A Probabilistic Framework for Ontology-Based Annotation in Neuroimaging Literature

Chayan Chakrabarti [1*], Thomas B. Jones [1*], Jiawei F. Xu [1], George F. Luger [1], Angie R. Laird [2], Matthew D. Turner [1, 3], and Jessica A. Turner [3]

[1] Department of Computer Science, University of New Mexico, Albuquerque, NM. [2] Department of Physics, Florida International University, Miami, FL. [3] Mind Research Network, Albuquerque, NM.

ABSTRACT

Ontologies encode domain relationships in formally robust data structures that can be used to annotate scientific literature regarding human neuroimaging studies and results. The annotation process requires significant time and effort when performed by humans. Text mining algorithms can facilitate this process, but they render an analysis mainly based upon keyword, synonym and semantic matching. They do not leverage information embedded in an ontology's structure. We present a probabilistic framework that facilitates the automatic annotation of literature by leveraging the dependencies between labels in different categories in the ontology. Our research focuses on annotating human functional neuroimaging literature using the Cognitive Paradigm Ontology. We use a hierarchical approach that combines the stochastic simplicity of naïve Bayes with the formal transparency of decision trees. Our data structure is easily modifiable to reflect changing domain knowledge. We compare our results to the baseline analysis returned by the National Center for Biomedical Ontology (NCBO) annotator.

1 INTRODUCTION AND MOTIVATION

Using ontologies to annotate scientific literature is an important knowledge management task in many scientific fields. Automating this process leads to reduction in human time and effort. We present an approach that goes beyond text mining, and demonstrate results on a specific ontology.

1.1 BrainMap and CogPO

BrainMap (www.brainmap.org) is one of the largest databases of human neuroimaging results. The BrainMap software suite provides toolsets needed to explore the different cognitive constructs underlying brain function in various disorders. In order to run large-scale meta-analyses a method to easily identify studies using the same (or similar) experimental methods and subjects is necessary.

Although the value of the BrainMap project has been proven through facilitating numerous meta-analyses of fMRI studies, the number of publications in the literature far outweighs the number of publications that have been included in the database. The task of annotating fMRI papers with experimental terms is traditionally performed with manual curation by human experts (Laird et al. 2005). This currently creates a bottleneck. In this research we propose to assist or even replace the human with automated suggestions for classifying the PubMed paper.

The BrainMap method for describing experiments has evolved into a taxonomy composed chiefly of structured keywords that categorize the experimental question addressed, the neuroimaging methods used, the behavioral conditions during imaging, and the statistical comparisons performed.. BrainMap forms the backbone of the Cognitive Paradigm Ontology (CogPO), which is incorporated as an ontology in the National Center for Biomedical Ontologies (NCBO) repository (Turner & Laird 2012). The standardized annotations allow published experiments to be linked and identified for meta-analyses, despite the use of alternate vocabularies.

1.2 Text Mining for annotation

Text mining algorithms for annotation are needed for the problem of multi-label classification; the general case in which there are more than two labels to choose from, and each instance can have more than one label (Read et al. 2011, Shi et. Al 2012).

In previous work using a similar dataset, we evaluated a version of k-nearest-neighbor (kNN) in performing automated annotations (Chakrabarti et al. 2012). We found that the performance was comparable with results on other textual annotation datasets, but fairly poor for the multi-label aspects of the problem. Naïve Bayes is a probabilistic learning method, based on Bayes' rule, which works surprisingly well on problems where its strong independence assumption is not met. Particularly it works well for supervised learning when the number of the instances in the training set is relatively small, which is the situation here. It has been extended to the multi-label scenario using various transformation techniques (Dembczynski et al 2010). Therefore, we start with a naïve Bayes approach.

The naïve Bayes technique across all categories and possible labels does not leverage the correlations between labels in different categories, which are implicitly encoded in

[*] Joint first authors, to whom correspondence should be addressed.

the domain ontology. Most text-mining techniques consider the labels to be anchors for clustering or topic modeling techniques, but have no way to use the fact that the terms may have implicit correlations to each other and object features. The features used to derive terms in traditional text mining are most often a set of high entropy keywords (Srivastava and Sahami 2009). We present a framework for the annotation task that makes use of the implicit information that is encoded in the ontology.

In many ontologies, there are often different categories from which a label may be drawn (Turner & Laird 2012). While naive Bayes is able to assign certain features in a training object to labels in a single category, it is unable to learn about correlations between labels and their associated features in different categories. Further, it is not possible for naive Bayes alone to increase or decrease its confidence in one label after it has been informed that some other label is a correct or incorrect annotation for the same object. Our method expands on naive Bayes by constricting training sets at each node in the tree to only those training objects pertinent to that node. This allows us to take advantage of any underlying correlations in the training set between labels of different categories, which would otherwise be hidden by building a separate classifier for each category.

