Informatics for Health: Connected Citizen-Led Wellness and Population Health
R. Randell et al. (Eds.)
© 2017 European Federation for Medical Informatics (EFMI) and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-753-5-256

Automated Classification of Semi-Structured Pathology Reports into ICD-O Using SVM in Portuguese

Michel OLEYNIK[a,1], Diogo F. C. PATRÃO[b] and Marcelo FINGER[c]
[a] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[b] *International Center for Research, A.C.Camargo Cancer Center, Brazil*
[c] *Institute of Mathematics and Statistics, University of São Paulo, Brazil*

Abstract. Pathology reports are a main source of information regarding cancer diagnosis and are commonly written following semi-structured templates that include tumour localisation and behaviour. In this work, we evaluated the efficiency of support vector machines (SVMs) to classify pathology reports written in Portuguese into the International Classification of Diseases for Oncology (ICD-O), a biaxial classification of cancer topography and morphology. A partnership program with the Brazilian hospital A.C. Camargo Cancer Center provided anonymised pathology reports and structured data from 94,980 patients used for training and validation. We employed SVMs with tf-idf weighting scheme in a bag-of-words approach and report F_1 score of 0.82 for 18 sites and 0.73 for 49 morphology classes. With the largest dataset ever used in such a task, our work provides reliable estimates for the classification of pathology reports in Portuguese and agrees with a few similar studies published in the same kind of data in other languages.

Keywords. Natural language processing, pathology report, support vector machines

1. Introduction

Clinical reports are usually written in natural language due to its descriptive power and ease of communication among specialists. Processing data for knowledge discovery and statistical analysis requires information retrieval techniques, already established for newswire texts, but still in development in the medical subdomain. Some studies [1,2] explored mapping techniques to obtain structured information from clinical data, usually mixing sets of rules with machine learning. Although the results are promising, major efforts are required to build medical corpora and to adapt general language rules to the clinical domain.

Pathology reports are a main source of information regarding cancer diagnosis [3] and are commonly written following semi-structured templates that include tumour localisation and behaviour. Since structured data are mostly not available from the electronic health record (EHR) with sufficient accuracy and completeness, cancer

[1] Corresponding author: Michel Oleynik, Institute for Medical Informatics, Statistics and Documentation, Auenbruggerplatz 2, 8036 Graz, Austria; E-mail: michel.oleynik@stud.medunigraz.at.

registries play an important role as containers of manually reviewed clinical content at patient level, in order to report cancer statistics to authorities by using controlled vocabularies. In this process, health professionals often employ the International Classification of Diseases for Oncology (ICD-O) [3], a biaxial classification of cancer topography and morphology maintained by the World Health Organization (WHO).

Probably the first work to evaluate the automated classification of pathology reports into more than one class, Martinez and Li [4] explored a Naïve Bayes classifier with named entities as features to classify 217 reports written in English into 11 different sites. They report a micro-averaged F_1 score of 0.58. Jouhet et al. [5] employed Naïve Bayes classifiers and Support Vector Machines (SVMs) in 5,121 French free-text pathology reports to classify them into the two ICD-O axes using ngrams as features. They reported an F_1 score of 0.72 for 26 topographic sites and 0.85 for 18 morphology classes in the ICD-O code attribution task. Later, Kavuluru et al. [6] applied Naïve Bayes classifiers, SVMs and Logistic Regression to a dataset of 56,426 English pathology reports in order to classify them into 57 primary sites from the ICD-O. They compared the efficiency of unigrams, bigrams and named entities as features and reported an F_1 score of 0.90. More recently, Oleynik et al. [7] applied Naïve Bayes classifiers to a set of pathology reports written in Portuguese and obtained $F_1 = 0.75$ for the recognition of 16 topographies and $F_1 = 0.62$ for 49 morphologies.

In this paper, we applied SVM to a large dataset of pathology reports in Portuguese and assessed its efficiency. We report results in the two ICD-O axes, viz. topography and morphology. To our knowledge, there are no previous studies related to the same language, method and type of data.

2. Materials and Methods

2.1. Pathology Reports and Cancer Registry Corpora

A partnership program with the Brazilian hospital A.C. Camargo Cancer Center provided anonymised pathology reports and structured data from 94,980 patients used for training and validation. The documents were created during routine operation between 1996 and 2010, with their text structure following the institution's editorial guidelines.

