Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Biomedical Informatics 38 (2005) 145–153

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier

Serguei V. Pakhomov*, James Buntrock, Christopher G. Chute

Division of Biomedical Informatics, Mayo Clinic College of Medicine, SW, Rochester, MN 55905, USA

Received 28 September 2004
Available online 24 December 2004

Abstract

This paper addresses a very specific problem of identifying patients diagnosed with a specific condition for potential recruitment in a clinical trial or an epidemiological study. We present a simple machine learning method for identifying patients diagnosed with congestive heart failure and other related conditions by automatically classifying clinical notes dictated at Mayo Clinic. This method relies on an automatic classifier trained on comparable amounts of positive and negative samples of clinical notes previously categorized by human experts. The documents are represented as feature vectors, where features are a mix of demographic information as well as single words and concept mappings to MeSH and HICDA classification systems. We compare two simple and efficient classification algorithms (Naïve Bayes and Perceptron) and a baseline term spotting method with respect to their accuracy and recall on positive samples. Depending on the test set, we find that Naïve Bayes yields better recall on positive samples (95 vs. 86%) but worse accuracy than Perceptron (57 vs. 65%). Both algorithms perform better than the baseline with recall on positive samples of 71% and accuracy of 54%.

Keywords: Automatic classification; Naïve Bayes; Perceptron; Machine learning; Congestive heart failure; Natural language processing; Medical informatics

1. Introduction

Epidemiological research frequently requires recruiting a set of human subjects that are deemed relevant for a particular study. Clinical trials constitute another area where human subject recruitment is necessary. The recruitment is a tedious and difficult process and still remains a bottleneck for clinical research [1]. In this paper, we focus on an epidemiological study where patients with acute congestive heart failure need to be identified, preferably as soon as they are diagnosed in the clinic, so that they may be recruited to participate in the study. One of the requirements for an epidemiological study of this kind Do we know what kind? (Maybe we can determine the kind and combine this sentence and the next) is the completeness of the subject pool. Incidence or prevalence studies rely on complete population cohort identification. The identification of the candidates relies on a large number of sources, some of which do not exist in an electronic format, but it may start with the clinical notes dictated by the treating physician.

Another aspect of candidate identification is prospective patient recruitment. Prospective recruitment is based on inclusion or exclusion criteria and is of great interest to physicians for enabling just-in-time treatment, clinical trial enrollment, or research study options for patients. At Mayo Clinic, most clinical documents are transcribed within 24 h of patient consultation, which creates an ideal resource for enabling prospective

* Corresponding author. Fax: +1 507 284 0360.
E-mail addresses: pakhomov@mayo.edu (S.V. Pakhomov), buntrock@mayo.edu (J. Buntrock), chute@mayo.edu (C.G. Chute).

1532-0464/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2004.11.016

recruitment based on criteria present in clinical documents.

Probably the most basic approach to identification of candidates for recruitment is to develop a set of terms the presence of which in the note may be indicative of the diagnoses of interest. This term set may be used as a filtering mechanism, either by searching an indexed collection of clinical notes or simply by doing term spotting if the size of the collection allows it. For example, in case of congestive heart failure, one could define the following set of search terms: "CHF," "heart failure" "cardiomyopathy" "volume overload" "fluid overload," and "pulmonary edema." The number of possible variants is virtually unlimited, which is the inherent problem with this approach. It would be hard to guarantee the completeness of this set to begin with, and the problem is further complicated by morphological and spelling variants. This problem is serious affecting recall and therefore completeness of the candidate pool (already established point).

Another problem is that such term spotting or indexing approach would have to be intelligent enough to identify the search terms in negated and other contexts that would render documents containing these terms irrelevant. A note containing "no evidence of heart failure" should not be retrieved, for example. Identifying negation and its scope reliably is far from trivial and is in fact a notoriously difficult problem in linguistics [2]. This problem is slightly less serious than the completeness problem since it only affects precision, which is less important in the given context than recall, but high precision is still very desirable because it would minimize the amount of manual review of false positives needed.

To be able to identify automatically whether a given patient note contains evidence that the patient is relevant to a congestive heart failure study, a computer system has to "understand" the note. Currently, there are no systems capable of human-like "understanding" of natural language; however, there are methods that allow at least partial solutions to the language understanding problem once the problem is constrained in very specific ways. One such constraint is to treat language understanding as a classification problem and to use available machine learning approaches to automatic classification to solve the problem. Clearly, this is a limited view of language understanding, but we hypothesize that it is sufficient for the purposes referred to in this paper.

2. Previous work

The classification problems that have been investigated in the past are just as varied as the machine learning algorithms that have been used to solve these problems. Linear least squares fit [3], support vector machines, decision trees, Bayesian learning [4,5], symbolic rule induction [6], maximum entropy [7], and expert networks [8] are just a few of the algorithms that have been applied to classifying e-mail, Web pages, news articles, and medical reports, among other documents.

Aronsky and Haug [5] have developed and tested a system based on Bayesian learning for identification of patients with pneumonia for a suggested clinical guideline. The evaluation of the system showed 68.5% specificity at 95% sensitivity which, according to the authors, was acceptable for a real-time diagnostic system that has more stringent requirements for recall than for precision. The requirements for our application are similar. A difference, however, is that we are not attempting to diagnose patients for congestive heart failure automatically; rather, we are using the diagnoses and observations already issued by physicians to identify candidates that meet specific selection criteria. Our application also relies on natural language processing for analyzing the text of clinical notes to extract salient features predictive of patients with ongoing CHF.

A number of successful approaches to similar problems have been taken in the past using natural language processing (NLP). Wilcox [9] have experimented with a number of classification algorithms for identifying clinical conditions, such as congestive heart failure and chronic obstructive pulmonary disease, in radiology reports. They found that using an NLP system such as Medical Language Extraction and Encoding System (MedLEE) and domain knowledge sources such as UMLS [10] for feature extraction can significantly improve classification accuracy over the baseline where single words are used to represent training samples.

Jain and Friedman [11] have also demonstrated the feasibility of using MedLEE for classifying mammogram reports. Unlike Wilcox [9], this work does not use an automatic classifier. Instead, it uses the MedLEE NLP system to identify findings that are considered suspicious for breast cancer directly to profile potential candidates. In both cases, automatic classification and profiling, NLP plays an important role, whether it is used for feature extraction or for term spotting.

Aronow et al. [12] have also investigated a problem with one particular aspect that is similar to the one described in the present work. This aspect is the acuity of the condition being identified. The authors developed an ad-hoc classifier based on a variation of relevance feedback technique for mammogram reports, where the reports were classified into three "bins": relevant, irrelevant, and unsure. One of the features of the text processing system they used had to do with the ability to detect and take into account negated elements of the reports. Another system developed by Aronow et al. [13] is designed for classifying electronic patient encounter notes for the purpose of identifying patients with acute cases of pediatric asthma exacerbation. The system uses an inference network information retrieval