MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Multi-label Text Classification of German Language Medical Documents

Stephan Spat[a], Bruno Cadonna[a], Ivo Rakovac[a], Christian Gütl[b], Hubert Leitner[c], Günther Stark[c], Peter Beck[a]

[a] *Institute of Medical Technologies and Health Management, JOANNEUM RESEARCH Forschungsgesellschaft mbH, Graz, Austria*
[b] *Institute for Information Systems and Computer Media, Graz University of Technology, Graz, Austria*
[c] *Steiermärkische Krankenanstaltenges. m.b.H., Graz, Austria*

Abstract

At nearly every patient visit, medical documents are produced and stored in a medical record, often in an unstructured form as free text. The growing amount of stored documents increases the need for effective and timely retrieval of information. We developed a multi-label text classification system to categorize free text medical documents (e.g. discharge letters, clinical findings, reports) written in German into predefined classes. A random sample of 1,500 free text medical documents was retrieved from a general hospital information system and was manually assigned to 1 to 8 categories by a domain expert. This sample was used to train and evaluate the performance of 4 classification schemes: Naïve Bayes, k-NN, SVM, and J48. Additional tests of the effect of text preprocessing were done. In our study, preprocessing improved the performance, and best results were obtained by J48 classification.

Keywords:

machine learning, classification, medical records, multi-label

Introduction

At nearly every patient contact with healthcare-providers, medical documentation is generated and stored in medical or nursing records, often as free text. With the increasing amount of stored, unstructured free text information, the need for effective and timely retrieval of relevant information is growing. In this work, we describe the development and the evaluation of an information system for multi-label classification of medical documents into predefined classes.

Methods

A random sample of 1,500 unstructured, free text documents written in German was extracted from an electronic medical record (EMR) of a general hospital in Austria. A domain expert (physician) manually classified the retrieved documents into one or more of the following classes: surgery, vascular surgery, casualty surgery, internal medicine, neurology, anesthesia and intensive care, radiology and physiotherapy. In average, 1.47 labels were assigned to a document. We built an automated multi-label text classification system in Java based on Weka [1], an open-source machine-learning framework. Four different kinds of classification schemes were compared: Naïve Bayes, k-NN, SVM and J48. 10-fold cross validation was used for evaluation. Moreover, the influence of text preprocessing (e.g. stop-word-removal, stemming, lowercasing) was studied.

Results

Evaluation results for the F-measure [2] with and without text preprocessing are illustrated in Figure 1.

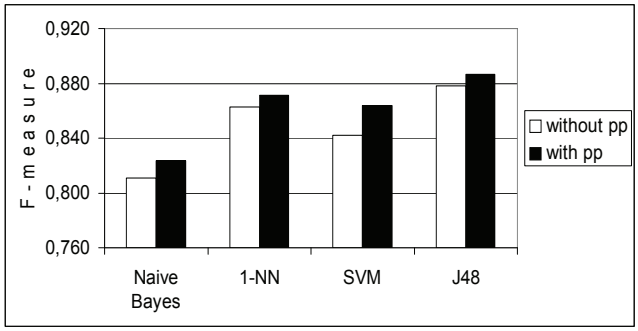

Figure 1 - F-measure with and without text preprocessing (pp)

Evaluation results *without* text preprocessing:

J48 classification model achieved best results with an F-measure of 0.877. The F-measure of the 1-NN model is 0.864 and SVM obtained a value of 0.850. Naïve Bayes achieved the worst result with a value of 0.811.

Evaluation results *with* text preprocessing:

Best result achieved again the J48 classification model with an F-measure of 0.886. 1-NN and SVM are middle-ranking models with values of 0.871 and 0.864. Naïve Bayes achieved an F-measure of 0.824, which was again the worst result of the models. Table 1 and shows the

improvement of the F-measure in percentage caused by text preprocessing

Table 1 - Improvement of the F-measure in percentage caused by text preprocessing (pp).

F-measure	Naïve Bayes	1-NN	SVM	J48
without text pp	0.811	0.864	0.850	0.877
with text pp	0.824	0.871	0.864	0.886
improvement	**+ 1.53** %	**+ 0.83** %	**+ 1.61** %	**+ 1.05** %

In order to compare classification models regarding their achieved F-measures more clearly, differences are given in Table 2. For this comparison F-measures are calculated for classifications with text preprocessing. For example, the 1-NN model achieved an F-measure better by 5.76 % than Naïve Bayes and the F-measure of Naïve Bayes is 4.66 % worse than the F-measure achieved by the SVM model.

Table 2 - Differences in percentages of F-measures achieved by evaluated classification models for classifications with text preprocessing

	Naïve Bayes	1-NN	SVM	J48
Naïve Bayes	0.00 %	**- 5.45** %	**- 4.66** %	**- 7.07** %
1-NN	**+ 5.76** %	0.00 %	**+ 0.84** %	**- 1.72** %
SVM	**+ 4.88** %	**- 0.83** %	0.00 %	**- 2.53** %
J48	**+ 7.61** %	**+ 1.75** %	**+ 2.60** %	0.00 %

Discussion

Evaluation results indicate that J48, SVM and 1-NN are preferable to Naïve Bayes. This finding corresponds to the results of [3]. Text preprocessing increased F-measures for all models. A reason for the rather small effect of text preprocessing on the evaluation results might be medical language. Medical language is quite specific with its own grammar and terms. Hence, for example standard stop word removal or stemming might not be that effective. [4] states that the result of text preprocessing strongly depends on the underlying terms. Thus, text preprocessing might be improved by using domain specific stop word lists and adapted stemming algorithms.

Conclusion

Results show that it is possible to classify medical documents written in German originating from a general hospital with automated machine-learning classification schemes with promising results, comparable with [5].

The classification system is used in a prototype of an information retrieval system for score-calculation, thus influencing the display order of search results. Further studies are needed to evaluate the accuracy of the developed system in other hospitals as well as the user-perceived benefits of this prototype.

References

[1] Witten IH, Frank E: Data Mining: Practical machine-learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[2] Hripcsak G, Rothschild AS: Agreement, the F-Measure, and Reliability in Information Retrieval. J Am Med Inform Assoc. 2005; 12(3): 296-298.

[3] Joachims T: Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, edited by C. Nédellec and C. Rouveirol, 1398. 1998; 137–142.

[4] Gonçalves T, Quaresma P: The impact of nlp techniques in the multilabel text classification problem. Intelligent Information Systems. 2004; 424–428.

[5] Wilcox A, Hripcsak G: Classification algorithms applied to narrative reports. Proc AMIA Symp. 1999; 455-459.