In order to train a supervised machine learning classifier, we programmatically unified reports of the same patient and associated their content using the patient identifier to the data available in cancer registries. These registries include manually encoded information of the cancer topography, morphology and the metastatic status. In the next step, we discarded those with confirmed metastasis or multiple classifications and used the structured information in the cancer registries as the classifier target.

The resulting dataset maps patients into the two ICD-O axes, viz. topography with n = 18,905 patients (18 code groups) and morphology with n = 18,599 patients (49 code groups). Due to the nature of the problem, patients are not uniformly distributed in the groups, as can be seen on the data presented in Section 3.

2.2. Support Vector Machines

In a pre-processing step, we lowercased all tokens (extracted with the Java `StringTokenizer` class) and kept only the remaining 5,000 most frequent ones to speed up

processing and reduce model overfitting. We then applied a Support Vector Machine (SVM) over the vector space representation of the data (in a bag-of-words approach), with tf-idf weighting scheme[2] and a linear kernel. SVMs as such are known to produce good results in text classification [8].

A SVM is a discriminative and non-probabilistic classifier that tries to maximise the decision margin between two given classes [8]. The decision function, seen in Eq. (1), assigns either $+1$ or -1 to an input vector $\vec{x}$ based on the decision hyperplane normal vector $\vec{w}$ and an intercept term b.

$$f(\vec{x}) = sign(\vec{w}^T \vec{x} + b) \quad (1)$$

We used Weka 3.6.6 [9] for most of the steps and LibSVM 3.17 [10] to perform SVM calculations under a one-*versus*-all approach, common in multi-class classification tasks. A regularisation parameter $C = 2^{-7}$ was empirically determined following [11].

3. Results

Tables 1 and 2 show a breakdown of sample size (n), precision (P), recall (R) and F_1 score (F_1) obtained via 10-fold cross-validation for the ten topographies and morphologies with best results, respectively. The last row of each table shows the overall micro-averaged efficiency measures. Due to space limitations, full tables and confusion matrices are only available online at https://goo.gl/iG41Ok and https://goo.gl/cTrJfI.

Table 1. Top ten F_1 scores in the ICD-O topography attribution task.

Code Group	Description	n	P	R	F_1
C44	Skin	3,858	0.88	0.94	0.91
C50	Breast	3,668	0.89	0.91	0.90
C73-C75	Thyroid and other endocrine glands	1,329	0.92	0.87	0.90
C60-C63	Male genital organs	1,536	0.93	0.81	0.87
C64-C68	Lymph nodes	660	0.86	0.78	0.82
C51-C58	Female genital organs	1,574	0.85	0.77	0.81
C69-C72	Eye, brain and other parts of central nervous system	536	0.83	0.70	0.76
C00-C14	Lip, oral cavity and pharynx	903	0.80	0.71	0.75
C15-C26	Digestive organs	2,159	0.67	0.84	0.75
C77	Lymph nodes	590	0.68	0.80	0.74
	Overall	**18,905**	**0.82**	**0.82**	**0.82**

[2] Tf-idf stands for *term frequency–inverse document frequency*.

Table 2. Top ten F_1 scores in the ICD-O morphology attribution task.

Code Group	Description	n	P	R	F_1
959-972	Hodgkin and non-Hodgkin lymphomas	859	0.85	0.87	0.86
850-854	Ductal and lobular neoplasms	3,410	0.85	0.87	0.86
855	Acinar cell neoplasms	1,059	0.87	0.85	0.86
809-811	Basal cell neoplasms	1,704	0.80	0.89	0.84
872-879	Nevi and melanomas	1,473	0.87	0.81	0.84
906-909	Germ cell neoplasms	208	0.89	0.71	0.79
812-813	Transitional cell papillomas and carcinomas	384	0.81	0.74	0.78
938-948	Gliomas	237	0.82	0.71	0.76
858	Thymic epithelial neoplasms	17	1.00	0.59	0.74
868-871	Paragangliomas and glomus tumors	26	1.00	0.58	0.73
	Overall	**18,599**	**0.74**	**0.74**	**0.73**

4. Discussion

Results obtained in the topography attribution task ($F_1 = 0.82$) were better than the ones in the morphology attribution task ($F_1 = 0.73$). As expected, the evaluated efficiency is better at simpler tasks, where the number of target classes is lower (18 *versus* 49). The results might be dependent on the non-uniform distribution of classes in the dataset, with the best results reported on the most common classes. Additionally, the most common cancer in women (C50: Breast) and men (C60-C63: Male genital organs) achieved high precision rates (0.89 and 0.93, respectively).

