Bayesian Classification of Triage Diagnoses for the Early Detection of Epidemics

Robert T. Olszewski

Center for Biomedical Informatics
University of Pittsburgh
Pittsburgh, Pennsylvania 15213
bobski@cbmi.upmc.edu

Abstract

The distribution of illnesses reported by emergency departments from hospitals in a region under surveillance is particularly informative for the early detection of epidemics. The most direct source of data for construction of such a distribution is the final diagnoses of patients being seen in the emergency departments, but the delay in their availability impinges on the requirement that detection be timely. Free-text descriptions of patients' symptoms, called triage diagnoses, and ICD-9 values that encode the symptoms are entered when patients are admitted and, consequently, are timelier sources of data. An experiment to evaluate the accuracy of Bayesian classification of triage diagnoses into syndromes (i.e., illness categories) was performed, resulting in areas under the ROC curve (AUC) between .80 and .97 for the various syndromes. The classification accuracies using triage diagnoses surpass the classification accuracies using ICD-9 codes reported by previous studies. Triage diagnoses, therefore, are a more accurate source of data than ICD-9 codes for the early detection of epidemics.

Introduction

The early detection of epidemics (either naturally-occurring or as the result of bioterrorism) is of critical concern for public health surveillance systems (Wagner et al. 2001). Such a system relies upon the real-time collection and analysis of various types of data indicative of the state of health of the region under surveillance. One type of data that is particularly informative for this task is the distribution of illnesses reported by emergency departments from region hospitals: unexplained deviations from the expected distribution may be evidence of an outbreak. Using such data, it has been shown that it is possible to detect an influenza outbreak earlier than with more traditional data (Tsui et al. 2001).

A distribution of illnesses could be most directly constructed from the final diagnoses of patients admitted to emergency departments. However, the delay between a patient's admission and the availability of the patient's final diagnosis has been shown by one study to have a mean of 7.5 hours and a maximum of 80.6 hours (Espino and Wagner 2001), and by another study to have a mean of 6.3 hours and a maximum of 14.5 hours (Ivanov et al. 2002). Information entered upon patient admission, therefore, would be a timelier source of data. Such information includes a free-text description of the patient's symptoms, called a triage diagnosis, which combines the patient's description of symptoms with medical terminology added by the admitting nurse, and an ICD-9 (International Classification of Diseases, ninth revision) value that encodes the contents of the triage diagnosis. Classification of patients into categories of illnesses, called syndromes, using ICD-9 codes has been reported by previous studies to suffer from poor accuracy (Espino and Wagner 2001; Ivanov et al. 2002). Using triage diagnoses for classification may result in an improvement in accuracy.

To test the hypothesis that classification using triage diagnoses is more accurate than with ICD-9 codes, an experiment was performed using statistical natural language processing techniques: a Bayesian classifier was employed to classify triage diagnoses into syndromes, and the results were compared with those previously reported for ICD-9 codes.

Methods

A designed experiment was used to evaluate the accuracy of Bayesian classification of triage diagnoses. A set of triage diagnoses was collected and classified by hand for the experiment. Bayesian classification of the triage diagnoses was performed using three different language models, and the classification accuracies were compared among the models.

Data

A set of 28,990 triage diagnoses was collected from an emergency department in Utah covering patient visits over the course of one calendar month. The triage diagnoses contained a large number of misspelled and abbreviated words. Commonly-used phrases were also regularly abbreviated. The triage diagnoses as entered by the admitting nurses were preprocessed to transform all letters into lowercase and to replace all punctuation with spaces, resulting in a collection of triage diagnoses ranging in length from one word to ten words.

A physician read and classified each individual triage diagnosis with one or more of eight syndromes used for this experiment. The syndromes included Gastrointestinal, Constitutional, Respiratory, Rash, Hemorrhagic, Botulinic, Neurological, and Other. The decision to classify a triage diagnosis with a particular syndrome was based on whether the words of the triage diagnosis suggested that the patient was suffering from symptoms indicative of that general category of illness: symptoms such as nausea, vomiting, and diarrhea were classified as Gastrointestinal; non-localized systemic problems like fever, chills, or influenza were classified as Constitutional; problems with the nose, throat, or lungs were classified as Respiratory; any description of a rash was classified as Rash; bleeding from any site was classified as Hemorrhagic; ocular abnormalities, and difficulty speaking or swallowing were classified as Botulinic; non-psychiatric complaints related to brain function were classified as Neurological; and any pain or process in a system excluded from surveillance (e.g., trauma, psychological evaluations) was classified as Other. Table 1 shows the distribution of the triage diagnoses across the syndromes.

