Optimizing Performance of an *Ngram* Method for Classifying Emergency Department Visits into the Respiratory Syndrome

Philip Brown[1], Sylvia Halasz[1] PhD, Dennis G. Cochrane[2] MD,
John R. Allegra[2] MD, PhD, Colin Goodall[1] PhD, Simon Tse[1] PhD
[1]ATT Research Labs; [2]Emergency Medical Associates of NJ Research Foundation

Introduction

A number of different methods are currently used to classify patients into syndromic groups based on the patient's chief complaint (CC). We previously reported results using an "*Ngram*" text processing program for building classifiers (adapted from business research technology at AT&T Labs). The method applies the ICD9 classifier to a training set of ED visits for which both the CC and ICD9 code are known. A computerized method is used to automatically generate a collection of CC substrings (or *Ngrams*), with associated probabilities, from the training data. We then generate a CC classifier from the collection of *Ngrams* and use it to find a classification probability for each patient. Previously, we presented data showing good correlation between daily volumes as measured by the *Ngram* and ICD9 classifiers.

Objectives

Our objective was to determine the optimized values for the sensitivity and specificity of the *Ngram* CC classifier for individual visits using a ROC curve analysis. Points on the ROC curve correspond to different classification probability cutoffs.

Methods

We used a computerized database of consecutive visits seen by ED physicians from 1-1-2004 to 3-31-2005 (800,993 visits). We used as our ICD9 classifier an existing ESSENCE filter for the respiratory syndrome, RESP. The ICD9 classifier was applied to a training set of all visits for the year 2004 to create the *Ngram* based CC classifier (a collection of 95 *Ngrams* with a preferred size of 4 characters).

We then used both the *Ngram* CC and ICD9 classifiers to categorize the test set of visits from 1-1-2005 to 3-31-2005 (166,490). To generate the ROC curve, we chose a set of cutoffs for the classification probability that matched the probabilities for the classifier *Ngrams*. The ROC curve shows the sensitivity and specificity for each of these cutoffs. For the optimum cutoff, we chose the point on the ROC curve closest in least squares distance to sensitivity=specificity=1.

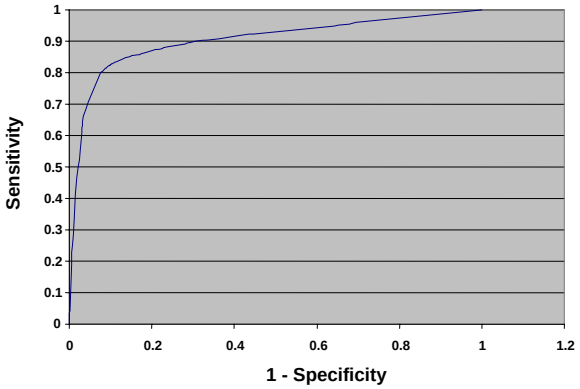
Results

The ROC analysis revealed for the RESP syndrome an area under the curve [AUC = 0.903 +/- 0.001] with an optimal sensitivity = 81% and specificity = 92%, for which the classification probability is 0.31 (comprising 45 *Ngrams*). Among the most predictive *Ngrams* were words starting with "ASTH" (.977) and words containing "HEEZ" (.963). The largest spikes in sensitivity were seen for words starting with "COUG" (.932), words containing "HEST" (.840) and words starting with "FEVE" (.467).

Conclusion

The sensitivity and specificity of an *Ngram* CC classifier for the RESP syndrome compares favorably to manually created CC classifiers. This approach has promise in that it may offer a complementary method to using manual and natural-language processing techniques to create CC classifiers. The approach has the advantages that it allows the rapid automated creation and updating of CC classifiers based on ICD9 groupings and may be independent of the spoken language or dialect.

ROC Curve



Sensitivity & Specificity versus Ngram Positive Predictive Value