Identifying population characteristics tables in full text articles

Garrick Sherman, MS, Catherine Blake, PhD, Jooho Lee, MS
University of Illinois, Urbana Champaign, IL, USA

Abstract: *Authors frequently use tables to provide population characteristics of patients who participate in a study. Automated indexing methods tend to focus on the narrative in an article, rather than tables, even though this data can help physicians determine whether a study is relevant to their current clinical encounters. We provide an automated method to identify tables that contain population characteristics, which can be extracted in a later step.*

Introduction

Several efforts have focused on standardizing the data reported in a clinical trial (CONSORT), a meta-analysis of observational studies (MOOSE), and an epidemiological study (STROBE). These new guidelines do not address how to extract information from the millions of already published studies. Our goal is to identify and extract demographics (e.g. age, gender, and ethnicity), behavioral factors (e.g. tobacco and alcohol consumption), and medical descriptors (e.g. disease stage) from tables, such as that shown in Figure 1. As a first step, we identify tables containing data corresponding to the patient aspect in PICO [1] and population, intervention or risk factor, and medical condition of the information synthesis framework [2].

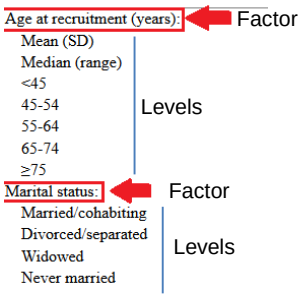

Figure 1. Factors and levels in population table.

Materials and Methods

Full text articles in 18 breast cancer journals from the open access subset of PubMed Central were downloaded. Tables were extracted from the 3,638 NXML files; only the first table is included in these experiments. A heuristic approach to processing tables, similar to [3], was used. Headers were identified using HTML tags, and factors were defined as the first non-empty cell in a row that included either a) only one non-empty cell, or b) at most half as many non-empty cells as total cells in the row (including empty cells). Once factors were identified, the first cell of the next row was interpreted as a level and all subsequent cells in that row were considered values. Differently structured tables were labeled with a generic "Main" factor, but otherwise parsed identically.

The task of identifying population characteristic tables was framed as a classification problem. Several features were considered: factor unigrams, factor *and* level unigrams, caption unigrams, and numbers. Only unigrams were considered due to the brevity of table text. Text was tokenized on whitespace and converted to lower case (no stemming was applied since terms were not expected to exhibit many forms). Stopwords were eliminated, as were terms occurring in ≤3 or >95% of the tables. A training set of 1,001 tables was annotated by hand (414 population tables), and information gain (see [4]) was used to identify the most discriminating terms. Four classifiers—support vector machine (SVM), naïve Bayes (NB), decision tree (DT), and the general linear model (GLM)—were used to classify the test set in the Oracle Data Miner (with default settings). The features identified in the training set were then used on an unseen test set comprising 497 tables.

Results

NB outperformed the other classifiers for all feature combinations, with a maximum accuracy of 87.6% on the judged subset, which included 50 random positive and 50 random negative examples. The most effective set of features were the top 50 terms, no captions or numbers and two columns (i.e. using terms from both the factor and levels). The NB model was applied to the test set. There were 29 false positives and 32 false negatives leading to an overall accuracy of 87.73% in the test set.

Term	Cap-tion	Num-ber	Cols	SVM	DT	NB	GLM
50	No	No	2	84.46	**84.97**	**87.56***	83.42
50	Yes	No	2	84.46	84.46	86.01	**84.97**
100	Yes	No	1	**85.01**	82.69	85.53	84.75
100	Yes	Yes	2	82.64	84.46	86.01	83.94
100	Yes	No	2	81.35	84.46	86.01	**84.97**
50	Yes	No	1	81.40	82.69	85.27	83.98
50	No	No	1	71.49	70.61	76.75	71.49

Table 1. Accuracy of selected feature subsets. Bold indicates highest per classifier; asterisk highest overall.

References

1. Richardson, W.S., et al., *The well-built clinical question: a key to evidence-based decisions.* ACP J Club, 1995. **123**(3): p. A12-3.
2. Blake, C. and W. Pratt, *Collaborative Information Synthesis I: A Model of Information Behaviors of Scientists in Medicine and Public Health.* Journal of the American Society for Information Science, 2006. **57**(13): p. 1740-9.
3. Pyreddy, P. and W.B. Croft, *TINTIN: A system for retrieval in text tables.* Acm Digital Libraries '97, 1997: p. 193-200.
4. Yang, Y. and J.P. Pedersen. *A Comparative Study on Feature Selection in Text Categorization*. in *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*. 1997.