Syndrome	Number
Gastrointestinal	4,082
Constitutional	1,848
Respiratory	3,438
Rash	317
Hemorrhagic	799
Botulinic	85
Neurological	2,521
Other	17,166

Table 1. The distribution of triage diagnoses.

Classification

A Bayesian classifier was used to assign one or more syndromes to a triage diagnosis. A training set of triage diagnoses was used to estimate the prior probability and the probabilities of unique words for each syndrome. Given a triage diagnosis, these probabilities were used to compute the posterior probability for each syndrome; the set of syndromes with posterior probabilities above a specified threshold was used to classify the triage diagnosis.

Given a triage diagnosis G consisting of a sequence of words $w_1\ w_2\ \ldots\ w_n$, the posterior probability of syndrome R, $P(R|G)$, can be expressed using Bayes' rule and the expansion of G into words as

$$P(R \mid G) = \frac{P(R)P(w_1 \mid R)P(w_2 \mid w_1 R)\cdots P(w_n \mid w_1 \ldots w_{n-1} R)}{\sum_R P(R)P(w_1 \mid R)P(w_2 \mid w_1 R)\cdots P(w_n \mid w_1 \ldots w_{n-1} R)}$$

An approximation of $P(R|G)$ was computed by employing language models that made assumptions about the conditional independence of the words in a triage diagnosis (Manning and Schütze 1999). For this experiment, three models were investigated: unigram (i.e., each individual word was assumed to be conditionally independent), bigram (i.e., each word pair was assumed to be conditionally independent), and mixture (i.e., a weighted combination of the unigram and bigram models). The mixture model applied a weight of .05 to the unigram model and a weight of .95 to the bigram model. While there are an infinite number of ratios of the unigram and bigram models that could have been used, this one was selected because it represents the basic bigram model with a small correction introduced by the unigram model to compensate for word pairs absent from the training set.

To apply the Bayesian classifier with a particular model, the triage diagnoses were separated into a training set and a test set, the prior and word probabilities were estimated from the training set, and each triage diagnosis in the test set was classified with the subset of syndromes having a posterior probability above a specified threshold. The sensitivity and specificity of the model were computed for each syndrome by comparing the true classifications of the triage diagnoses in the test set with those assigned by the classifier. Classification with the same training and test sets over a range of thresholds resulted in a collection of sensitivity and specificity pairs that described the performance of the classifier under different conditions.

Experiment Design

A ten-fold cross-validation experiment was performed to estimate the accuracy of the Bayesian classifier across different training sets. First, the set of triage diagnoses was divided into ten randomly-selected, disjoint subsets. Then the following steps were performed for each iteration: (1) one subset was selected as the test set and the remaining nine subsets were concatenated to form a training set; (2) the Bayesian classifier was trained with one of the three models using the training set; (3) the trained classifier was used to assign syndrome classifications to each triage diagnosis in the test set over a range of thresholds; and (4) the sensitivity and specificity at each threshold for each syndrome were computed. Ten iterations were performed, each with a different subset used as the test set. The entire experiment was repeated with each of the models using the same randomly-selected subsets of triage diagnoses.

Analysis

For each combination of model, syndrome, and iteration, the end result of the experiment was a set of sensitivity and specificity pairs which determined an ROC curve that described the performance of the classifier under different conditions (Zweig and Campbell 1993). To transform each ROC curve into a single-number representation of the classification accuracy, the area under the ROC curve (AUC) was computed (Bradley 1997; Hanley and McNeil 1982). Then, to obtain an estimate of the overall classification accuracy for each combination of model and syndrome, the mean and standard deviation of the AUC across the ten iterations was computed and used as the

basis for comparison among the models for each syndrome. Since this methodology did not result in an average ROC curve for each combination of model and syndrome, the pooled sensitivity and specificity values (i.e., the average sensitivity and specificity values at each threshold across the ten iterations) were used to display the ROC curves (Bradley 1997).

Results

Table 2 shows the mean (and standard deviation) of the AUC for each combination of syndrome and model from the ten-fold cross-validation experiment. A pairwise comparison of the mean AUC for each syndrome between models shows that the unigram model consistently resulted in a larger mean AUC than did the bigram and mixture models. One-tailed paired t-tests confirm that the pairwise differences in mean AUC between the unigram and bigram models are statistically significant at the $p < .0001$ level (except for the Rash syndrome which is significant at the $p < .001$ level), and that the pairwise differences in mean AUC between the unigram and mixture models are significant for only the Constitutional, Respiratory, and Neurological syndromes (at the $p < .0001$ level).

