Biomedical Metadata Prediction

Maryam Panahiazar, Michel Dumontier, Olivier Gevaert

Objective

A crucial and limiting factor in data reuse is the lack of accurate, structured, and complete descriptions of data about the 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 and can be used to predict unknown metadata values.

Data

We used data from GEO , a database of gene expression data which contains experimental metadata authored by the original data submitters.
We used 5 structural elements: platform, type, organism, molecule, and label.

Element	Description
Platform	A platform is a list of probes that define what set of molecules may be detected (GPLxxx).
Type	Type of the sample of experiment.
Organism	The organism(s) from which the biological material was derived for experiment.
Molecule	Type of molecule that was extracted from the biological material.
Label	The compound used to label the extract.

This table lists the structured metadata elements along with a description of each element.

The GEO metadata contains 1,368,682 individual sample records.
It contains around 50,000 studies, called “series”.
A series is identified with a series id (i.e. GSExxxxx) consists of one or more samples.
A sample (identified with GSMxxxxx) describes the set of molecules that are being probed.

Element Name	Number of classes	Number of selected classes	Example Values
Platform	12431	2697	gpl570, gpl1261, gpl96, gpl10558
Type	7	7	rna, genomic, sra, protein
Organism	2434	537	homo sapiens, zea mays
Molecule	9	9	total rna, genomicdna, polya rna, protein
Label	1641	454	biotin, cy3, cy5 and cy3, alexa fluor 647

Number of classes in our experimental setup. This table shows the number of classes which constitute as well as example values, for each structured element.

Rule Learning Algorithms

Apriori

Apriori algorithm was proposed by Agrawal and Srikant. Rule discovery is based on making frequent item sets. An item set is called frequent when its support is above a defined minimum support. Apriori requires many database scans.

Predictive Apriori

Predictive Apriori algorithm was proposed by Scheffer. Predictive Apriori returns the n best rules that maximize the accuracy and minimize the number of database scans compare with Apriori.

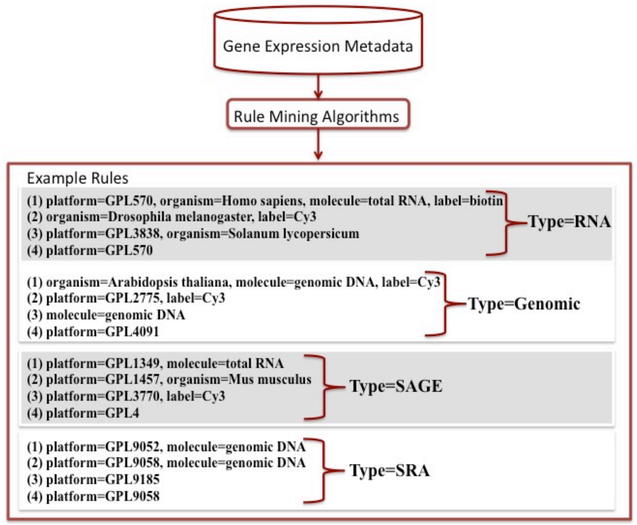
Decision Table

Decision Table was proposed by Kohavi. This algorithm is based on a decision tree where each node represents a feature and each branch represents a value that the node can assume.

PART

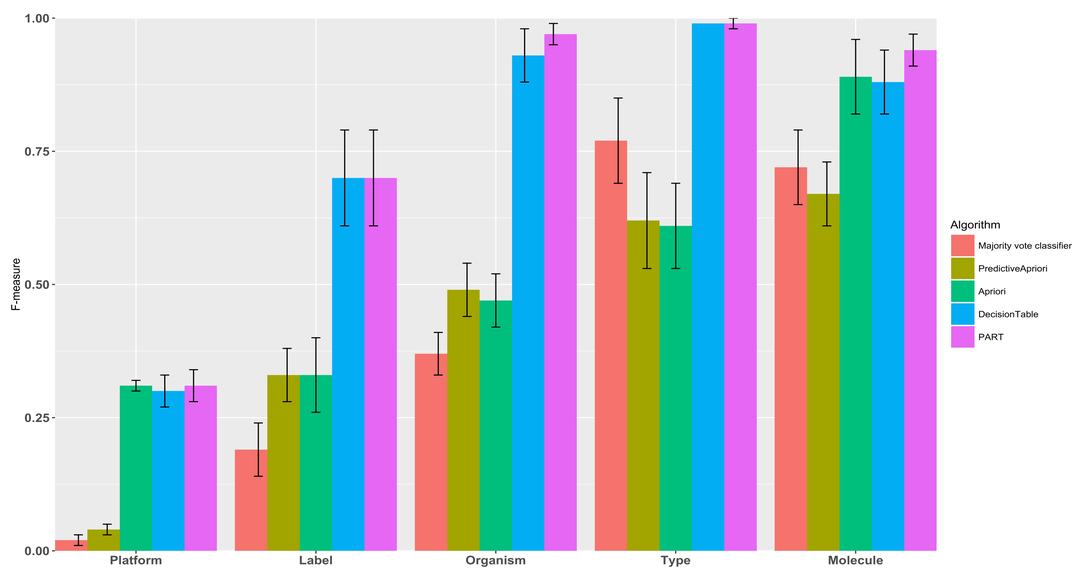
PART was proposed by Frunkranz. This algorithm uses partial trees to generate near-optimal decision list and a single rule is extracted from the list.

Rules Example



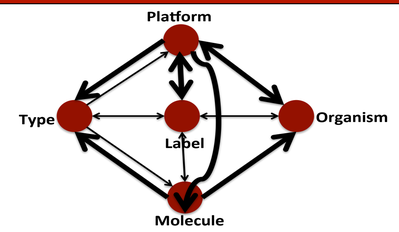
Set of the rules to predict different types. The average of 5000 rules are generated to predict the value of type, organism, molecule, platform and label with all algorithms.

Evaluation



F-measure for weighted class averages for each element.

Model Network



This model shows the association between elements for PART algorithm.

Association between element for PART algorithm. thin link : 0.05 < association <= 0.5, thick link: association > 0.5

Conclusion

Our work suggests that experimental metadata can be accurately predicted using rule mining algorithms. It has implications for augmentation of metadata quality.

Acknowledgement

This research is supported by CEDAR (U54 AI117925) awarded by the National Institute of Allergy and Infectious Diseases.