Moreover, the analysis of the confusion matrix reveals that the most common mistake in the topography attribution task is the classification of C51-C58: Female genital organs as C15-C26: Digestive organs, which also accounts for its low precision (0.67) and is responsible for 5% (171/3380) of the incorrect classified patients. In the morphology axis, classification of 805-808: Squamous cell neoplasms as 814-838: Adenomas and adenocarcinomas is the most frequent source of error and accounts for 8% (381/4796) of the misclassifications.

Although a strict comparison is hard due to the lack of public data in the medical domain, SVM shows improvements over the Naïve Bayes approach of Martinez and Li [4] applied to the same kind of data. While they report an F_1 score of 0.58 with 11 sites, we achieved a higher F_1 score of 0.82 with seven more classes. Likewise, our study achieved results comparable to those reported by Jouhet et al. [5]. Even though we report a lower F_1 score of 0.74 (*versus* 0.85) with 31 more classes in the morphology code attribution task, we achieved a higher F_1 score of 0.82 (*versus* 0.72) over only eight less topography classes. Compared to the work of Kavuluru et al. [6], we obtained a lower F_1 score (0.82 *versus* 0.93) with four more target classes. Lastly, SVM performed much better than a prior work done in the same dataset with Naïve Bayes classifiers [7], with an F_1 score improvement of 0.07 and 0.11 in the topography and morphology groups, respectively. One common approach in these studies is to remove rare groups and therefore achieve better efficiency rates. In contrast, we report results in the standard set of 18 topography groups and 49 morphology groups, as defined by the WHO [3].

5. Conclusion

Although our classifier is in general agreement with other works reported in literature, we can see some limitations. A more precise analysis would have been done if we had classified a subset of the reports by a team of specialists without access to other patient data. Assessing Cohen's kappa factor among them would provide a smaller upper bound to the algorithm and perhaps reproduce high discordance rates described in literature. The unavailability of such a team for an extended period also grounds our automated process. Nonetheless, we could have tested other learning models known for providing better results, like neural networks. However, given its elevated complexity and high computational cost, we chose a simpler and more manageable approach.

Our study may improve recall rates in tasks such as cohort building for clinical trials, as it creates additional structured information over textual data. Moreover, it could be employed to ease manual classification of pathology reports via the generation of probability ordered code lists. The research showed that the automatic classification of pathology reports is not only feasible, but also achieves high efficiency rates comparable to those found in similar papers. We believe that our work provides a successful baseline for future research, not only for the classification of medical documents written in Portuguese, but also to be extended and applied to other domains.

Acknowledgments and Legal Aspects

We would like to thank Prof. Stefan Schulz for the paper revision. Our work is funded by the Brazilian National Research Council - CNPq (project number 206892/2014-4) and is approved by the committee on ethics on research of the A.C. Camargo Cancer Center (registered under number 1418/10).

References

[1] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program", in *Proceedings of the AMIA Symposium*, p. 17, AMIA, 2001.

[2] C. Friedman, P. O. Alderson, J. H. M. Austin, J. J. Cimino, and S. B. Johnson, "A general natural-language text processor for clinical radiology", *JAMIA*, vol. 1, no. 2, p. 161, 1994.

[3] A. Fritz, C. Percy, A. Jack, K. Shanmugaratnam, L. Sobin, D. M. Parkin, and S. Whelan, eds., *International Classification of Diseases for Oncology*. WHO Press, third ed., 2000.

[4] D. Martinez and Y. Li, "Information extraction from pathology reports in a hospital setting", in *Proceedings of the 20th CIKM*, pp. 1877–1882, ACM, 2011.

[5] V. Jouhet, G. Defossez, A. Burgun, P. Le Beux, P. Levillain, P. Ingrand, V. Claveau, et al., "Automated classification of free-text pathology reports for registration of incident cases of cancer", *Methods of Information in Medicine*, vol. 51, no. 3, p. 242, 2012.

[6] R. Kavuluru, I. Hands, E. Durbin, and L. Witt, "Automatic extraction of ICD-O-3 primary sites from cancer pathology reports", in *Clinical Research Informatics AMIA symposium (forthcoming)*, 2013.

[7] M. Oleynik, M. Finger, and D. Patrão, "Automated classification of pathology reports", *StudHealth Technol Inform*, vol. 216, p. 1040, 2015.

[8] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features", in *European conference on machine learning*, pp. 137–142, Springer, 1998.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update", *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[10] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[11] C.-W. Hsu, C.-C. Chang, C.-J. Lin, *et al.*, "A practical guide to support vector classification", 2003.