2 DATA SETS AND FORMAL FRAMEWORK

Our data set consists of annotated abstracts from PubMed. We work on the abstracts and not the full papers because we want to interface our tool directly with the eUtils toolkit of PubMed that can fetch a set of text abstracts in batch mode.

2.1 Multi-label text corpus

Our corpus consists of 247 expertly annotated abstracts from fMRI and PET human neuroimaging papers, which are part of the BrainMap database. We consider annotations in 5 distinct categories for each abstract – Stimulus Modality (SM), Stimulus Type (ST), Response Modality (RM), Response Type (RT) and Instructions (I). Each of these categories comprises from 7 to over 40 allowable labels as described in CogPO (Turner & Laird 2012). Each abstract has been annotated by a human expert, and the label set for a single abstract includes at least one label from each of the SM, ST, RM, RT, or I categories, and possibly multiple labels from each.

The average number of labels per category per abstract ranged from 1.15 to 1.85 depending on the category. The CogPO ontology explicitly includes constraints on the labels, e.g. a “Tone” as a Stimulus Type label entails that the Stimulus Modality must include “Auditory”, or the Instruction label “Smile” entails a Response Modality “Facial” label. The reader may notice that a flat text mining approach would be unable to make these distinctions, i.e., it would not be able to tell that label a changes the probability of label b in some other category. This multi-label, multi-class, multi-category corpus serves as a gold standard against which we test our stochastic approach to literature annotation.

2.2 Naïve Bayes

More formally, we define the set of abstracts, the feature vector (representing words from the corpus that are stems and not stop words), and the set of feature vectors as follows:

$$D = \{d_i | d_i \text{ is an abstract in the corpus}\}$$

$$F = < f_x | f_x \text{ is a feature representing a word} >$$

$$V = \left\{ v_c | v_c = < b_{c1} \dots b_{cn} >, b_{cj} = \begin{cases} 1, & c \in a_i \\ 0, & otherwise \end{cases} \right\}$$

By definition, the length or size (| |) of

$$|v_c| = |f| \text{ and } |V| = |D| = \text{number of abstracts}$$

We can define CogPO as a set of categories

$$C = \{SM, ST, RM, RT, I\}$$

and each category using their individual labels.

$SM = \{l_1, l_2, \dots\}$ (other categories can be defined similarly)

From the domain knowledge we know the following values

$$P(M(d_x, l_1)|b_{xi}), P(b_{xi}|M(d_x, l_1)), P(b_{xi})$$

where $M(d_x,l_1)$ is an indicator function that is 1 if d_x is labeled with l_1, and 0 otherwise.

We want to find the following value

$$P(M(d_i|l_1), V_i) = \frac{P(M(d_i|l_1) \cup V_i)}{P(V_i)}$$

It follows from Bayes' rule that

$$P(M(d_x, l_1)| V_x) \propto P\big(M(d_x, l_1)\big) \prod_{i=1}^{|F|} P(b_{xi} | M(d_x | l_1))$$

Similarly, we can calculate the probability for all the other labels in SM as well as ST, RM, RT, and I. We used binary relevance in a single category to solve the multi label classification problem (Read et al. 2011). Our method takes the raw probability calculated by the Bayesian classifier using the above equations for each label, and accepts all labels, which receive a probability greater than a pre-defined cutoff α.

2.3 Bayesian Decision Trees

Decision trees are discrete models that can predict the output labels of samples in a data set, based on several input variables arranged in a tree-like structure with nodes and branches. Nodes in the tree represent a decision variable and the branches correspond to the next decision variable to be queried based on the outcome of the previous decision variable. We use the Bayesian classifiers to make decisions about which labels to include while traversing down the tree.

Definition 1. $B_{C,S}$ *is a Bayesian classifier trained on set* $S \subseteq D$ *over category* C .

Definition 2. *If* S *is a training set and* $s \in S$ *then* label(s) *is the set of correct labels attached to item* s.

