Statistical significance of Delta-based authorship attribution

Stefan Evert

15 May 2019

Some preliminaries

Data sets

Load relative frequencies and z-scores for the German, English and French data set. For technical reasons, the data structures store the transposed document-term matrices $\mathbf{F}^T$ and $\mathbf{Z}^T$

```
load("data/delta_corpus.rda")
## FreqDE, FreqEN, FreqFR ... text-word matrix with absolute and relative frequencies
## zDE, zEN, zFR          ... standardized (z-transformed) relative frequencies
## goldDE, goldEN, goldFR ... gold standard labels (= author names)
```

- $\mathbf{F}^T$ is available under the names `FreqDE$S`, `FreqEN$S` and `FreqFR$S`
- $\mathbf{Z}^T$ is available under the names `zDE`, `zEN` and `zFR`
- absolute frequencies $n_{D_j} \cdot f_i(D_j)$ can be found in `FreqDE$M`, `FreqEN$M`, `FreqFR$M`

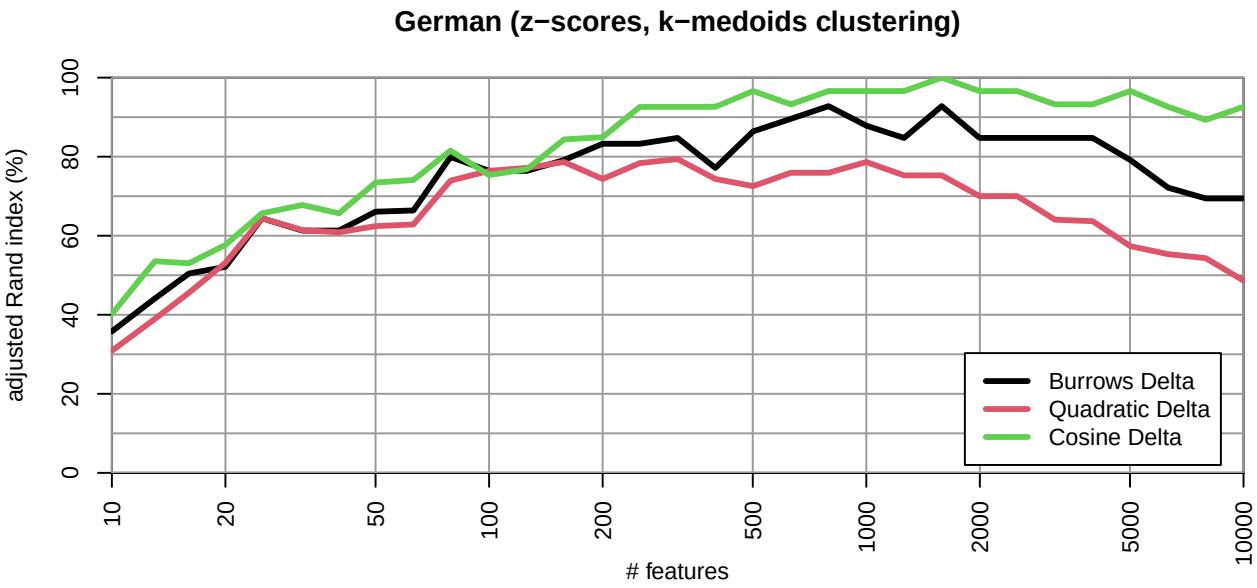
Setup for plots

```
n.vals <- round(10 ^ seq(1, 4, .1)) # logarithmic steps
draw.grid <- function () { # corresponding grid for plot region
  abline(h=seq(0, 100, 10), col="grey60")
  abline(v=c(10,20,50,100,200,500,1000,2000,5000,10000), col="grey60")
}
```

