CWI

CWI - Information Access Group

PERLE

a method for Misclassification-Aware Data Analysis

Emma Beauxis-Aussalet, Arjen de Vries, Lynda Hardman emma@cwi.nl, lynda.hardman@cwi.nl

Classifiers & Counting Tasks

Although imperfect, classifiers are key assets for mining data, including cases where uncertainty is critical (e.g., e-Science, medicine). They can be used to count, e.g., people, cells, species of animals, behaviours or pixel coverage. Our use case is the core task of counting items over classes, and comparing counts over time, locations or other conditions.

Our Novel Method: PERLE

With **P**airwise **E**rror **R**ates (1) and **L**inear **E**quations (2), the end-results counts are corrected by solving a linear system.

Error Rate in Test Set / Error Rate in End-Results

$$\frac{n_{ik}}{n_{i.}} = \frac{n'_{ik}}{n'_{i.}}$$

$$n'_{ik} = n'_{i.}\frac{n_{ik}}{n_{i.}} \quad (1)$$

$$\begin{cases} n'_{.1} = n'_{1.}\frac{n_{11}}{n_{1.}} + n'_{2.}\frac{n_{21}}{n_{2.}} + \dots + n'_{i.}\frac{n_{i1}}{n_{i.}} \\ n'_{.2} = n'_{1.}\frac{n_{12}}{n_{1.}} + n'_{2.}\frac{n_{22}}{n_{2.}} + \dots + n'_{i.}\frac{n_{i2}}{n_{i.}} \\ \dots = \dots + \dots + \dots + \dots \\ n'_{.i} = n'_{1.}\frac{n_{1i}}{n_{1.}} + n'_{2.}\frac{n_{2i}}{n_{2.}} + \dots + n'_{i.}\frac{n_{ii}}{n_{i.}} \end{cases} \quad (2)$$

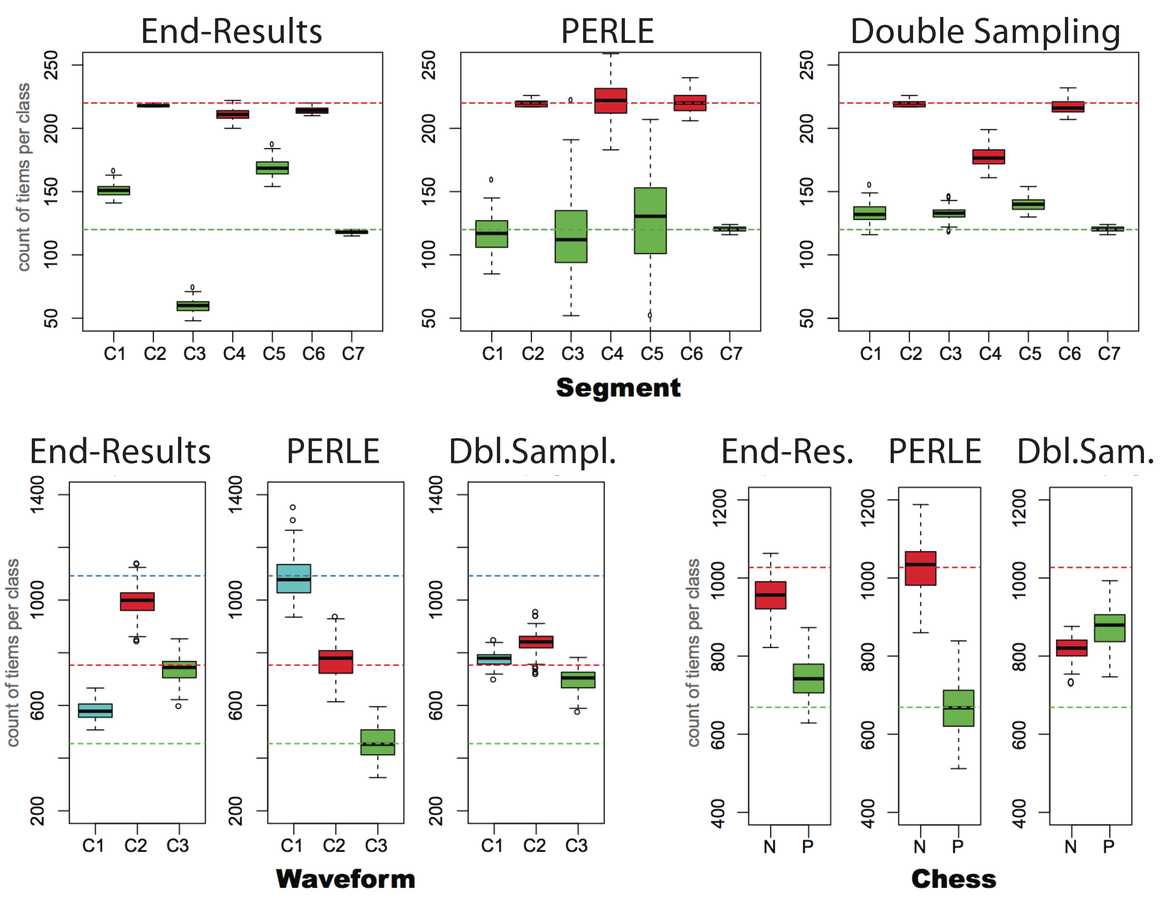
Corrected Counts / Counts in End-Results / Error Rates

