Predicting Outcome Measures in Active Learning

Sasikiran Kandula[a], Rosa Figueroa[b], Qing Zeng-Treitler[a]

[a] *Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah.*
[b] *Dep. Ing. Eléctrica, Fac. de Ingeniería, Universidad de Concepción, Concepción, Chile.*

Abstract and Objective

Text classification requires labeled data which is generally scarce and creating additional labeled data requires manual annotation which is an expensive and time consuming task. In our recent work with medical text classification we adopted an iterative labeling process where physicians label instances in batches. Each batch is incrementally added to the training set and changes in outcome measures, such as area under receiver operating characteristics curve (AUC) and accuracy, are calculated. We realized that being able to predict the outcome at a given sample size would be useful and can inform our decisions to terminate labeling and/or learning. In this paper, we describe one such method based on non-linear curve fitting.

Keywords:

Active learning, Non-linear regression, Curve fitting

Methods

We use a dataset of 11000 sentences each of which was manually reviewed and labeled with its' smoking status: non-smoker or smoker. The dataset was used to test a few active learning selection algorithms with fixed batch size of 16 i.e. the size of the training set increases in increments of 16. At each sample set size, AUC and accuracy are calculated and noted. This process yields n data points $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$, where x_i is the sample size and y_i is an outcome measure observed at x_i. Using these data points, we attempt to predict y_k at any given x_k ($x_k > x_n$).

We experimented with different functions that relate observed outcome measures with sample size and found the following hyperbolic function to be most suitable:

$$y_i = \frac{a * x_i}{b + x_i} + c * x_i \qquad (1)$$

In the above equation, a, b and c are parameters that define the fitted curve and are estimated using non-linear regression analysis.

To test the curve fitting process, we use an iterative process and partition the available n data points into two sets. In the i^{th} iteration ($n/3 \leq i \leq n$), points (x_1, y_1) to (x_i, y_i) are used to fit the curve and data points (x_{i+1}, y_{i+1}) to (x_n, y_n) are used to test the curve. After each iteration the goodness of fit is calculated using mean absolute error:

$$\sum_{j=i+1}^{n} \left| y_j^{predicted} - y_j^{actual} \right| \qquad (2)$$

It is generally recognized that outcome measures observed at larger sample sizes are more accurate. Hence a better fit can be obtained by assigning higher weights to the data points with larger x values. We use a normalized linear weighting scheme wherein for each (x_i, y_i) used to fit the curve, we assign weight i/k where k is the number of data points used to fit the curve.

Results

For the dataset considered here using a distance-based active learning algorithm [1], the mean absolute error observed for AUC at $i = n/3$ is 0.06. The error decreases with increasing i and at $i = n-1$ the error reduces to 0.001(*Distance* in Figure 1)

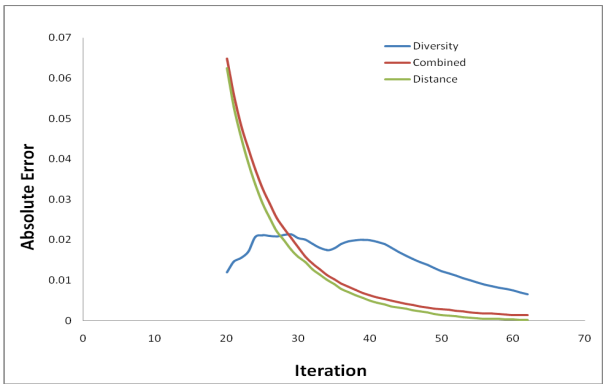

Figure 1-Rate of change of error observed in three active learning algorithms

Small differences in rate of change of error were observed for a diversity-based algorithm (*Diversity*) [2] and a modified algorithm that combines both distance and diversity.

Conclusion

We have discussed a simple method that can be used to predict the improvement in outcome measures with increasing sample size. Our preliminary results show that the hyperbolic function we used is representative of the response of outcome measures to training sample's size.

References

[1]. Tong S, Koller D. Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research. 2001;2:45-66.

[2]. Brinker K. Incorporating Diversity in Active Learning with Support Vector Machines. In Proceedings of the Twentieth International Conference on Machine Learning (ICML); 2003; p. 59-66.