Identifying ILI Cases from Chief Complaints: Comparing Keyword and Support Vector Machine Methods[1]

Darrell Ferguson[2], MCS; Norman G. Vinson[2], PhD; Jason Morin[2], MSc; Joel Martin[2] PhD; Susan McClinton, BScN[3]; Richard Davies[3], MD PhD

[2]*National Research Council of Canada, Institute for Information Technology*

[3]*University of Ottawa Heart Institute (UOHI)*

Ottawa, Canada

OBJECTIVE

We compared the accuracy of two methods of identifying ILI cases from chief complaints: a method based on keywords and one based on support vector machine learning.

BACKGROUND

The rapid spread of the novel H1N1 virus prompted Ottawa Public Health (OPH) to monitor Emergency Department Chief Complaints (EDCC) specifically for influenza-like illness (ILI). Note that data from ED visits is the most common data source for syndromic surveillance systems in the US [1].

METHODS

Our data set was formed of 149910 case records composed of free text EDCC and accompanying patient age. Each EDCC was typed into the system by the triage nurse at the time of the visit. The data set covered ED visits from May 2008 to June 2009 (approx.), which includes about 3 months of the H1N1 outbreak.

Our keyword method was based on human expert identification of ILI case records. Because the EDCC were free text, they contained misspellings, synonyms, and truncations. To compensate, we incorporated the EDCC variations into our keyword list [2].

The support vector machine (SVM) method uses a training set to learn to classify input items according to their features [3]. In this sense, it is similar to Naïve Bayes, which has been used previously to classify EDCC [2]. Unlike the keyword method, the SVM method does not require any compensation for misspellings, synonyms, or truncations.

We developed our training set by having human experts identify all the ILI records in our data set. We then used 10-fold cross validation to estimate our SVM model's performance. Specifically, we broke the dataset into 10 subsets of equal size, trained SVM on 9 of the subsets, and tested SVM on the remaining subset. This process was repeated 10 times, each with a different test set, providing a mean accuracy.

RESULTS

ILI Identification Accuracy for Keywords and SVM.

Method	Precision	Recall/Sensitivity	Specificity
Keywords	96.8%	97.0%	99.6%
SVM	97.4%	97.5%	99.7%

CONCLUSIONS

SVM proved slightly superior to the keyword method in identifying ILI cases in ED case records each composed of an EDCC and accompanying age. Moreover, SVM took misspellings, synonyms, and truncations into account automatically, while the keyword method required a human analysis of the EDCC to uncover the variation and create compensating keywords.

Because SVM is *based* on human expert classification, it cannot perform better than human experts. Unfortunately, there is evidence that human experts are quite poor at identifying febrile respiratory (similar to ILI) cases in EDCC data [4]. [4]'s expert only flagged 0.32% of the EDCC as febrile respiratory cases, while the estimated true incidence was 12.33%. In contrast, our human experts flagged 10.01% of our case records as ILI. This suggests that ILI is more detectible in our EDCC than febrile respiratory was in [4]'s. Consequently, our SVM model may not have been hampered by the limitations reported in [4].

REFERENCES

[1] Buehler, J.W.; Sonricker, A.; Paladini, M.; Soper, M. & Mostashari, F. Syndromic Surveillance Practice in the United States: Findings from a Survey of State, Territorial, and Selected Local Health Departments. Adv. in Disease Surveillance, 2008; 6(3)

[2] Dara, J.; Dowling, J.N.; Travers, D.; Cooper, G.F.; Chapman, W.W. Chief Complaint Preprocessing Evaluated on Statistical and Non-Statistical Classifiers. Adv. in Disease Surveillance 2007; 2:4.

[3] de Bruijn, B.; Cranney, A.; O'Donnell, S.; Martin, J.D.; Forster, A.J. Identifying Wrist Fracture Patients with High Accuracy by Automatic Categorization of X-ray Reports, Journal of the American Medical Informatics Association, 2006, 13(6), 696-698.

[4] Chapman, W.W. & Dowling, J.N. Can Chief Complaints Identify Patients with Febrile Syndromes? Adv. in Disease Surveillance, 2007, 3(6)

[1] This work was supported by CRTI (Chemical, Biological, Radiological-Nuclear, and Explosives (CBRNE) Research and Technology Initiative) grant #06-0234TA