Quick Permutation Test: feature filtering of n-gram data

Piotr Sobczyk[1]*, Michał Burdukiewicz[2], Chris Lauber[3], Paweł Mackiewicz[2]
*Piotr.Sobczyk@pwr.edu.pl

[1]Wrocław University of Technology, Department of Mathematics, Poland
[2]University of Wrocław, Department of Genomics, Poland
[3]Dresden University of Technology, Institute of Medical Informatics and Biometry, Poland

Introduction

N-grams (k-tuples) are vectors of n characters derived from input sequence(s). They may form continuous sub-sequences or be discontinuous. Another important n-gram parameter is its position. Instead of just counting n-grams, one may want to count how many n-grams occur at a given position in multiple (e.g. related) sequences.

	P1	P2	P3	P4	P5	P6
S1	4	3	1	1	2	3
S2	4	4	3	3	4	4
S3	3	1	1	4	4	2

Sample sequences.

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

Unigram counts.

P1_1	P2_1	P3_1	P4_1	P5_1	P6_1	P1_2	P2_2	P3_2	P4_2	P5_2	P6_2	P1_3
0	0	1	1	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	0	0	0	0	1	1

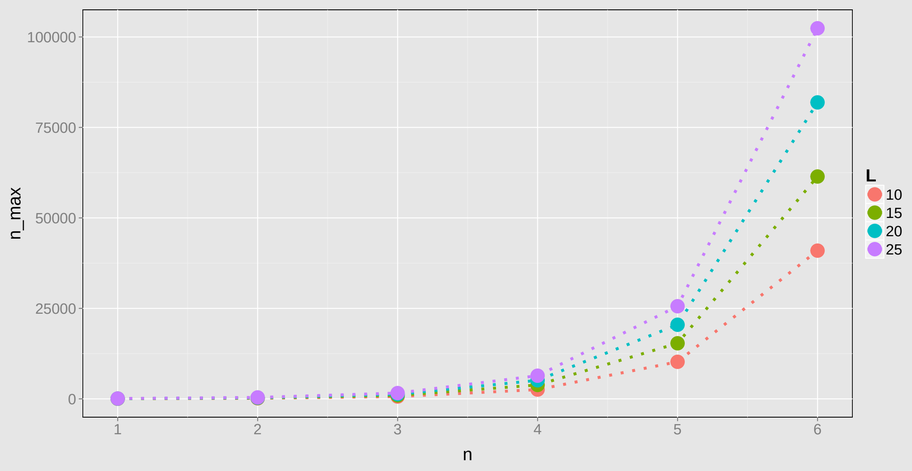
A fraction of possible unigrams with position information.

Originally developed for natural language processing, n-grams are also used in proteomics, genomics and transcriptomics.

Curse of dimensionality

Even when we limit ourselves to only continuous positioned n-grams, feature space growths rapidly with the number of elements in n-gram ($\boldsymbol{n}$) and length of the sequence ($\boldsymbol{L}$).
Number of possible positioned n-grams:

$$\boldsymbol{n}_{\text{max}} = \boldsymbol{L} \times \boldsymbol{m}^{\boldsymbol{n}}$$



Feature selecting permutation tests

Model and statistic independent permutation tests can be used to filter features obtained through counting n-grams.
During a permutation test class labels are randomly exchanged during computation of significance statistic. p-values are defined as:

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

where $\boldsymbol{N}_{\boldsymbol{T_P} > \boldsymbol{T_R}}$ is number of times when $\boldsymbol{T_P}$ (permuted test statistic) was more extreme than $\boldsymbol{T_R}$ (test statistic for non-permuted data).
Permutation tests are computationally expensive (especially precise estimation of low p-values, because the number of permutations is inversely proportional to the interval between p-values).

QuiPT algorithm

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

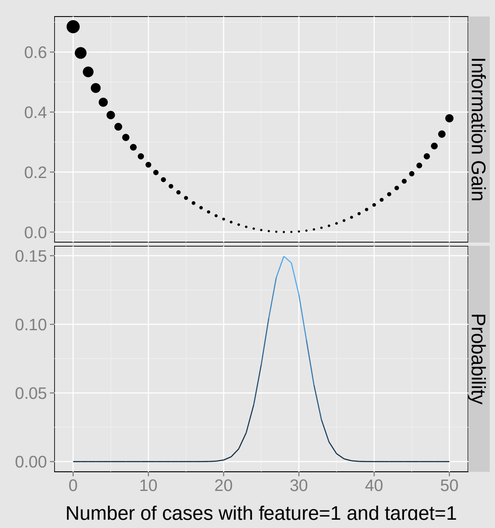
$$\boldsymbol{P}(\boldsymbol{Target}, \boldsymbol{Feature}) = (1,1)) = \boldsymbol{p} \cdot \boldsymbol{q}$$
$$\boldsymbol{P}(\boldsymbol{Target}, \boldsymbol{Feature}) = (1,0)) = \boldsymbol{p} \cdot (1 - \boldsymbol{q})$$
$$\boldsymbol{P}(\boldsymbol{Target}, \boldsymbol{Feature}) = (0,1)) = (1 - \boldsymbol{p})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

