MEDINFO 2015: eHealth-enabled Health
I.N. Sarkar et al. (Eds.)

doi:10.3233/978-1-61499-564-7-1070

Evaluating Methods for Identifying Cancer in Free-Text Pathology Reports Using Various Machine Learning and Data Preprocessing Approaches

Suranga Nath Kasthurirathne, BEng[a], Brian E. Dixon, MPA, PhD[b,c], Shaun J. Grannis, MD, MS[b,d]

[a]*Indiana University School of Informatics and Computing, Indianapolis, IN, USA*
[b]*Regenstrief Institute, Indianapolis, IN, USA*
[c]*Indiana University Fairbanks School of Public Health, Indianapolis, IN, USA*
[d]*Indiana University School of Medicine, Indianapolis, IN, USA*

Abstract

Automated detection methods can address delays and incompleteness in cancer case reporting. Existing automated efforts are largely dependent on complex dictionaries and coded data. Using a gold standard of manually reviewed pathology reports, we evaluated the performance of alternative input formats and decision models on a convenience sample of free-text pathology reports. Results showed that the input format significantly impacted performance, and specific algorithms yielded better results for presicion, recall and accuracy. We conclude that our approach is sufficiently accurate for practical purposes and represents a generalized process.

Keywords:

Public health reporting; decision models; ontologies; cancer; pathology; data preprocessing.

Introduction

Cancer case reporting is often delayed and incomplete [1]. Automated methods for identifying public health reportable cases can address this issue [2], yet a substantial amount of cancer case-related data are captured as free-text making it challenging to interpret [3]. We sought to assess approaches to identify cancer cases from free-text pathology reports to (a) determine whether we could achieve acceptable accuracy using a generalizable approach that does not require complex dictionaries, grammars or ontologies; (b) compare various candidate decision models; and (c) evaluate how data input format affects decision model accuracy.

Methods

We identified seven keywords associated with the presence of cancer in pathology reports. Each free text report was parsed and separate counts tabulated for the presence of each keyword either in positive or negated contexts using the Negex algorithm. We evaluated two preprocessed data input vectors. The first input vector ("raw count") contained positive counts (C_p) and negated counts (C_n) for keywords in each report. The second ("four-state") reduced these to a single value per keyword: 1=($C_p > C_n$); 2=($C_n > C_p$); 3=($C_p = C_n$) and 4=keyword absent. We evaluated logistic regression, naïve bayes (NB), k-nearest neighbor, random forest (RF), and J48 decision tree decision models implemented in Weka software version 3.6.10. The precision, recall, and accuracy was calculated for each input format combination.

Results

Each decision model and input format combination yielded satisfactory results. However, the "raw count" input format outperformed the "four-state" input for all three performance measures. The NB decision model produced statistically significant lower results for accuracy ($p < 0.01$); the remaining methods showed no difference as a group. For recall, all decision models showed no difference as a group. For precision, both RF and NB showed lower values ($p < 0.01$); the remaining methods were indistinguishable.

Discussion

Overall results indicated that the "raw count" input format outperformed the "four-state" format. Although we achieved reasonable performance while avoiding the use of complex dictionaries or ontologies, this approach occasionally failed to identify cases when text reports contained only disease specific terms and the seven generic keywords were absent. We conclude that our approach represents a generalized process that can be adapted for many additional clinical use cases, and is accurate enough for practical purposes.

References

[1] Zanetti R, Schmidtmann I, Sacchetto L, Binder-Foucard F, Bordoni A, Coza D, et al. Completeness and timeliness: Cancer registries could/should improve their performance. European journal of cancer. 2014.

[2] Overhage JM, Grannis S, McDonald CJ. A comparison of the completeness and timeliness of automated electronic laboratory reporting and spontaneous reporting of notifiable conditions. American journal of public health. 2008;98(2):344.

[3] Fidahussein M, Friedlin J, Grannis S. Practical challenges in the secondary use of real-world data: the notifiable condition detector. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:402-8.

Address for correspondence

S.N. Kasthurirathne, Indiana University School of Informatics and Computing, 535 W. Michigan Street, IT 475, Indianapolis, IN 46202, USA, (317)-278-4636, snkasthu@iupui.edu