Syndrome	Unigram	Mixture	Bigram
Gastrointestinal	.945 (.012)	.941 (.013)	.891 (.013)
Constitutional	.931 (.012)	.916 (.015)	.853 (.017)
Respiratory	.957 (.009)	.949 (.009)	.888 (.013)
Rash	.910 (.030)	.904 (.030)	.837 (.055)
Hemorrhagic	.926 (.012)	.919 (.012)	.866 (.026)
Botulinic	.781 (.061)	.774 (.053)	.702 (.062)
Neurological	.924 (.015)	.915 (.017)	.830 (.023)
Other	.957 (.003)	.956 (.003)	.896 (.007)

Table 2. The AUC means (and standard deviations).

An examination of the misclassified triage diagnoses suggests three general causes for misclassification—misspelled words, use of nonstandard terminology, and compound complaints. Table 3 lists example triage diagnoses that were misclassified by the unigram model and the reason that each was misclassified. Triage diagnoses that contained misspelled words or nonstandard terms tended to be misclassified because the prior probabilities of these words were zero which, in turn, caused the posterior probabilities of such triage diagnoses to be zero. Similarly, triage diagnoses that contained multiple complaints were commonly misclassified because the prior probabilities of a majority of the words in such triage diagnoses were small for each syndrome, thereby resulting in small posterior probabilities.

The problems of misspelled words and nonstandard terminology can be addressed by translating the words in each triage diagnosis into an established vocabulary—e.g., by using a domain- and application-specific spell checker—but the problem of compound complaints would require separating each triage diagnosis into its component complaints. Focusing on the first two problems, a list of substitutions was created to correct misspellings and nonstandard terminology in the triage diagnoses used in this experiment. While specific to these triage diagnoses, the list of substitutions approximated the effect of a more sophisticated and automated approach that would normally be developed for use in a real-time data-analysis application. The list of substitutions was used to translate each triage diagnosis, and the ten-fold cross-validation experiment was re-run on the transformed triage diagnoses.

Triage Diagnosis	Misclassification Reason
cugh fever	misspelled word
nausea vomtiting	misspelled word
naus vomiting	nonstandard terminology
syncopal episode	nonstandard terminology
dizzy nausea	compound complaint
fever cough vomiting	compound complaint

Table 3. Examples of misclassified triage diagnoses.

Table 4 shows the mean (and standard deviation) of the AUC for each combination of syndrome and model from the ten-fold cross-validation experiment when using substitutions. The pairwise relationships of the mean AUC for each syndrome between models are parallel to those when substitutions were not used: the unigram model consistently outperformed both the bigram and mixture models. Moreover, one-tailed paired t-tests confirm that precisely the same pairwise differences in mean AUC between the models are significant at the same p levels as were significant when substitutions were not used (except for the difference in mean AUC between the unigram and bigram models for the Botulinic syndrome which is significant at the $p < .001$ level when using substitutions).

Syndrome	Unigram	Mixture	Bigram
Gastrointestinal	**.954 (.010)**	**.952 (.009)**	**.904 (.008)**
Constitutional	**.941 (.013)**	.927 (.015)	**.864 (.018)**
Respiratory	**.968 (.008)**	**.961 (.010)**	**.902 (.015)**
Rash	.922 (.020)	.917 (.018)	.842 (.052)
Hemorrhagic	.934 (.010)	.923 (.011)	.874 (.024)
Botulinic	.798 (.073)	.797 (.073)	.721 (.083)
Neurological	**.935 (.015)**	**.927 (.017)**	**.849 (.023)**
Other	**.964 (.001)**	**.963 (.002)**	**.910 (.006)**

Table 4. The AUC means (and standard deviations) when using substitutions. Bold values indicate statistically-significant increases in mean AUC.

The mean AUC for each combination of syndrome and model when using substitutions was uniformly higher than the corresponding mean AUC when not using substitutions. One-tailed paired t-tests confirm that the intra-model pairwise differences for all three models are significant at the $p < .001$ level for the Other, Gastrointestinal, Respiratory, and Neurological syndromes. Additionally, the intra-model pairwise differences for the unigram and

