Top-level MeSH Disease Terms Are Not Linearly Separable in Clinical Trial Abstracts

Joël Kuiper[1] and Gert van Valkenhoef[1,2]

[1] Faculty of Economics and Business, University of Groningen, Groningen, The Netherlands
[2] Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

Abstract. Assessments of the efficacy and safety of medical interventions are based on systematic reviews of clinical trials. Systematic reviewing requires the screening of vast amounts of publications, which is currently done by hand. To reduce the number of publications that are screened manually, we propose the automated classification of publications by disease category using Support Vector Machines. We base our classification on the ontological structure of the Medical Subject Headings (MeSH) by treating all terms as their top-level disease category. Unfortunately the resulting classifier lacks sufficient sensitivity for use by systematic reviewers. We argue that this is partially due to the inseparability of the terminology into the disease categories and discuss how future work could address this problem.

1 Introduction

Randomized Controlled Trials (RCTs) provide the most reliable assessments of the efficacy and safety of medical interventions and as such they should inform treatment decisions [1]. This information is complicated by the massive numbers of trials that are conducted; for example, the Cochrane Library indexes 286,418 trials over the last decade [2]. To reduce the amount of information they need to process, decision makers often rely on systematic reviews to summarize the RCTs that concern a specific disease and/or intervention. Systematic reviewing consists of searching the literature, screening the results and summarizing the relevant evidence. In screening, the inclusion of trials depends on what they study, i.e. the disease, interventions and patient population. In some cases this means reducing thousands of publications to only a handful. There have been attempts to optimize search queries to reduce the number of false positive results [3], but for a comprehensive review a broader query is required [4]. Therefore systematic reviewers maximize sensitivity and sacrifice specificity.

Accurate meta-data could reduce the number of false positives while maintaining a high sensitivity. For example, PubMed employs the Medical Subject Headings (MeSH) to provide meta-data such as anatomical terms, organisms, and diseases. However, like most meta-data initiatives it relies on the goodwill of the authors and publishers to provide annotations. Consequently, only a fraction

of the publications are completely and correctly annotated. To fill these gaps in meta-data, machine learning techniques like automatic classification can be applied. Automatic classification is a type of supervised learning where a mapping of observations to predetermined categories is built, based on a "training" set of observations for which the class is known in advance. This methodology is potentially both useful and appliccable for PubMed MeSH annotations: filtering by disease is an important sub-task of screening and PubMed provides a large subset of abstracts annotated with MeSH disease terms that can be used as a training set. In this paper, we explore the application of automatic classification to the problem of tagging abstracts with the appropriate MeSH disease terms.

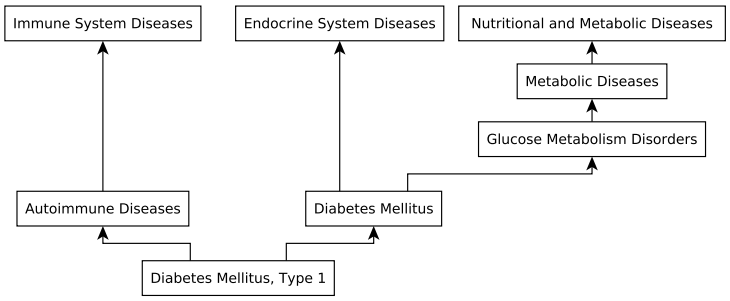

Fig. 1. *Diabetes Mellitus, Type 1* sub-graph of the MeSH disease terms

The MeSH disease terms are structured as an ontology, meaning that terms are encoded as nodes in a directed acyclic graph where its connections represent relations such as instance-of (`is-a`). For example, Diabetes Mellitus `is-a` Glucose Metabolism Disorder, which in turn `is-a` Metabolic Disease (see Fig. 1). MeSH (2013 edition) places 4308 unique disease terms in only 26 top level categories. However, since MeSH is not a strict hierarchy, most terms correspond to several top level categories; in other words the top-level categories are not mutually exclusive. *Diabetes Mellitus, Type 1*, for example, maps to the top level categories *Immune System Diseases*, *Endocrine System Diseases*, and *Nutritional and Metabolic Diseases* (see Fig. 1). This raises the question which level of description is most appropriate for applying machine learning, and how the ontological structure can be used to both simplify the classification task aid the researcher in screening by providing familiar categorizations.