$$\begin{pmatrix} n'_{1.} \\ n'_{2.} \\ \dots \\ n'_{i.} \end{pmatrix} = \begin{pmatrix} \frac{n_{11}}{n_{1.}} & \frac{n_{21}}{n_{2.}} & \dots & \frac{n_{i1}}{n_{i.}} \\ \frac{n_{12}}{n_{1.}} & \frac{n_{22}}{n_{2.}} & \dots & \frac{n_{i2}}{n_{i.}} \\ \dots & \dots & \dots & \dots \\ \frac{n_{1i}}{n_{1.}} & \frac{n_{2i}}{n_{2.}} & \dots & \frac{n_{ii}}{n_{i.}} \end{pmatrix}^{-1} \begin{pmatrix} n'_{.1} \\ n'_{.2} \\ \dots \\ n'_{.i} \end{pmatrix}$$

What If Class Proportions Vary?

If one class represents, e.g., 50% of all test set items but only 25% of the end-results dataset, then Double Sampling results are biased. PERLE results remain unbiased.

Evaluation



How Many Misclassifications in End-Results?

Data analysts are provided with imperfect classifiers, and to estimate classification uncertainty, with groundtruth evaluations. Classification errors are measured in test sets, but end-users are left with no estimate of the potential errors in end-results datasets. PERLE and Double Sampling address this problem. They estimate counts of items per class corrected for potential classification errors.

		True Class				Output Count
		c_1	c_2	...	c_i	
Output Class	c_1	n_{11}	n_{21}	...	n_{i1}	$n_{.1}$
	c_2	n_{12}	n_{22}	...	n_{i2}	$n_{.2}$
	...	...	...	...	...	...
	c_i	n_{1i}	n_{2i}	...	n_{ii}	$n_{.i}$
True Count		$n_{1.}$	$n_{2.}$	...	$n_{i.}$	$n_{..}$

Confusion matrix and notation (n for test set variables, n' for end-results)

Existing Method: Double Sampling

Developed in the 70s, it was not intended for machine learning classifiers. But it is applicable to our problem.

Corrected Count / Count in End-Results / Error Rate

$$n'_{i.} = \sum_k \frac{n_{ik}}{n_{.k}} n'_{.k}$$

$$TP' + FN' = \frac{TP}{TP+FP}(TP' + FP') + \frac{FN}{TN+FN}(TN' + FN')$$

$$= Precision \times (TP' + FP') + False\ \ Omission\ \ Rate \times (TN' + FN')$$

What If Error Rates Vary?

If different test sets were sampled, error rates may vary. End-results datasets may also involve varying error rates. We propose Sample-to-Sample estimates (3) to account for the variance over both test and end-results sets.

Sample-to-Sample estimate / Error Rate in Test Set / Variance in Test Sets / Variance in End-Results Sets

$$\hat{r}_{ik} \sim N\left(r_{ik}, \frac{r_{ik}(1-r_{ik})}{n_{i.}} + \frac{r_{ik}(1-r_{ik})}{n'_{i.}} \right) \quad (3)$$

We derived methods to estimate the variance of PERLE and Double Sampling results, for which Sample-to-Sample estimates were required. We found that error rate variance impacts PERLE results more than Double Sampling results.

Future Work

- Apply to real-world use cases
- Estimate optimal test set size
- Estimate which error correction method is preferable

Left Figure - Counts of items per class (boxplots) over random data subsets with fixed true counts (dashed lines)

[1] Tenenbein A.: A double sampling scheme for estimating misclassified multinomial data with application to samplign inspection (1972)
[2] Boom B.J. et al.: Uncertainty-aware estimation of population abundance using machine learning (2015)
[3] Beauxis-Aussalet E. et al.: Multifactorial uncertainty assessment for monitoring population dynamics (2015)