SyCo: A Probabilistic Machine Learning Method for Classifying Chief Complaints into Symptom and Syndrome Categories

Jeremy U. Espino[1], John Dowling[2], John Levander[2], Peter Sutovsky[2], Michael M. Wagner[1,2], Gregory F. Cooper[2]

General Biodefense, Pittsburgh, PA[1]

RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh, PA[2]

OBJECTIVE

Design, build and evaluate a symptom-based probabilistic chief complaint classifier for the Real-time Outbreak and Disease Surveillance System (RODS).

BACKGROUND

Scientists have utilized many chief complaint (CC) classification techniques in biosurveillance including keyword search,[1,2] weighted keyword search,[3] and naïve Bayes.[4] These techniques may utilize CC-to-syndrome or CC-to-symptom-to-syndrome classification approaches. In the former approach, we classify a CC directly into syndrome categories. In the latter approach, we first classify a CC into symptom categories. Then, we use a syndrome definition, a combination of one or more symptoms, to determine whether or not a chief complaint belongs in a particular syndrome category. One approach to CC-to-symptom-to-syndrome classification uses manually weighted keyword search and Boolean operations to build syndrome classifiers.[3] A limitation to this approach is that it does not address uncertainty in the data and the system is manually parameterized. A CC-to-symptom-to-syndrome approach that is both probabilistic and utilizes machine learning addresses these limitations.

METHODS

We constructed SyCo — a CC-to-symptom-to-syndrome probabilistic chief complaint classifier. SyCo learns a Naïve Bayes model of the relationship between words and symptoms given a training set of labeled chief complaints.[5]

To perform a classification, SyCo first computes the posterior probability of each symptom using the odds formulation of Bayes rule. SyCo can compute the posterior probability traditionally or in single word mode. When single word mode is enabled SyCo will only use the likelihood ratio of the word (given a symptom) that maximizes the posterior probability.

Finally, SyCo uses the posterior probabilities from the first step to compute the posterior probability of a syndrome given a chief complaint. A syndrome is defined as any combination of symptoms and Boolean operations. SyCo supports the operations AND, OR, and NOT by using the rules of conjunction, disjunction and negation of independent events. For example, P(A) AND P(B) = P(A) x P(B).

A board certified infectious disease physician [JD] read 16718 chief complaints and indicated the presence or absence of seventeen symptoms for each chief complaint.

We measured the performance of SyCo when classifying seventeen individual symptoms and three syndromes with and without the single word mode using leave-one-out cross validation. We measured the area under the curve (AUC) of the resultant receiver operator characteristic (ROC) curves and established 90% confidence intervals using 100 iterations of non-parametric bootstrapping.

RESULTS

The area under the curve without and with the single word assumption ranged from 0.785 to 0.9918 and 0.7442 to 0.9916, respectively. The single word mode improved performance significantly in 7 out of 20 cases and degraded performance in 2 out of the 20 cases.

CONCLUSION

SyCo is a symptom-based probabilistic chief complaint classifier that has excellent discriminatory ability for classifying chief complaints into symptom categories and syndromes. We have made SyCo available in RODS Version 4.2.

ACKNOWLEDGEMENTS

This research was supported in part by DARPA/Mellon/Pitt grant F30602-01-2-0550, NSF grant IIS-0325581 and PA Department of Health grant ME-01-737

Table 1. Area under the curves (AUC) of the receiver operator characteristic curves of SyCo when classifying a chief complaint. Superscripts define syndromes which are Boolean disjunctions of symptoms. Shaded cells indicate significant difference between AUC values without and with the single word assumption.

Symptom	AUC [90% CI]	AUC w/single word assumption [90%CI]
Fever or Chills[2]	0.9918 [0.9887-0.9947]	0.9906 [0.9862-0.9956]
Sweats[2]	0.785 [0.6887-0.8768]	0.8943 [0.7945-0.9676]
Fatigue or Malaise[2]	0.9517 [0.9411-0.9629]	0.962 [0.9453-0.9793]
Cough[1]	0.9849 [0.98-0.9888]	0.9809 [0.9712-0.992]
Nausea or Vomiting[2]	0.9914 [0.9887-0.9943]	0.9895 [0.984-0.995]
Respiratory Distress[1]	0.99 [0.9876-0.9926]	0.9821 [0.9753-0.9892]
Chest Discomfort or Pleuritic Pain[1]	0.8167 [0.7792-0.8518]	0.8922 [0.8605-0.9271]
Myalgia[3]	0.9229 [0.8996-0.9433]	0.931 [0.901-0.9668]
Headache	0.9634 [0.9576-0.9692]	0.9791 [0.9712-0.9863]
Meningitis Symptoms	0.9377 [0.9155-0.9526]	0.9515 [0.9287-0.9738]
Abdominal Pain[2]	0.9806 [0.9759-0.984]	0.9866 [0.9824-0.9908]
Sore Throat[1]	0.9671 [0.9564-0.9755]	0.9916 [0.9846-0.9976]
Upper Respiratory Infection Symptoms[1]	0.9528 [0.9426-0.9636]	0.9836 [0.9724-0.9938]
Hemoptysis	0.9059 [0.8473-0.9481]	0.8858 [0.7955-0.97]
Infectious Symptoms	0.8997 [0.8776-0.9211]	0.7442 [0.6815-0.8218]
Sepsis or Shock	0.8617 [0.8241-0.899]	0.7602 [0.6989-0.8225]
Tachycardia	0.9409 [0.9121-0.967]	0.9289 [0.8689-0.9693]
Syndrome 1 - Respiratory	0.9606 [0.9551-0.9649]	0.9815 [0.978-0.9856]
Syndrome 2 - Gastrointestinal	0.9823 [0.9797-0.9851]	0.9885 [0.9858-0.9919]
Syndrome 3 – Constitutional	0.9293 [0.9207-0.9383]	0.9555 [0.9493-0.9619]

REFERENCES

1. Hutwagner L, Thompson W, Seeman GM, Treadwell T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). J Urban Health 2003;80(2 Suppl 1):i89-96.
2. Das D, Weiss D, Mostashari F, Treadwell T, McQuiston J, Hutwagner L, et al. Enhanced drop-in syndromic surveillance in New York City following September 11, 2001. J Urban Health 2003;80(2 Suppl 1):i76-88.
3. Sniegoski C. Automated syndromic classification of chief complaint records. Johns Hopkins APL Technical Digest 2004;25(1):68-75.
4. Olszewski R. Bayesian classification of triage diagnoses for the early detection of epidemics. In: 16th International FLAIRS Conference; 2003; Florida; 2003. p. 412-416.
5. Manning CD, Schèutze H. Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press; 1999.