	X4_1_0	X5_1_0	X6_1_0	X1_2_0
1	0	1	0	0	1	0
0	1	0	0	0	0	0
0	1	0	0	0	0	0

A fraction of possible unigrams with position information.

Positioned n-gram data is binary.

Number of possible positioned n-grams:

$$n_{max} = L \times m^n$$


20000
10000
0
n_max
1
2
3
4
5
n
L
10
15
20
25

Outline

1. Calculate test statistic for the given positioned n-gram and etiquettes (T_R).

1. Calculate test statistic for the given positioned n-gram and etiquettes (T_R).
2. Permute counts of n-grams and calculate permuted test statistic (T_P).

1. Calculate test statistic for the given positioned n-gram and etiquettes (T_R).
2. Permute counts of n-grams and calculate permuted test statistic (T_P).
3. Repeat step 2. N times.

1. Calculate test statistic for the given positioned n-gram and etiquettes (T_R).
2. Permute counts of n-grams and calculate permuted test statistic (T_P).
3. Repeat step 2. N times.
4. Calculate p-value using:

$$\text{p-value} = \frac{N_{T_P > T_R}}{N}$$

$N_{T_P > T_R}$ is number of times when T_P was bigger than T_R

- Model independent.

- Model independent.
- Statistic independent.

- Computationally expensive (number of cases, number of features).

- Computationally expensive (number of cases, number of features).
- Single feature analysis (no feature interaction).

- Computationally expensive (number of cases, number of features).
- Single feature analysis (no feature interaction).
- Unfeasible precise estimation of low p-values.

The number of permutations is inversely proportional to the interval between p-values.

Example: with 10×10^6 permutation the smallest possible p-values are: 0, 1×10^{-6}, 2×10^{-6} and so on.

Outline

The binary positioned n-gram data tabulated by binary label can be easily described in 2d contingency table.

sequence ID	feature	target
1	1	0
2	1	0
3	0	0
4	1	1
5	0	1
...	...	...

Positioned n-grams with a label.

	target	feature
0	$n_{1,1}$	$n_{1,0}$
1	$n_{0,1}$	$n_{0,0}$

Contingency table.

Test statistics used by QuiPT (information gain, Kullback-Leibler divergence) measure inbalance of contingency tables.

If probability that target equals 1 is p and probability that feature equals 1 is q and feature and target are independent then each of them has the following probabilities

$$P(Target, Feature) = (1, 1)) = p \cdot q$$

$$P(Target, Feature) = (1, 0)) = p \cdot (1 - q)$$
$$P(Target, Feature) = (0, 1)) = (1 - p) \cdot q$$
$$P(Target, Feature) = (0, 0)) = (1 - p) \cdot (1 - q)$$

$$
\begin{aligned}
F(n_{1,1}, n_{1,0}, n_{0,1}, n_{0,0}) = \binom{n}{n_{1,1}} (p \cdot q)^{n_{1,1}} \binom{n - n_{1,1}}{n_{1,0}} (p \cdot (1-q))^{n_{1,0}} \\
\binom{n - n_{1,1} - n_{1,0}}{n_{0,1}} ((1-p) \cdot q)^{n_{0,1}} \\
\binom{n - n_{1,1} - n_{1,0} - n_{0,1}}{n_{0,0}} ((1-p) \cdot (1-q))^{n_{0,0}}
\end{aligned}
$$

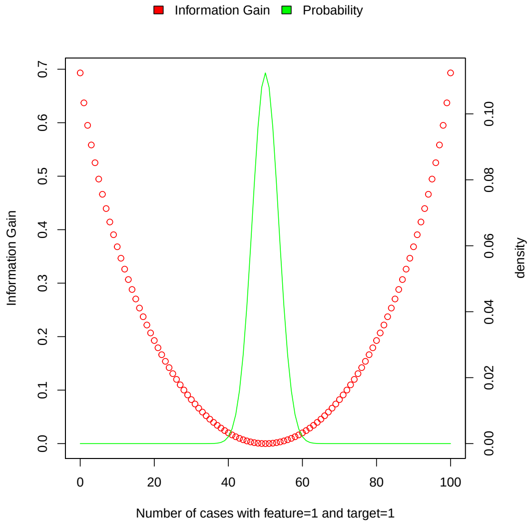
In addition to this: $n_{1,\cdot} = n_{1,1} + n_{1,0}$ and $n_{\cdot,1} = n_{1,1} + n_{0,1}$ are known and fixed.

- $n_{1,1}$ is from range $[0, min(n_{\cdot,1}, n_{1,\cdot})]$.

- $n_{1,1}$ is from range $[0, min(n_{\cdot,1}, n_{1,\cdot})]$.
- The probability of certain contingency table is given as the conditional distribution, as impose restrictions on two parameters $n_{\cdot,1}$ and $n_{1,\cdot}$.

- $n_{1,1}$ is from range $[0, min(n_{\cdot,1}, n_{1,\cdot})]$.
- The probability of certain contingency table is given as the conditional distribution, as impose restrictions on two parameters $n_{\cdot,1}$ and $n_{1,\cdot}$.
- The test statistic is computed for each possible value of $n_{1,1}$.

- $n_{1,1}$ is from range $[0, min(n_{\cdot,1}, n_{1,\cdot})]$.
- The probability of certain contingency table is given as the conditional distribution, as impose restrictions on two parameters $n_{\cdot,1}$ and $n_{1,\cdot}$.
- The test statistic is computed for each possible value of $n_{1,1}$.
- The distribution of test statistics under hypothesis that target and feature are independant is computed using values from 3.



