SAS® GLOBAL FORUM 2018

USERS PROGRAM

Minimum information for training a classifier

Catherine Halsey

April 8 – 11 | Denver, CO

#SASGF

MINIMUM INFORMATION FOR TRAINING A CLASSIFIER

Catherine Halsey, Frans Kanfer and Sollie Millard

University of Pretoria

ABSTRACT

- Classifier accuracy is extremely important and can be improved by increasing the size of the training data set. However, in experimental studies it might be very costly to survey cases; therefore, limiting sample size to a minimum is essential. Sometimes very large data sets might not contain enough information, and additional computer resources do not improve accuracy. Stopping at the optimal iteration saves computer time and sampling costs.

- For this reason, a sequential method of training classifiers is suggested. This method seeks to sample the minimum number of observations necessary to train a classifier to estimate the feasible minimum rate of misclassification, the Bayes error.

- Using SAS/IML® Studio, this method of classifier training proves ideal as it gives the researcher more control over the process by specifying when the sequential procedure should be stopped. It is not restricted to any single method of classification, and it never seeks to obtain an unfeasibly low misclassification rate.

RESEARCH PROBLEM

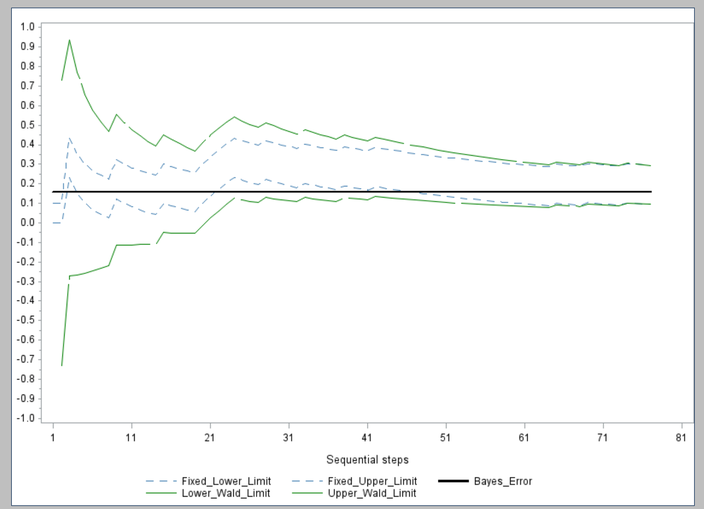
- Classifier accuracy can be improved by increasing training data sample size. This can prove difficult as factors such as time, affordability, information availability and level of computer intensity come into play. Therefore it is ideal to train a classifier using the least amount of observations.

- A sequential training procedure is proposed in which an algorithm, based on a derived stopping criteria, is used to update the classification rule following each sequential step until it can be guaranteed with a prescribed level of confidence that the probability of an incorrect classification is within a certain pre-specified level of the minimum feasible error [3].

- The process repeatedly samples, trains and classifies observations until the misclassification rate that lies within *h* of the Bayes error with a certain probability, ensuring that the classifier is always at optimum performance.

SUGGESTED TRAINING ALGORITHM

1. Obtain an initial sample of size S_0 and select a desired half-width $h \in \{0.1, 0.05, 0.01\}$ and the associated values of a and α.
2. At the i^{th} iteration, train the classifier using all observations obtained thus far.
3. Sample an additional random observation and classify it using the classifier trained in step 2.
4. Test whether the new observation has been correctly classified. If a correct Classification is made set $Q_i = 0$. If the observation is misclassified set $Q_i = 1$.
5. Calculate $\hat{p}$, the proportion of observations misclassified thus far, and as discussed in Frey [1] evaluate the stopping rule: $\frac{\tilde{p}_a(1-\tilde{p}_a)}{n} \leq \left(\frac{h}{\frac{Z_\alpha}{2}}\right)^2$.
6. If the stopping rule is met, stop the sequential training procedure, if not repeat steps 2 to 5.



Graphical representation of the convergence of the adapted Wald confidence interval towards the fixed-width interval for one iteration of a simulation study conducted using SAS/IML® studio

Minimum information for training a classifier

Catherine Halsey, Frans Kanfer and Sollie Millard

University of Pretoria

APPLICATION

- As an example the procedure is applied to a set of handwritten digits from the ZIP codes on envelopes from U.S postal mail [2].
- Based on the pixel intensity of 16x16 eight bit greyscale images of single digits the goal is to train a classifier using the least amount of observations to classify an image of a handwritten digit into one of the groups 0, 1, 2, 3, 4, 5, 6, 7, 8 or 9 as accurately as possible.
- Two data sets were used for this study, a training data set consisting of a spread of 2 000 handwritten digits and a testing data set consisting of 5 000.
- Initially, using the full training data set and *PROC DISCRIM* a LDA classifier is trained and the resulting classifier used to classify the observations from the testing data set.
- Thereafter, using the suggested algorithm in SAS/IML® Studio the procedure was applied to the training data set using LDA classification and the resulting classifier used to classify the observations from the testing data set.
- Results showed that with only a maximum of 301 training observations when h=0.1 or 519 observations when h=0.05, it is possible to obtain an observed misclassification rate that is within h of that obtained when the full 2 000 training observations were used.