So, this shows that what we actually get is a contingency table that needs to tested for independance.

Independence test

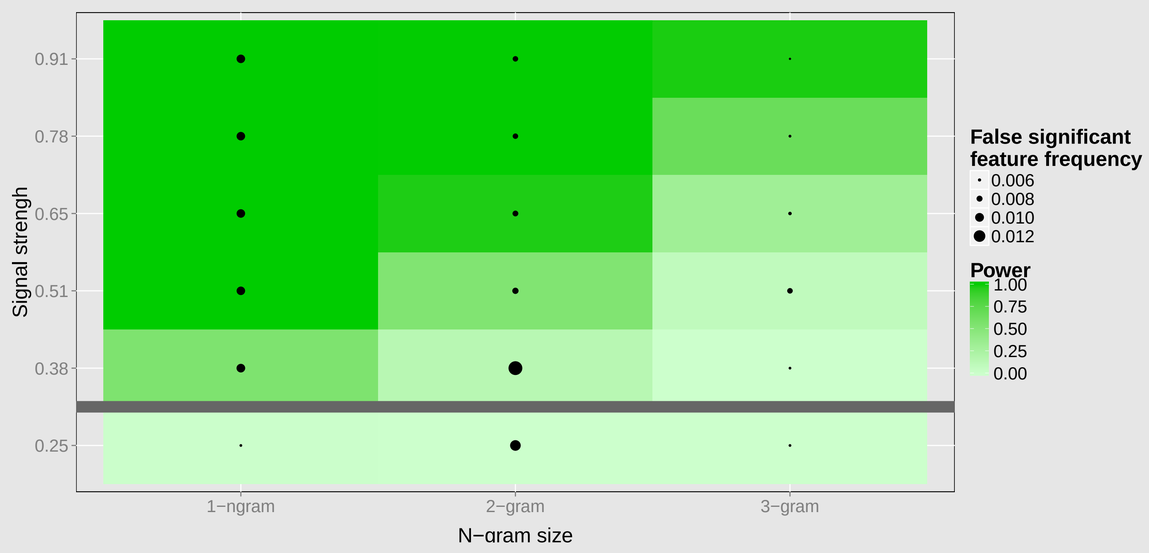
$\mathrm{F}(\mathrm{n}_{1,1}, \boldsymbol{n}_{1,0}, \boldsymbol{n}_{0,1}, \boldsymbol{n}_{0,0}) = \binom{n}{n_{1,1}} (\boldsymbol{p} \cdot \boldsymbol{q})^{n_{1,1}} \binom{n-n_{1,1}}{n_{1,0}} (\boldsymbol{p} \cdot (1 - \boldsymbol{q}))^{n_{1,0}}$
$\binom{n-n_{1,1}-n_{1,0}}{n_{0,1}} ((1 - \boldsymbol{p}) \cdot \boldsymbol{q})^{n_{0,1}} \binom{n-n_{1,1}-n_{1,0}-n_{0,1}}{n_{0,0}} ((1 - \boldsymbol{p}) \cdot (1 - \boldsymbol{q}))^{n_{0,0}}$
Our data gives constraints: $\boldsymbol{n}_{1,\cdot} = \boldsymbol{n}_{1,1} + \boldsymbol{n}_{1,0}$ and $\boldsymbol{n}_{\cdot,1} = \boldsymbol{n}_{1,1} + \boldsymbol{n}_{0,1}$. Thus, conditioning on $\boldsymbol{n}_{1,1}$ we get hypergeometric distribution.
This is in fact exact two-sided Fisher's test. Information Gain is used here as a way of deciding which contingency tables are more extreme.

Validation procedure



	Target	Feature	Freq
1	0	0	40
2	1	0	10
3	0	1	25
4	1	1	40

Power and False discoveries



Summary

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

Avaibility

biogram R package:
`http://cran.r-project.org/web/packages/biogram/`

Bibliography