Definition 3. *If* t *is a node in a tree* T *such that each node in* T *contains a label or an empty label, then* l_{t*} *is a set that contains the label of node* t *and all of the labels of each ancestor of* t, *with no addition made if the label of a node is empty. In practice, the root is the only node that will have an empty label, since the training set is unlimited on the root node.*

Definition 4. T *is a* Bayesian Decision Tree *if each node* t *of* T *consists of a category* C_t *which is not the same category as any of the ancestors of* t, and which is shared among the siblings and cousins of t; *a label* l_t *which comes from the category of the parent of* t *and which is not shared with any of the siblings of* t; *and a multi-label Bayesian classifier* $B_{Ct,St}$ *using definition 1. The training set* S_t *has the following restriction:* $\forall s \in S_t$ label(s) $\subseteq l_{t*}$. *Finally, we require that the label of the root node is empty.*

Definition 5. *If* B_t *is the Bayesian classifier associated with node* t *and* I *is an object which maybe categorized by* B_t, *then* $B_t(I)$ *is the list of all labels which* B_t *returns upon classifying* I.

Definition 6. *If* l *is a label and* t *is a node in a tree then* Child(l, t) *is the child of* t *which contains label* l.

Using these definitions, we construct a formal framework for annotating the CogPO human neuroimaging abstracts with labels from the CogPO ontology categories of SM, ST, RM, RT, and I.. We limit the training set on the naïve Bayes classifiers, in the tree, in order to leverage the dependencies between labels in different categories. By limiting the training set on an internal node to only those abstracts which have the labels of that node's parents, we change the underlying probabilities of the Bayesian classifier to better fit any dependencies between labels in different categories. This "less is more" approach helps the Bayesian classifier to focus on features, which are important to the current node in the tree.

3 METHODOLOGY AND ALGORITHMS

The NCBO Annotator, a component of the Bio-Portal, (http://bioportal.bioontology.org/annotator) is a Web service that annotates textual metadata with relevant ontology concepts (Shah et al 2009). The NCBO Annotator consists of several standard annotations draws from many ontologies for its annotations, including CogPO. We use this to annotate our corpus of abstracts as a baseline. The performance is measured using the F1-micro score, based on precision and recall. In all our calculations, we set $\beta = 1$. (results are presented in Table 1).

$$F_\beta = (1+\beta^2)\frac{precision * recall}{\beta^2 * precision + recall}$$

We first construct 5 separate naïve Bayes classifiers for each of the 5 categories as formalized in section 2.2. Each classifier is trained and tested on the entire corpus of abstracts using 10-fold cross-validation, and their f1-micro scores are calculated. Abstracts in the testing set are annotated with a label if the label had a probability score greater than $\alpha = 0.1$.

Next we construct the Bayesian Decision Trees (BDTs) as formalized in section 2.3.. Given that we have 5 categories, we build all 120 possible BDTs. We annotate the corpus of abstracts using the BDTs, with the criterion that if the probability of a label is greater than 0.1 for some abstract, then that abstract is tagged with that label. Then we aggregate the labels across each of the 5 categories and calculate a mean f-score for each category to determine the quality of the annotations for each instance of the category across the 120 trees.

Finally, we consider the case when a human annotator is using our algorithm to quickly annotate abstracts. In practice an expert sometimes can accurately guess the label for at least one of the categories just with a quick glance at an abstract (e.g., the abstract states explicitly that the experiment used picture of faces as the stimulus, or that subjects pushed a button with their foot to respond). To model this, we trained our BDTs with the condition that the root node has already been decided. We call this the Constrained Decision Tree (CDT). As a result we have trees rooted at SM, ST, RM, RT, and I, corresponding to the cases where the human expert assigns the label for that category. The rest of the tree is constructed exactly as before except that, when the mean f-score is calculated for each category across all possible CDTs, we remove the instances where the f1-micro score is 1.0. These instances correspond to the annotations assigned by the expert, and we do not want them to influence the quality of the results returned by our algorithm.

Input*: Un-Labeled Item I, Bayesian Decision Tree T*
Output*: Label Vector in Multiple Categories L*
t = Root(T)
SearchList = NULL
while do *t/= NULL*
    L = L : Bt(I)
    for *l* $\in$ *$B_t(I)$* **do**
        SearchList = SearchList : Child(l, t)
    end for
    t = SearchList[0]
    x : SearchList = SearchList
end while
return L

Algorithm 1: Generalization of Bayesian Decision Trees

	SM	*ST*	*RM*	*RT*	*I*
F-score	0.0000	0.0094	0.0000	0.0023	0.0043

Table 1: F1-micro results of the NCBO Annotator recreating the correct annotations for the fMRI abstracts.

4 RESULTS AND DISCUSSION

We observe from Table 1 that annotations returned by the NCBO annotator leave a huge scope for improvement. Given the paucity of synonyms listed in CogPO, correct results occurred only when the annotation term occurred itself in the abstract text.

