Identifying Wrist Fracture Patients with High Accuracy by Automatic Categorization of X-ray Reports

BERRY DE BRUIJN, PhD, ANN CRANNEY, MD, MSc, SIOBHAN O’DONNELL, MSc, JOEL D. MARTIN, PhD, ALAN J. FORSTER, MD, MSc

Abstract The authors performed this study to determine the accuracy of several text classification methods to categorize wrist x-ray reports. We randomly sampled 751 textual wrist x-ray reports. Two expert reviewers rated the presence \((n = 301)\) or absence \((n = 450)\) of an acute fracture of wrist. We developed two information retrieval (IR) text classification methods and a machine learning method using a support vector machine (TC-1). In cross-validation on the derivation set \((n = 493)\), TC-1 outperformed the two IR based methods and six benchmark classifiers, including Naive Bayes and a Neural Network. In the validation set \((n = 258)\), TC-1 demonstrated consistent performance with 93.8\% accuracy; 95.5\% sensitivity; 92.9\% specificity; and 87.5\% positive predictive value. TC-1 was easy to implement and superior in performance to the other classification methods.


Methods

Material

Two expert reviewers (AC, SO’D) classified 751 x-ray reports describing findings of a forearm, wrist, or hand x-ray as consistent with the presence or absence of an acute wrist fracture. Eligibility criteria included patients 40 years or older, seen at the Ottawa Hospital (an academic hospital in Eastern Ontario, Canada) between October 1, 2003 and October 31, 2004. From the entire set of x-ray reports in this period, we randomly selected 493 for derivation (included first x-ray report in the study period) and 258 for validation (included first and last x-ray report). We used this selection method for the validation set as we felt that, in future use, the text classifier might need to evaluate repeat x-ray reports. The prevalence of acute wrist fractures was 43.2\% (213/493) and 34.1\% (88/258) in the derivation and validation set respectively. The average length of all documents was 64 words, 95\% had fewer than 150 words.

Evaluation

We calculated the accuracy of the different text categorization methods for correctly mimicking our manual classifications. Metrics were accuracy, sensitivity (or recall), specificity, and positive predictive value (PPV, or precision). We used a ten-fold cross-validation design that involved randomly assigning reports to 1 of 10 sections or ‘folds’; the model is then trained on 9 folds and tested on the 10th. Training and testing is repeated ten times so that each fold serves as a test set once. Differences in accuracy between classifiers in the 10-fold cross-validations were tested for significance with paired t-tests, pairing the folds.

Phase I

We implemented three text classification methods to automatically categorize the x-ray reports. The first used an information retrieval algorithm \((IR-I)\) with a set of key terms compiled by a radiologist. This lexicon contained 57
Table 1 • Performance of IR-1, IR-1’, and TC-1 in the Derivation Set, and Validation Set (TC-1)

<table>
<thead>
<tr>
<th>X-ray Reports</th>
<th>Method</th>
<th>TP FP</th>
<th>FN TN</th>
<th>Sensitivity, % (95% CI)</th>
<th>Specificity, % (95% CI)</th>
<th>PPV, % (95% CI)</th>
<th>Accuracy, % (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derivation set (n = 493)</td>
<td>IR-1</td>
<td>110</td>
<td>103</td>
<td>51.6 (45.0–58.3)</td>
<td>97.5 (94.9–98.8)</td>
<td>94.0 (88.2–97.1)</td>
<td>77.7 (73.7–81.1)</td>
</tr>
<tr>
<td></td>
<td>IR-1’</td>
<td>7</td>
<td>273</td>
<td>68.1 (61.6–74.0)</td>
<td>91.8 (88.0–94.5)</td>
<td>86.3 (80.3–90.7)</td>
<td>81.5 (77.9–84.7)</td>
</tr>
<tr>
<td></td>
<td>IR-2</td>
<td>23</td>
<td>257</td>
<td>87.3 (82.2–91.1)</td>
<td>91.1 (87.2–93.9)</td>
<td>88.2 (83.1–91.8)</td>
<td>89.7 (86.4–91.9)</td>
</tr>
<tr>
<td></td>
<td>TC-1</td>
<td>25</td>
<td>255</td>
<td>93.9 (89.8–96.4)</td>
<td>93.6 (90.1–95.9)</td>
<td>91.7 (87.3–94.7)</td>
<td>93.7 (91.2–95.5)</td>
</tr>
<tr>
<td>Validation set (n = 258)</td>
<td>TC-1</td>
<td>84</td>
<td>4</td>
<td>95.5 (88.9–98.2)</td>
<td>92.9 (88.1–95.9)</td>
<td>87.5 (79.4–92.7)</td>
<td>93.8 (90.2–96.2)</td>
</tr>
</tbody>
</table>

TP/FN/FP/TN = True/False Pos/Negative; Sensitivity = Recall, PPV = positive predictive value-precision. 95% confidence intervals are calculated using the Wilson procedure.

terms (words or phrases) indicative of an acute wrist fracture. Reports containing any of these predefined terms were classified as ‘positive.’ To improve sensitivity, we truncated the terms to 4 characters and appended them with a wildcard character (method IR-1’).

