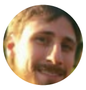
Automated Classification of Pathology Reports

Conference Paper *in* Studies in health technology and informatics · August 2015

DOI: 10.3233/978-1-61499-564-7-1040

3 authors:



Michel Oleynik
Medical University of Graz
15 PUBLICATIONS 5 CITATIONS
SEE PROFILE

Diogo Patrão
Hospital A. C. Camargo
25 PUBLICATIONS 323 CITATIONS
SEE PROFILE

Marcelo Finger
University of São Paulo
125 PUBLICATIONS 1,079 CITATIONS
SEE PROFILE

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

doi:10.3233/978-1-61499-564-7-1040

Automated Classification of Pathology Reports

Michel Oleynik[a], Marcelo Finger[a], Diogo F. C. Patrão[b]

[a] *Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil*
[b] *International Center for Research, A. C. Camargo Cancer Center, São Paulo, Brazil*

Abstract

This work develops an automated classifier of pathology reports which infers the topography and the morphology classes of a tumor using codes from the International Classification of Diseases for Oncology (ICD-O). Data from 94,980 patients of the A.C. Camargo Cancer Center was used for training and validation of Naive Bayes classifiers, evaluated by the F_1-score. Measures greater than 74% in the topographic group and 61% in the morphologic group are reported. Our work provides a successful baseline for future research for the classification of medical documents written in Portuguese and in other domains.

Keywords:

Natural Language Processing; Medical Records; Data Mining.

Introduction

Pathology reports are an important data source for diagnosis in the cancer domain because reports provide clinical evidence on the topography of the tumor, its histological type, and morphology. The information may be encoded with the International Classification of Diseases for Oncology (ICD-O). A recent work [1] evaluated the use of Naive Bayes classifiers and Support Vector Machines, and we applied a modified version of these techniques on pathology reports written in Portuguese and assessed its efficiency.

Methods

We used a collection of pathology reports written in Portuguese from the A.C. Camargo Cancer Center in São Paulo, Brazil. The dataset comprises 94,980 patients from the years 1196 to 2010.

We applied a multinomial Naive Bayes classifier [2], seen in Figure 1, with add-alpha smoothing to the resulting data (we chose α = 0.5 since our belief in uniformity is weak).

$$c_{map} = \operatorname*{argmax}_{c\in C}[\log\hat{P}(c) + \sum_{1\le k\le n_d}\log\hat{P}(t_k \mid c)]$$

Figure 1 – Naïve Bayes Classifier

Results

While one work [1] reported 71.5% F_1-measure on the classification of 26 topographic groups, we presented a higher F_1-measure (75.1%) on a task of 16 groups. On the morphology axis of the ICD-O classification, we achieved a lower F_1-measure (61.3% *versus* 85.4%) on a wider classification task (49 *versus* 18 groups). The results were micro-averaged with 10-fold cross-validation.

Table 1 – Micro-averaged efficiency (%) of Naive Bayes classifiers trained on increasing levels of detail of the ICD-O

	Topography			**Morphology**		
	P	**R**	**F_1**	**P**	**R**	**F_1**
Group	78.928	73.451	75.304	68.813	60.172	62.256
Category	73.180	69.284	69.284		N/A	
Concept	43.239	40.820	39.941	49.151	39.705	40.673

When analyzing the classifier efficiency on the topographic group, we achieved 98.3% and 97.3% precision on the groups *C50: Breast* and *C60-C63: Male genital organs*, respectively the most common cancer for women and men. We also analyzed the resulting confusion matrix and observed that the most common cause of error is the incorrect classification of *C50: Breast* as *C51-C58: Female genital organs*. This is probably due to the fact that the diagnosis of breast cancer is usually accompanied by a screening test for cervical cancer.

Conclusion

Our work has immediate implications to knowledge discovery and statistical analysis in the medical domain, as it eases the process of obtaining structured information over textual data. It also accelerates the work of physicians when classifying patient data.

Acknowledgments

This research was funded by the Brazilian National Research Council - CNPq (project number 133862/2010-0) and is approved by the committee on ethics on research of the A.C. Camargo Cancer Center (registered under number 1418/10).

References

[1] Jouhet V, Defossez G, Burgun A, Le Beux P, P. Levillain, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. Methods of Information in Medicine. 2012;51(3):242.

[2] Pearl J. Bayesian Networks: a model of self-activated: memory for evidential reasoning. Computer Science Department, University of California; 1985.