Author's response to reviews

Title: Predicting Sample Size Required for Classification Performance

Authors:

Rosa L Figueroa (rosfigue@udec.cl)
Qing Zeng-Treitler (q.t.zeng@utah.edu)
Sasikiran Kandula (sasi.kandula@utah.edu)
Long H Ngo (lngo@bidmc.harvard.edu)

Version: 2 Date: 9 November 2011

Author's response to reviews: see over
Dear Natalie Pafitis,

On behalf of my co-authors, I’m submitting a revised version of the manuscript titled: “Predicting Future Classification Performance based on Sample Size”. We have thoroughly revised the manuscript to incorporate the reviewers’ comments within the manuscript. Additionally, we include in this cover letter a detailed summary of the changes made in response to the reviewers’ comments.

Thank you for your consideration. I look forward to hearing from you.

Qing Zeng-Treitler, PhD
Department of Biomedical Informatics,
26 S 2000 E, HSEB 5700,
University of Utah,
Salt Lake City, UT 84112-5750.
Email: q.t.zeng@utah.edu

Enclosures:

Response to reviewers' comments
Reviewer 1: Chuan LU

Major Compulsory Revisions

Comment 1: It is unclear what classifiers have been used for these examples. Maybe SVM were used in this work as hinted by the active sampling methods. However, it should be explained explicitly and more details should be given on e.g. parameter settings for building of the classifiers and the active sampling methods.

Our response: The reviewer was correct. The learning curves obtained from the datasets presented on the paper were obtained using active learning as an iterative sampling method. A SVM implementation from Weka was used. More details about the classifier and the active learning methods have been included within the manuscript (page 11).

Comment 2: For the three active sampling methods mentioned in the Method- Evaluation section, although references were given, it would still be better if more details can be provided, as some references are not accessible to readers (even after google search).

Our response: More details about the methods have been provided within the text. We apologize for the inconvenience with the references. The reference “Exploring Active Learning in Medical Text Classification” by Figueroa and Zeng-Treitler used to be available from the AMIA website but it appears they have introduced a new way to explore and navigate within the AMIA proceedings and somehow the abstract is not there anymore. The references will be made available online on our website. (page 11)

Comment 3: It would be useful to report the prevalence of each class for all three datasets. It would be also useful to compare the confidence intervals estimated from the Hessian, to the empirical confidence intervals estimated from 100 runs of random sampling and testing.

Our response: The class prevalence for each dataset has been added to the manuscript (beginning of page 11 together with data description). We also added the confidence intervals from the observed data (estimated from 100 runs of random sampling and testing) to the figure 2 and created a new figure 5. It can be observed that the width of the observed confidence intervals changes only slightly along the learning curve, showing that performance variance among experiments are not strongly impacted by the sample size. The predicated confidence interval on the other hand narrows dramatically as more samples are used and the predication becomes more accurate (brief description of comparison added to the last paragraph of page 15).

Reviewer 2: Stephen Luther

Major Compulsory Revisions
None

Minor Essential Revisions

Comment 1: They described the data sets pretty well, although I wondered why the waveform-5000 data set was split evenly between training and evaluation.
Our response: The dataset waveform5000 used in this study came from a prior computer science publication. We used the waveform5000 datasets to reproduce the same learning curve from that study in which the dataset was split 100 times into a training set and a test set of equal size. Splitting the data differently may result in different curves. The same curve-fitting algorithm would still apply.

Comment 2: The model fitting section was complex but well written. One minor note, I believe there is a word missing at the end of the first sentence “after equation (1).”

Our response: The sentences has been reviewed and changed (page 9, Model fitting and parameter identification subsection).

Comment 3: One thing that i did not get from the methods section was an explicit description of the method used to classify each sentence/case. They describe how the information, from each iteration is used to estimate the power curve but what classifier did they use to get the data. Mukherejee used SVMs. Is that what is used here? This should be clarified in the paper.

Our response: As Mukherejee’s study, we also used SVM as our base classifier. Details about the classifier were included in the manuscript (page 11 second paragraph).

Comment 4: The data are sound and represent an expansion from previous work. The authors correctly state that including three different data sets is a plus for the study. However, the two smoking datasets would appear to be very similar and have a large number of features available to classify a relatively simple attribute in the data. Classification is often impacted by the rate of occurrence of positive sentences/cases in documents. It would be useful to know how many positive sentences/cases were in each set.

Our response: The class prevalence was detailed in the manuscript together with the data description (beginning of page 11 on the Datasets description subsection).

Comment 5: The authors acknowledge that the confidence provided through their method may be optimistic given that they are calculated using the variance-covariance matrix. They mention that the learning curves reported in this paper are based on 10 fold cross validations repeated 100 times and that less iterations would likely provide less smooth curves. It would be informative to the reader to describe the time needed to perform the reported analysis.

Our response: The following sentences have been added to the discussion section(page 17 second paragraph): “Although the curve fitting can be done in real time, the time to create the learning curve depends on the classification task, batch size, feature number, processing time of the machine among others. The longest experiment we performed to create a learning curve run on a single core laptop for several days, though most experiments need only a few hours.”

Comment 6: I do not program in R myself but wondered if the algorithm provided included the cross validation and iterative options. If not, this would be a nice addition to the algorithm.

Our response: The method presented in this paper was designed to predict performance based in sample size, in order to help researchers to make informed decisions about data
labeling. Cross-validation and interactive sampling are part of the active learning processing methods that are used to generate the curves in the paper.

Descretionary Revisions
In truth the title is not particularly informative to me. It does not reflect the aim of the study which is to improve sample size estimation.

Our response: The method we present here can be used to predict sample size based on target performance or predict performance for a sample size. We propose a new title “Predicting Sample Size Required for Classification Performance” to address the reviewer’s concern.