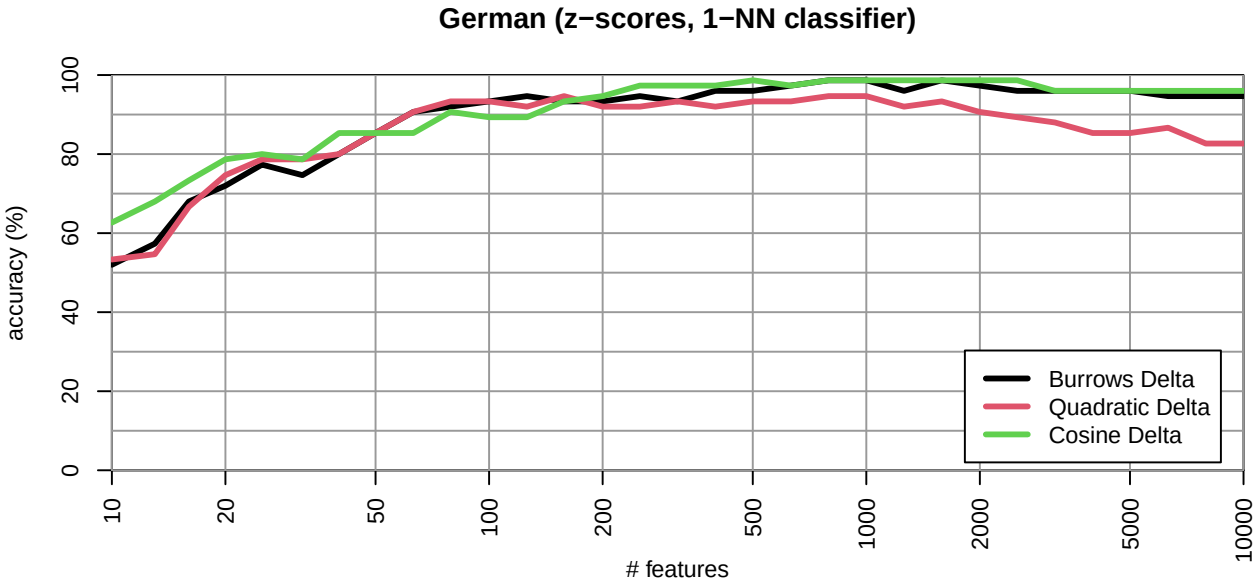
Cross-validation

The standard approach in supervised classification tasks is to estimate sampling variation from cross-validation runs. In a leave-one-out setting, this amounts to treating the test items as a random sample ($\rightarrow$ binomial confidence interval) and ignoring variability due to training data and the choice of authors.

Let us start with a standard evaluation plot:

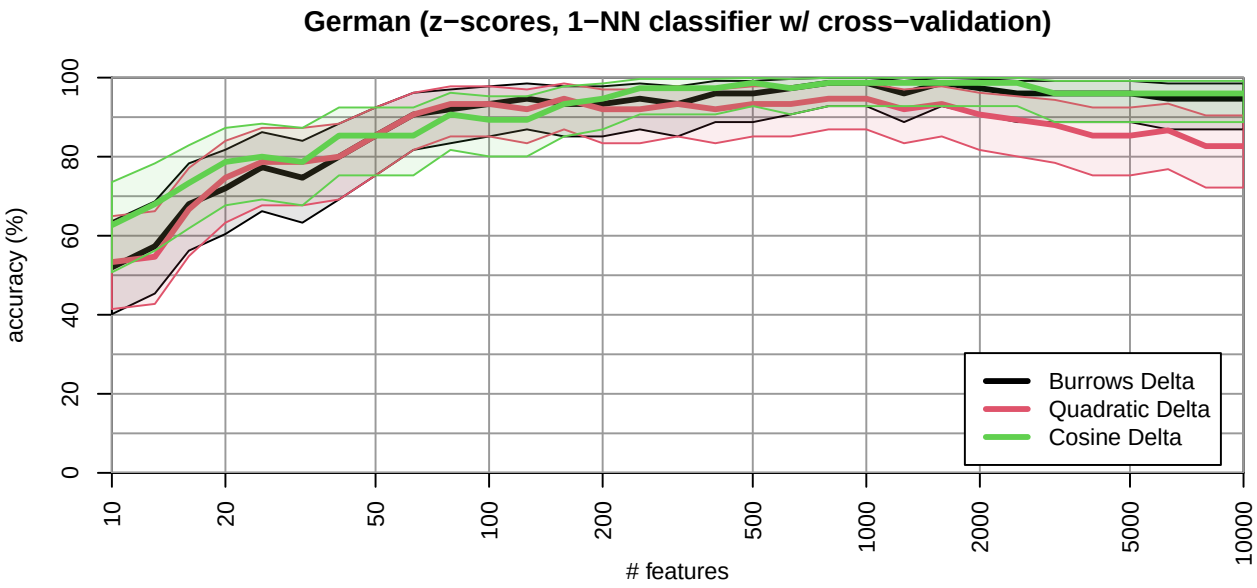


For the cross-validation approach, we re-evaluated based on the accuracy of a 1-NN classifier:



We now define a function that re-constructs the number of correct assignments and generates binomial confidence intervals.

```
plot.cv <- function (x, acc, n=75, col=1, alpha=.1) {
  correct <- round(n * acc / 100)
  cint <- prop.cint(correct, n, method="binomial")
  bg.col <- adjustcolor(palette()[col], alpha=alpha)
  polygon(c(x, rev(x)), 100 * c(cint$upper, rev(cint$lower)),
          border=col, col=bg.col)
  lines(x, acc, lwd=3, col=col)
}
```


German (z−scores, 1−NN classifier w/ cross−validation)
accuracy (%)
0
20
40
60
80
100
10
20
50
100
200
500
1000
2000
5000
10000
features
Burrows Delta
Quadratic Delta
Cosine Delta

References