	Target	Feature	Freq
1	0	0	50
2	1	0	50
3	0	1	50
4	1	1	50

- QuiPT is faster.

- QuiPT is faster.
- Using the exact distribution of possible values of the criterion QuiPT yields precise small p-values without increasing the computation time.

Outline

1. n-grams
   - n-gram definition
   - Positioned n-grams
2. Permutation test
   - Testing framework
   - Advantages
   - Drawbacks
   - p-value resolution
3. QuiPT
   - Contingency tables
   - Multinomial distribution of target-feature relationship
   - Advantages over permutation test
4. Simulation scheme
   - Power of the test
   - False significant features
5. Conclusion

1. Random 4000 sequences (20 nucleotides each). The half of the sequences has label 0.

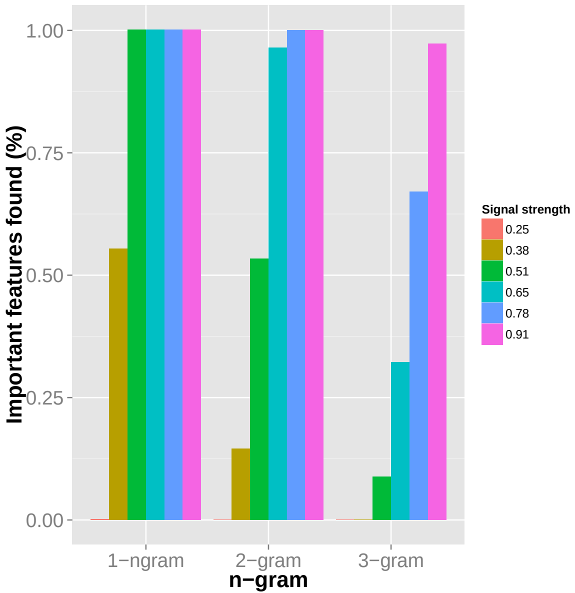
1. Random 4000 sequences (20 nucleotides each). The half of the sequences has label 0.
2. Choose a single position between 3 and 18 (to avoid border cases).

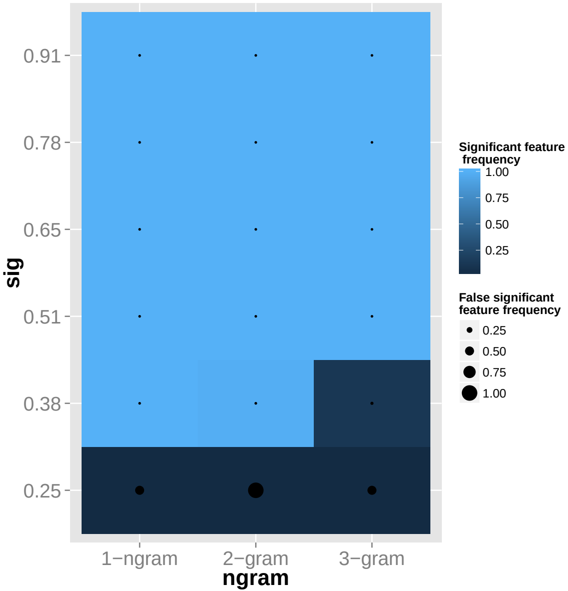
1. Random 4000 sequences (20 nucleotides each). The half of the sequences has label 0.
2. Choose a single position between 3 and 18 (to avoid border cases).
3. Resample nucleotides at chosen position. The dominant nucletoide has probabilitiy of occurence $p_d = 0.25$. Other nucleotides have probability of occurence $p_o = (1 - p_d)/3$.

1. Random 4000 sequences (20 nucleotides each). The half of the sequences has label 0.
2. Choose a single position between 3 and 18 (to avoid border cases).
3. Resample nucleotides at chosen position. The dominant nucletoide has probabilitiy of occurence $p_d = 0.25$. Other nucleotides have probability of occurence $p_o = (1 - p_d)/3$.
4. Perform QuiPT (Information Gain) and choose significant features (with p-value < 0.001).

1. Random 4000 sequences (20 nucleotides each). The half of the sequences has label 0.
2. Choose a single position between 3 and 18 (to avoid border cases).
3. Resample nucleotides at chosen position. The dominant nucletoide has probabilitiy of occurence $p_d = 0.25$. Other nucleotides have probability of occurence $p_o = (1 - p_d)/3$.
4. Perform QuiPT (Information Gain) and choose significant features (with p-value < 0.001).
5. Iterate steps 1-4 over other values of p_d - 0.38, 0.51, 0.65, 0.78, 0.91.

1. Random 4000 sequences (20 nucleotides each). The half of the sequences has label 0.
2. Choose a single position between 3 and 18 (to avoid border cases).
3. Resample nucleotides at chosen position. The dominant nucletoide has probabilitiy of occurence $p_d = 0.25$. Other nucleotides have probability of occurence $p_o = (1 - p_d)/3$.
4. Perform QuiPT (Information Gain) and choose significant features (with p-value < 0.001).
5. Iterate steps 1-4 over other values of p_d - 0.38, 0.51, 0.65, 0.78, 0.91.
6. Repeat steps 1-5 200 times.


1.00
0.75
0.50
0.25
0.00
Important features found (%)
1−ngram
2−gram
3−gram
n−gram
Signal strength
0.25
0.38
0.51
0.65
0.78
0.91


0.91
0.78
0.65
0.51
0.38
0.25
sig
1−ngram
2−gram
3−gram
ngram
Significant feature frequency
1.00
0.75
0.50
0.25
False significant feature frequency
0.25
0.50
0.75
1.00

Outline

Quick permutation test is a powerful and quick equivalent of permutation test in binary feature-binary target testing scenario.