To start exploring the possibility of using the MeSH disease categories for automated classifications of RCTs, we attempt to automatically classify publications by their corresponding top level MeSH categories using Support Vector Machines (SVMs). A resulting classifier with sufficient performance could reduce the number of false positives from literature searches and provide a starting point for more fine-grained classification of terms deeper in the ontology.

2 Methods

To obtain a representative corpus, PubMed was queried for all publications tagged with the publication type "Randomized Controlled Trial" and the MeSH term "Human". The publications' MeSH annotations were then matched against the MeSH disease terms based on exact string match. The ontological structure of MeSH was taken into account by treating all disease terms as their top level ancestor(s). This was done by first transforming the MeSH terms into the Open Biological and Biomedical Ontologies (OBO) format [5] and then extracting relationships from the resulting ontology using the OntoCAT library [6]. The disease categories *Animal Diseases*, *Disorders of Environmental Origin*, *Pathological conditions, Sings and Symptoms* and *Occupational Diseases* were excluded from the ontology because they were either irrelevant, only included very few terms, or had substantial overlap with other categories.

The title and abstract of each publication were used as input for text classification. To reduce dimensionality common English words were filtered from these texts and the words were stemmed using a kstem filter [7]. Each publication was treated as a bag-of-words, a representation in which the order of the words in the document is ignored. For each of the words the Term Frequency normalized by the Inverse Document Frequency (*tf-idf* [8]) was calculated. The term frequency is the number of occurrences of a word in a document, the Inverse Document Frequency is the logarithm of the total amount of documents divided by the total occurrences of a word across all documents:

In general, classification algorithms attempt to find a mapping between labels y_i and instances $\mathbf{x}_i$ where $\mathbf{x}_i \in R^n$ and $y_i \in \{-1, +1\}$. Here SVMs were used to classify the annotated publications, because they have been successful in fast large scale text classification [9,10]. SVMs attempt to perform classification by finding a maximum margin-separating hyperplane. This is done by solving the following unconstrained optimization problem:

$$\min_w \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi(\mathbf{w};\mathbf{x}_i,y_i) \ , \tag{1}$$

with a loss function ξ and $C > 0$ as a penalty parameter. The LIBLINEAR [11] SVM implementation was used to solve the optimization in its primal form with the L2 loss function:

$$\xi(\mathbf{w};\mathbf{x}_i,y_i) = (\max(1 - y_i\mathbf{w}^T\mathbf{x}_i, 0))^2 \tag{2}$$

The penalty factor C was varied between 0.0001 and 10 to determine the effects of introducing a softer margin. For each of the categories a binary one-vs-the-rest classifier was constructed, because publications could belong to multiple categories.

Performance of the classifiers was assessed using 10-fold cross-validation with averaged specificity and sensitivity as performance measures.

3 Results

The PubMed query resulted in 404,371 publications of which the full records were retrieved in XML format (on 2012-10-29). Only 13,918 publications were annotated as a top-level disease category. When treating all terms as their top-level ancestor(s) 226,710 publications were annotated with a disease.

The specificity of the binary classifiers for each top-level disease category was consistent around 0.98 for all values of the penalty parameter C. The median sensitivity of the classifiers was 0.53 ($C = 1.0$). The classifier for *Stomatognathic Diseases* was the most sensitive (0.822, $C = 1.0$). The least sensitive classifiers were those for *Hemic and, Lymphatic Diseases* ($0.208, C = 1.0$) and *Congenital, Hereditary, and Neonatal Diseases and Abnormalities* ($0.270, C = 1.0$). Relaxing the condition for the optimal hyper-plane by decreasing the penalty factor C (i.e. introducing a soft margin) did not substantially improve the sensitivity. From this we conclude that the data on the margin was not particularly noisy, and it indicates that the data might be linearly inseparable.