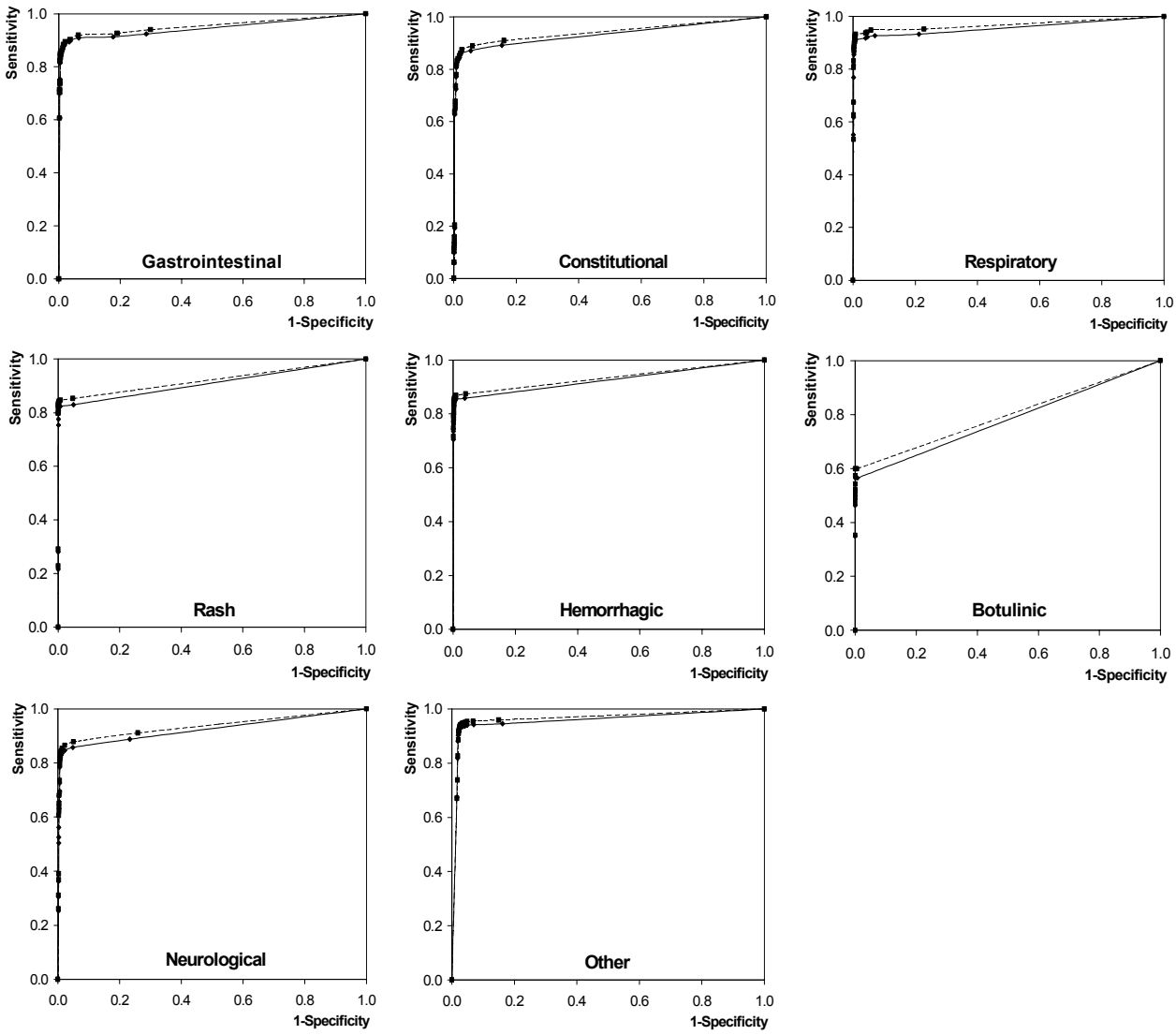

Figure 1. ROC curves when using (dashed lines) and not using (solid lines) substitutions for the unigram model.

bigram models are significant at the $p < .001$ level for the Constitutional syndrome. The bold values in Table 4 indicate statistically-significant increases in mean AUC as compared to the corresponding mean AUC when not using substitutions.

The ROC curves when using (dashed lines) and not using (solid lines) substitutions are illustrated in Figure 1 for the unigram model. The ROC curves when using substitutions are generally as good as or better than the ROC curves when not using substitutions. However, the effective improvements—while statistically significant in some cases—are small. The ROC curves for the mixture and bigram models exhibit similar pairwise differences.

Discussion

The large mean AUC values reported by the ten-fold cross-validation experiment demonstrate that it is possible to achieve high classification accuracies of triage diagnoses using a Bayesian classifier. The unigram model clearly outperformed the bigram model, suggesting that each word in a triage diagnosis has high information content independent of the other words. The poorer performance of the bigram model might be attributable to the relative increase in sparseness of the training set when considering word pairs as opposed to individual words; a larger training set might have improved the performance of the bigram model. While the unigram and mixture models performed comparably well for some syndromes, the relative simplicity of the unigram model makes it arguably preferable to the mixture model. It should be noted, however, that a different ratio of the unigram and bigram models in the mixture model might have performed better. A parameter estimation algorithm such as Expectation Maximization (EM) can be used to find a ratio that improves the classification accuracy (Mitchell 1997).