The quality of the results for the naïve Bayes (NB), Bayesian Decision Trees (BDT), and Conditional Decision Trees (CDT) are shown in figure 1. The error bars shown are twice the standard deviation on both directions of the mean of the f1-micro score for each category. All the methods show a significant improvement over the NCBO results. The f1-micro scores for Stimulus Type (ST) and Instructions (I) are lower than in the other categories because of the large number of labels they incorporate, leading to lower

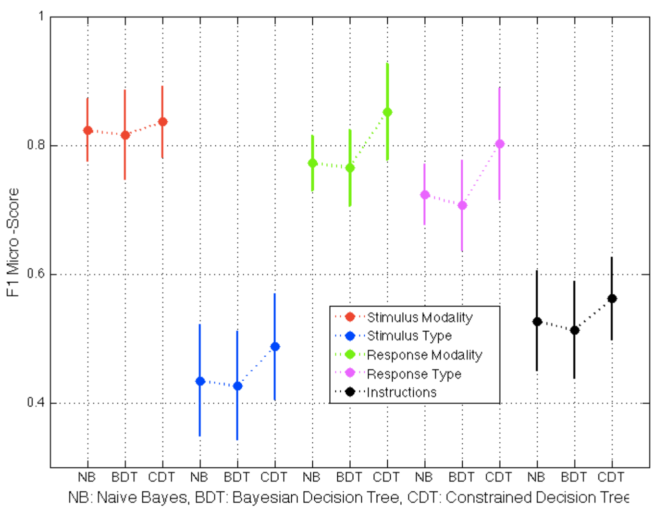

Figure 1: Comparison of Naïve Bayes, Bayesian Decision Tree, and Constrained Decision Tree, by CogPO category. The error bars are twice the standard deviation.

sample size for each label. Stimulus Modality (SM), Response Modality (RM), and Response Type (RT) have fewer labels and thus better performance. For every category, the Bayesian Decision Tree f1-micro score is slightly lower than that of the naïve Bayes. This is due to fact that our sample size constriction for the training sets at each level of the decision tree decreases precision and recall for labels lower down in the tree and any increases due to underlying correlations are not sufficient to make up for this decrease. The Constrained Decision Tree always has a higher f1-micro score than the other 2 methods because the guarantee of correct labels in the first category of each tree is leveraged through the cascading correlations among labels in different categories further down the tree and labels in the root node's category.

5 CONCLUSIONS AND FUTURE WORK

We have demonstrated a formally rigorous stochastic framework for annotating BrainMap literature using the Cognitive Paradigm Ontology. Unlike text mining algorithms, our framework references the knowledge encoded by the dependencies in the ontology. The framework can be easily modified and updated to reflect changes in domain knowledge.

Our results from naïve Bayes analysis significantly improve upon the baseline analysis of the NCBO annotator. The constrained decision tree architecture improves upon the naïve Bayes results. When we fix the first node of the decision tree, there is a significant improvement in the annotation accuracy. This is a useful tool for aiding a human expert in annotation because the expert can accurately select one annotation from several categories with a quick skim of an abstract. Our technique can then annotate the remaining categories with high accuracy. Although this approach does not eliminate the human expert from the loop, it complements their decision-making and significantly reduces the time and effort for the annotation task.

We next plan to apply our techniques to different ontologies with more complex structures. We believe the modular nature of our framework will scale well to these new ontologies. We also want to algorithmically learn gaps (missing labels) in the ontology through literature matching analysis.

REFERENCES

Chakrabarti C, Luger GF, Laird AR, and Turner JA (2012). Automated Annotation of Abstracts for Cognitive Experiments.Bio-Ontologies SIG, ISMB.

Dembczynski K, Waegeman W, Cheng W, and Eyke Hullermeier, (2010). Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss. European Conference on Machine learning and Knowledge Discovery in Databases.

Laird AR, Lancaster JL, Fox PT, (2005*) BrainMap: the social evolution of a human brain mapping database.* Neuroinformatics. 3(1):65-78.

Read, J, Pfahringer B, Holmes G, Frank E, (2011), Classifier Chains for Multi-label Classification. Machine Learning.

Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA, (2009). Ontology-driven Indexing of Public Datasets for Translational Bioinformatics. BMC Bioinformatics, Volume 10.

Shi C, Kong X, Philip SY and Wang B, (2012) Multi-Objective Multi-Label Classification. OMNI Books.

Srivastava, A., and Sahami. M, (2009). *Text Mining: Classification, Clustering, and Applications*. Boca Raton, FL: CRC Press.

Turner JA, Laird AR, (2012) *The cognitive paradigm ontology: design and application.* Neuroinformatics. 10(1):57-66.