For the second technique (IR-2), we used terms selected by a computer based algorithm instead of an expert. This feature selection algorithm used Information Gain scores which are based on the number of times words or phrases are observed in a positive or a negative document. The terms with the highest Information Gain were selected as features.

For the third technique (TC-1), we used a machine-learning algorithm called a ‘support vector machine’ (SVM) which has been successful in data categorization. Information Gain scores were used to select features for all reports, as with IR-2. The SVM then used these features to create a decision criterion between ‘acute wrist fracture’ or ‘no such fracture’ for new reports. Various parameter settings for the SVM and the feature selection phase were tested.

Phase II
We compared the TC-1 results with 6 other machine learning algorithms, as available in a benchmark toolkit named Weka. Weka runs were repeated with various feature set sizes, including a small (50) and large feature set (5000).

Phase III
We validated TC-1 on a separate set of x-ray reports. This validation set was kept quarantined until parameter optimizing and training of the categorization model was completed.

Results
Phase I
Although IR-1 showed high precision (or PPV) on the derivation set (94.0%, Table 1) it missed almost half of the wrist fracture cases (sensitivity = 51.6%). The overall accuracy of IR-1 was 77.7%. Using wildcards (IR-1’) improved the sensitivity to 68.1% at the expense of PPV (86.3%). Accuracy rose to 81.5%. IR-2 had higher sensitivity (87.3%), while the specificity remained high at 91.1% and PPV slipped to 88.2%. IR-2 achieved an accuracy of 89.5%. Applying the truncation and wildcard strategy within IR-2 did not improve its performance (results not shown). The TC-1 method was significantly more accurate than the IR methods achieving an accuracy of 93.7% (p < 0.001, paired t-test). The sensitivity and PPV were very good at 93.9% and 91.7% at the threshold setting that optimizes accuracy.

The optimal parameters for the SVM were identified as: linear kernel, with the cost parameter that balances errors with a hard margin set to 0.005. Features were strings of one to four words, no truncation was done and a stop word could not be a feature by itself but could be part of a multi-word feature. A maximum of 5,000 features was used.

The cut-off (threshold) parameter was set to optimize accuracy; this parameter can be adjusted so that false-positives and false-negatives trade off.

Phase II
Using the same derivation set and cross-validation design, accuracies for the Weka machine learning classifiers were: 84.6% (AdaBoost), 86.8% (K-Start), 87.2% (J48), 87.4% (Naive Bayes), 87.6% (J48), and 88.0% (Multi-layer Perceptron neural network). All were significantly lower than TC-1 accuracy (paired t-test, p < 0.001). The SMO algorithm in Weka, an equivalent implementation of SVMs, achieved the same result as TC-1.

Phase III
TC-1 was applied to the previously quarantined validation set, after it was retrained on the entire derivation set using the optimal parameter settings found in phase 1. Accuracy of TC-1 was 93.8% at the default threshold (Table 1). A comparison of the derivation and validation ROC curves demonstrated that the TC-1 method discriminated between positive and negative wrist fractures reports as accurately in both sets.

Discussion
We developed and validated an SVM based text classifier method (TC-1) which accurately categorizes textual wrist x-ray reports into those with and without an acute wrist fracture. Operating at 94% accuracy, TC-1 appears to be as or more accurate than classifiers used on x-ray reports in other studies such as Dreyer et al and Thomas et al, although a direct comparison of results is limited due to different study designs and data sets. TC-1 did outperform a number of well reputed classifiers on the same data set, including Naive Bayes and a neural network. TC-1 test characteristics stayed consistent when applied to a separate validation set of x-ray reports, despite the different prevalence of positive x-rays.
An additional strength is that the TC-1 algorithm does not rely on domain specific vocabularies or manual input. Instead, it learns from pre-categorized examples using only superficial textual features—which does have as its drawback that domain experts may need to invest time in labeling examples.

We believe that the system has proven accurate enough to use in automated decision support for osteoporosis management of wrist fracture patients. Its characteristics make it a strong candidate for use in other clinical settings. Future work includes validation of TC-1 for other fractures (i.e., hip and spine), and validation of TC-1 in a real world hospital setting.

References