CONCLUSIONS

- In experimental cases it is often not possible to obtain large amounts of data with which to accurately train a classifier. A sequential method, as proposed in Potgieter [3], of training a classifier to estimate the Bayes error is able to ensure with a certain level of confidence that the probability of the classifier making an error is within a pre-specified level of this classifier making an error is within a pre-specified level of this whilst only requiring the smallest possible number of observations.
- The method is ideal as it gives the researcher more control over the process by specifying when the sequential procedure should be stopped, it is also not restricted to any single method of classification due to the constant updating of classification rules at each step. This classifier training method can prove useful in many real-world situations, saving on required observations and computational time.

REFERENCES

[1] J. Frey. Fixed-width sequential confidence intervals for a proportion. *The American Statistician*, 64(3):242-249, 08 2010.

[2] T Hastie, R Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, 2009.

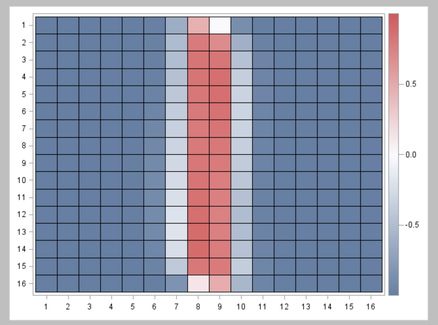
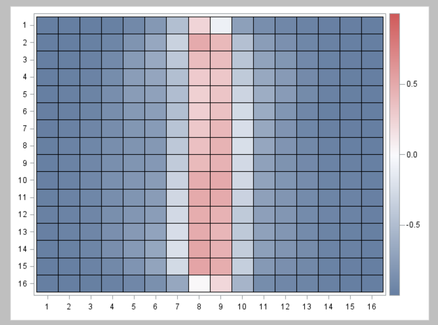
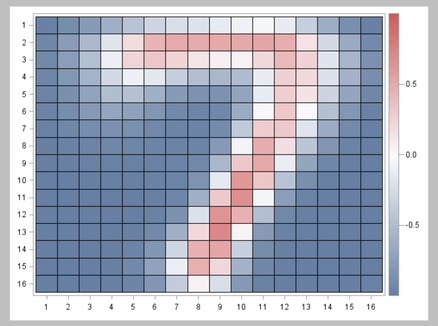
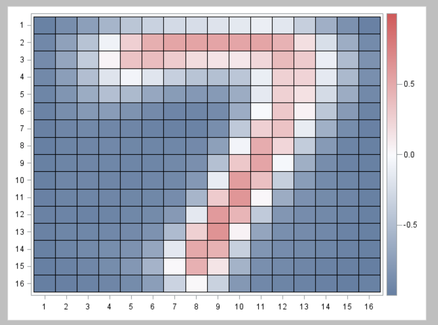
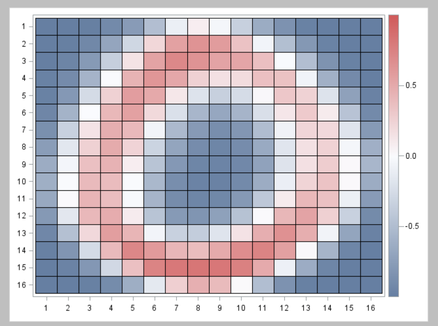
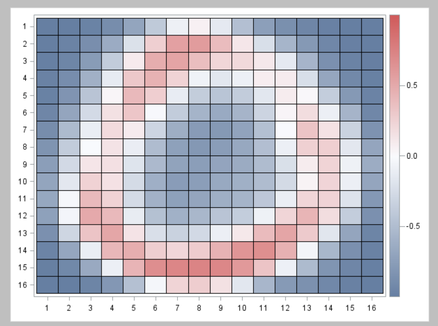
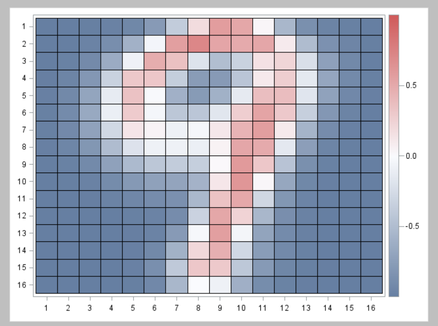
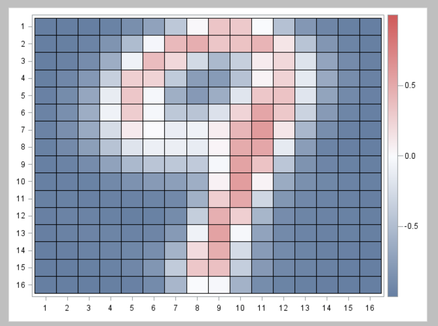
[3] R. Potgieter. Minimum sample size for estimating the bayes error at a predetermined level. Master's thesis, University of Pretoria, 2013.

Minimum information for training a classifier

Catherine Halsey, Frans Kanfer and Sollie Millard

University of Pretoria

Comparison of heat maps generated using SAS/IML® of original observed digits from testing data set (left) and the digits predicted by the sequentially trained classifier (right) for one simulation of the procedure for *h*=0.1

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF