Journal of Biomedical Informatics 72 (2017) 132–139

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Predicting biomedical metadata in CEDAR: A study of Gene Expression Omnibus (GEO)

Maryam Panahiazar, Michel Dumontier, Olivier Gevaert *

Stanford Center for Biomedical Informatics Research, Center for Data Annotation and Retrieval, Department of Medicine, Stanford University, Stanford, 94305, United States

ARTICLE INFO

Article history:
Received 26 January 2017
Revised 1 June 2017
Accepted 14 June 2017
Available online 16 June 2017

Keywords:
Data mining
Prediction
Metadata
GEO
CEDAR

ABSTRACT

A crucial and limiting factor in data reuse is the lack of accurate, structured, and complete descriptions of data, known as metadata. Towards improving the quantity and quality of metadata, we propose a novel metadata prediction framework to learn associations from existing metadata that can be used to predict metadata values. We evaluate our framework in the context of experimental metadata from the Gene Expression Omnibus (GEO). We applied four rule mining algorithms to the most common structured metadata elements (sample type, molecular type, platform, label type and organism) from over 1.3 million GEO records. We examined the quality of well supported rules from each algorithm and visualized the dependencies among metadata elements. Finally, we evaluated the performance of the algorithms in terms of accuracy, precision, recall, and F-measure. We found that PART is the best algorithm outperforming Apriori, Predictive Apriori, and Decision Table.

All algorithms perform significantly better in predicting class values than the majority vote classifier. We found that the performance of the algorithms is related to the dimensionality of the GEO elements. The average performance of all algorithm increases due of the decreasing of dimensionality of the unique values of these elements (2697 platforms, 537 organisms, 454 labels, 9 molecules, and 5 types). Our work suggests that experimental metadata such as present in GEO can be accurately predicted using rule mining algorithms. Our work has implications for both prospective and retrospective augmentation of metadata quality, which are geared towards making data easier to find and reuse.

1. Introduction

Biomedical data is increasingly being viewed as a valuable commodity that can be mined for new insights beyond that for which it was created. Large community-focused databases such as the Gene Expression Omnibus (GEO) [1] or the database of Genotypes and Phenotypes (dbGAP) [2] offer a wealth of omics' data that have been used in developing diagnostic, prognostic, and therapeutic models [3,4]. One crucial and limiting factor in the reuse of data lies in having access to accurate descriptions about the data – known as metadata. Community standards to describe an experiment (e.g. Minimum Information About a Microarray Experiment; MIAME [5]) are being widely promoted to highlight essential metadata, but creating good metadata can be challenging [6,7].

Indeed, metadata is often of low quality, and many entries are absent, erroneous or inconsistent. The largest database of gene expression studies, the GEO microarray database, contains 50,000 studies, over 1.3 million samples, and is still growing [1]. Yet the description of these samples suffers from a lack of consistency and completeness. For example, a preliminary analysis revealed that are 32 different ways to specify the age in GEO (e.g. age, Age, Age years, age year). Yet, these metadata are essential for researchers to find and reuse datasets of interest. When metadata are incomplete or inaccurate, researchers will miss relevant hits while being forced to sift through irrelevant results - resulting in lower productivity and potentially weaker scientific analyses. These issues are often attributed to lack of appropriate supporting infrastructure [8].

Metadata authoring applications such as ISA-Tools [9] or RightField [10] can be used to codify guidelines that specify multiple metadata elements and require users to use a set of controlled terms, such as terms from specified ontologies contained in the NCBO BioPortal [11]. Yet even with such tools, authoring good metadata is tedious and error-prone, and could benefit from more automation. The development of more effective platforms for metadata authoring and discovery is one of the goals of the Center for Expanded Data Annotation and Retrieval (CEDAR) [7,8].

In this study, we examine the utility of supervised machine learning to predict metadata from existing metadata. This will help

* Corresponding author.
E-mail address: olivier.gevaert@stanford.edu (O. Gevaert).

http://dx.doi.org/10.1016/j.jbi.2017.06.017
1532-0464/

metadata submitter during the submission process. Predicting metadata could be a guideline for template authors during the process of metadata definition. This facility will not only significantly facilitate the template definition task but also will make the resulting templates more comprehensive and reflective of the actual data. In CEDAR we also take advantage of emerging community-based standard templates for describing different kinds of biomedical datasets, and we investigate the use of computational techniques to help investigators to assemble templates and to fill in their values [7].

Learning value sets from data will help ensure that template authors do not miss important value sets that appear frequently in the data. Thus, data submitters will be able to find the terms they need, hence improving the quality of the metadata.

We use the increasing amounts of structured metadata to learn from as the project progresses and learn value sets conditional on the experimental level metadata. This incorporation of structural knowledge into the learning technology will allow us to infer common metadata patterns and their value sets in the context of technology platform, organism, molecule, label or sample type. Our key goal is to facilitate as much of the metadata collection process as possible, by suggesting possible value sets for the fields based on available data. This process will limit the value options, will reduce the burden of entering metadata terms and will significantly shorten the time that is needed for investigators to enter metadata.

We found that experimental metadata such as present in GEO can be accurately predicted using rule mining algorithms. Our work has implications for both prospective and retrospective augmentation of metadata quality, which are geared towards making data easier to find and reuse.

2. Background

Supervised learning uses classification algorithms to learn from data and make predictions. The goal of supervised learning is to build a model of the distribution of class labels from instances [12]. The classifier can then assign class labels to instances in which the values of the predictor features are known, but the value of the class label is unknown. Numerous supervised classification techniques have been developed including decision trees, artificial neural networks, and statistical techniques such as bayesian networks [12]. Machine learning has been widely applied across domains including the biomedical domain [13], such as protein function prediction [14], clinical outcome prediction [15] and survival analysis [16].

As we mentioned earlier, this study specifically is about metadata and association between them. Therefore, using machine learning will be helpful to mine the data, learn from the data, and find this association. In our study, we wanted to find the correlation between metadata elements and their values. Association rules are the main technique for data mining to find these correlations. Sharma et al., compared association rule mining algorithms (e.g. AIS and FP-Growth, and Apriori) [17]. Each algorithm has advantages and disadvantages according to their comparison. For example, AIS requires multiple scanning of the database, only rules that have one item in right side can be generated, and too many candidate itemsets are generated. FP-Growth also has some disadvantages such as the resulting FP-Tree is not unique for the same logical database and it cannot be used in interactive mining system. Apriori is scanning the complete database multiple times but still, it is easy to implement. Predictive Apriori algorithm overcomes this disadvantage of the Apriori algorithm with scanning the beast n rules instead of scanning all rules. PART algorithm uses partial decision trees to generate the decision list that is shown in the output, but only this final list is what is used to make classifications and with that, we have better performance.

In previously published manuscript [18], we proposed a framework to predict structured metadata terms from unstructured metadata for improving quality and quantity of metadata, using the Gene Expression Omnibus (GEO) microarray database. Our framework consists of classifiers trained using term frequency-inverse document frequency (TF-IDF) features and a second approach based on topics modeled using a Latent Dirichlet Allocation model (LDA) to reduce the dimensionality of the unstructured data. Our results based on GEO database showed that structured metadata can be predicted with TF-IDF more accurate than LDA. And both TF-IDF and LDA are outperforming the majority vote baseline as well. Overall this is a promising approach for metadata prediction that is likely to be applicable to other datasets and has implications for researchers interested in biomedical metadata curation and metadata prediction. Considering that metadata is structured and unstructured in GEO and other resources, we decided to find the correlation between structured metadata. In this study, we found the correlation between selected structured metadata elements versus in previous work we predicted structure metadata from the free text. Structure metadata has a potential to be predicted and suggested to metadata template author or metadata submitter during the submission process based on each other.

Several studies have been done regarding GEO metadata prediction. For instance Buckberry et al. [19] presented a method for predicting the sex of samples in gene expression microarray datasets. They believe that the metadata associated with many publicly available expression microarray datasets often lacks sample sex information, therefore limiting the reuse of these data in new analyses or larger meta-analyses where the effect of sex is to be considered. The package called massiR provides a method for researchers to predict the sex of samples in microarray datasets. “This package implements unsupervised clustering methods to classify samples into male and female groups, providing an efficient way to identify or confirm the sex of samples in mammalian microarray datasets” [19]. As it is clear this study is just about particular field in GEO data and it is specialized to predict the sex of the samples.

In this study, we propose methods to predict structured metadata. This method is applicable to any structured metadata in biomedical field. We use association rule mining (ARM) algorithms due to their interpretability and good performance [20]. ARM is a method for discovering relations between variables in large databases. [21]. ARM was defined by Agrawal in the early 90s in relation to a so called market basket analysis using APRIORI [20]. Since then, multiple studies have used this technique successfully to model data [22]. For example, ARM has been used to predict infection detection [23], to detect common risk factors in pediatric diseases [24], to understand the interaction between proteins [25], to discover frequent patterns in gene data [22], and to understand what drugs are co-prescribed with antacids [26]. To the best of our knowledge, ARM has not yet been applied for predicting experimental metadata.

3. Objective

We hypothesized that there are strong correlations between metadata elements and their values that can be used to predict metadata. The goal of this study is to predict the metadata based on the correlation between them. For example, there is a correlation between platforms, organism, and type. For GPL570 as a platform and *Homo Sapiens* as an organism a possible type of the study is RNA. We used four algorithms: Apriori, Predictive Apriori, Decision Table and PART (see below). We used these algorithms to find the association between metadata elements and to predict the value of each element of interest. We then evaluated our approach using a standard cross-validation of experimental metadata from GEO, a primary repository of gene expression data.