4 Discussion

Our classifiers had poor sensitivity (median of 0.53, whereas acceptable sensitivity would be 0.8 or higher). One explanation could be that some disease terms lack specific terminology which does not overlap with the other categories, so that the SVM fails to find a separating hyperplane. It could be that in general for medical terminology the ontological descendants of a top-level item do not sufficiently generalize that top-level item, i.e. the terminology of Diabetes and Hyperglycemia do not cluster under Metabolic Diseases. Indeed, it seems that the sensitivity of a classifier correlates with how well-defined a term is. For example, the classifier for the well-defined term *Stomatognathic Diseases* is sensitive (0.822) whereas the MeSH even suggests not to use the term *Hemic and Lymphatic Diseases* because it is too general. This raises the question for which level of description the classification problem would be easiest to solve.

To assess the separability of the data, clustering techniques could be applied. This would require the dimensionality of the data to be drastically reduced through techniques such as Latent Semantic Analysis [12] or Principal Component Analysis. Subsequent manual analysis could reveal associations between the identified clusters and MeSH disease terms. This could aid with a more classical classification task or be used as an alternative way of filtering RCTs.

However, an open question remains whether using ontologies derived from human expertise to guide supervised classification algorithms is feasible way altogether. It could very well be that while ontologies provide familiar and accepted labels for the (human) end-user, far better categorization can be achieved without a fixed hierarchy, instead leaving the categorization up-to the algorithm at hand.

Topic modelling techniques [13] such as Probabilistic Latent Semantic Analysis or Latent Dirichlet allocation, could provide techniques for finding relevant publications without the aid of a established ontology.

References

1. Evidence-Based Medicine Working Group: Evidence-based medicine. A new approach to teaching the practice of medicine. Journal of the American Medical Association **268**(17) (1992) 2420–2425
2. van Valkenhoef, G., Tervonen, T., de Brock, B., Hillege, H.: Deficiencies in the transfer and availability of clinical evidence in drug development and regulation. BMC Medical Informatics and Decision Making (2012) (in press).
3. Haynes, R.B., McKibbon, K.A., Wilczynski, N.L., Walter, S.D., Werre, S.R.: Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ **330**(7501) (2005) 1179
4. Higgins, J., Green, S., eds.: Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]. The Cochrane Collaboration (2009) Available from http://www.cochrane-handbook.org.
5. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology **25**(11) (2007) 1251–1255
6. Adamusiak, T., Burdett, T., Kurbatova, N., Velde, K.J.v.d., Abeygunawardena, N., Antonakaki, D., Kapushesky, M., Parkinson, H., Swertz, M.A.: OntoCAT – simple ontology search and integration in java, r and REST/JavaScript. BMC Bioinformatics **12**(1) (May 2011) 218
7. Krovetz, R.: Viewing morphology as an inference process. In: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '93, New York, NY, USA, ACM (1993) 191202
8. Salton, G., Fox, E.A., Wu, H.: Extended boolean information retrieval. Commun. ACM **26**(11) (November 1983) 10221036
9. Joachims, T.: Text categorization with suport vector machines: Learning with many relevant features. In: Proceedings of the 10th European Conference on Machine Learning. ECML '98, London, UK, UK, Springer-Verlag (1998) 137142
10. Joachims, T.: Training linear SVMs in linear time. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. (2006) 217226
11. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874
12. Dumais, S.T., Furnas, G.W., Landauer, T.K., Deerwester, S., Harshman, R.: Using latent semantic analysis to improve access to textual information. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '88, New York, NY, USA, ACM (1988) 281285
13. Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4) (2012) 7784