While the unigram model performed well overall, the classification accuracy was poorest for the Botulinic syndrome. As shown in Table 1, the number of triage diagnoses for this syndrome was much smaller than for the other syndromes. A larger number of triage diagnoses representative of the Botulinic syndrome would most likely result in an improvement in its classification accuracy.

The use of substitutions to correct misspellings and nonstandard terminology did increase the classification accuracies, however the relative improvement when compared to the effort needed to preprocess the data may make these transformations expensive. Such preprocessing may make sense only when maximal classification accuracy is critical or when easily performed with pre-existing software. The extent to which addressing the problem of compound complaints can improve classification accuracy remains an open question. Undoubtedly, separating each triage diagnosis into its component complaints would allow more precise models to be constructed and thus more accurate classification decisions to be made.

The early detection of epidemics relies upon the timely and reliable identification of increases in illnesses. Triage diagnoses and their associated ICD-9 codes are generally available earlier than final diagnoses, making them timelier sources of data for surveillance. Previous studies reported the classification accuracy using ICD-9 codes as having a sensitivity of 0.32 and a specificity of 0.99 for gastrointestinal illnesses (Ivanov et al. 2002), and as having a sensitivity of 0.44 and a specificity of 0.97 for respiratory illnesses (Espino and Wagner 2001). Using triage diagnoses, the classification accuracies with the unigram model without substitutions for the Gastrointestinal and Respiratory syndromes surpass this benchmark with sensitivities of 0.86 and 0.91 for specificities of 0.99 and 0.97, respectively. Since triage diagnoses have an earlier availability than final diagnoses and, at least for gastrointestinal and respiratory illnesses, a superior classification accuracy than ICD-9 codes, triage diagnoses are a better source of data for the early detection of epidemics.

Finally, an important issue which remains to be addressed is how well a classifier trained with triage diagnoses from one emergency department will classify those from another. As a first step, a set of triage diagnoses was similarly assembled from an emergency department in Pennsylvania and was used as the training set to classify triage diagnoses collected in Utah during the Olympic games (Tsui et al. 2002). While a cursory inspection indicated an acceptable performance, designed experiments are required to better investigate this issue of portability.

Acknowledgements

I would like to thank Drs. Wendy Chapman, Greg Cooper, Oleg Ivanov, Alon Lavie (Language Technologies Institute, Carnegie Mellon University), and Mike Wagner for the many enlightening discussions that helped direct the course of this work. I also would like to thank Dr. John Dowling for classifying the triage diagnoses and for providing the list of substitutions.

References

Bradley, A. P. 1997. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition* 30(7):1145-1159.

Espino, J. U.; and Wagner, M. M. 2001. Accuracy of ICD-9-Coded Chief Complaints and Diagnoses for the Detection of Acute Respiratory Illness. In *Proc of the AMIA Annual Symposium*, 164-168. Philadelphia: Hanley & Belfus, Inc.

Hanley, J. A.; and McNeil, B. J. 1982. The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143(1):29-36.

Ivanov, O.; Wagner, M. M.; Chapman, W. W.; et al. 2002. Accuracy of Three Classifiers of Acute Gastrointestinal Syndrome for Syndromic Surveillance. In *Proc of the AMIA Annual Symposium*, 345-349. Philadelphia: Hanley & Belfus, Inc.

Manning, C. D.; and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Mitchell, T. M. 1997. *Machine Learning*. New York: McGraw-Hill.

Tsui, F.-C.; Espino, J. U.; Wagner, M. M.; et al. 2002. Data, Network, and Application: Technical Description of the Utah RODS Winter Olympic Biosurveillance System. In *Proc of the AMIA Annual Symposium*, 815-819. Philadelphia: Hanley & Belfus, Inc.

Tsui, F.-C.; Wagner, M. M.; Dato, V.; et al. 2001. Value of ICD-9-Coded Chief Complaints for Detection of Epidemics. In *Proc of the AMIA Annual Symposium*, 711-715. Philadelphia: Hanley & Belfus, Inc.

Wagner, M. M.; Tsui, F.-C.; Espino, J. U.; et al. 2001. The Emerging Science of Very Early Detection of Disease Outbreaks. *J Pub Health Management Practice* 7(6):51-59.

Zweig, M. H.; and Campbell, G. 1993. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chem* 39(4):561-577.