Klarigi: Explanations for Semantic Groupings

Luke T Slater[1,2,4*], John A Williams[1], Andreas Karwath[1,2,4], Sophie Russell[1], Samantha C Pendleton[1], Hilary Fanning[2,4], Simon Ball[2,4], Paul Schofield[5], Robert Hoehndorf[3], Georgios V Gkoutos[1,2,4]

1 Institute of Cancer and Genomic Sciences, University of Birmingham, UK,
2 University Hospitals Birmingham, NHS Foundation Trust,UK,
3 Computational Bioscience Research Center, King Abdullah University of Science and Technology, SA
4 MRC Health DataResearch UK (HDR UK) Midlands, UK
5 Dept of Physiology, Development, and Neuroscience, University of Cambridge, UK

*** l.slater.1@bham.ac.uk**

Keywords: ontology, semantic analysis, mimic-iii, text mining, cluster explanation

Abstract

Summary: Semantic annotation facilitates the use of background knowledge in analysis. This includes approaches that sort entities into groups, clusters, or assign labels or outcomes that are typically difficult to derive semantic explanations for. We introduce Klarigi, a tool that creates semantic explanations for groups of entities described by ontology terms implemented in a manner that balances multiple scoring heuristics. We demonstrate Klarigi by using it to identify characteristic terms for text-derived phenotypes of emergency admissions for two frequently conflated diagnoses, pulmonary embolism and pneumonia. Klarigi provides a universal method by which entity groups or labels can be explained semantically, and thus contributes to improved explainability of analysis methods.

Availability and Implementation: Klarigi is freely available under an open source licence at `http://github.com/reality/klarigi`. Supplementary data is available with this article.

Contact: l.slater.1@bham.ac.uk

1 Introduction

Over the last two decades, biomedical knowledge has increasingly been represented in the form of ontologies. Ontologies provide a large corpus of formalized knowledge, facilitating the use of background knowledge in analysis and knowledge synthesis across many biomedical disciplines. Ontology-based analysis has been leveraged across many tasks including prediction of protein interaction and rare disease variants [1]. In the clinical space, similar analysis methods have been applied across a wide range of applications including diagnosis of rare and common diseases [2,3], as well as the identification of subtypes of diseases, such as autism [4]. In addition, the synthesis of ontology-based methods and machine learning is increasingly common [5]. Despite the increasing use of semantics in analysis, the anticipated subsequent derivation of semantic explanations for classifications, outcomes, labels, or groups generated by those analyses, remains a challenging task, and a major practical, ethical, and technical issue in biomedical analysis.

Semantic explanation is the task of producing, given a set of entities described by ontology terms, a set of terms that characterises the set of entities. Several previous methods have been developed to achieve this, such as semantic regression, which seeks to describe the functionality of clustered genes or protein arrays [6]. These approaches are often concerned with genetics, focusing on the measurement of the probability of terms appearing in a group. For example, gene enrichment analysis coupled with with a hypergeometric test identifies terms that are significantly over-expressed in a set of genes [7].

Our approach improves upon these methods in several ways. By introducing several heuristics that measure a candidate term's explanatory power, the approach provides multiple metrics for configuration and interpretation. Furthermore, hypergeometric gene enrichment is a univariate method, while Klarigi produces sets of terms which, considered individually or together, exclusively characterises multiple groups. We have previously applied this approach to faceted clusters of text-derived phenotypes [8]. However, in this work, we generalise the algorithm, and present a standalone application that can be used with any group or set of groups of entities associated with ontology classes.

2 Approach

Our approach calculates three heuristics to measure the explanatory power of candidate terms: inclusivity, exclusivity, and specificity. Inclusivity measures the proportion of entities in a group of terms where at least one is subclass of or equivalent to the candidate term. Conversely, exclusivity measures the proportion of entities in *other* groups of terms with at least one being a subclass of or equivalent to the candidate term. Specificity is a measure of term specificity, calculated through a configurable information content measure. These scores are calculated for all classes associated with all members of a group and their superclasses.

Klarigi then uses these heuristics to identify explanatory sets of terms for the group. To evaluate explanatory sets, we further define measurements of overall inclusivity and exclusivity. Overall inclusivity measures the proportion of group members that contain at least one term that is a subclass of a term in the explanatory set. Conversely, overall exclusivity measures the proportion of members of other groups that are excluded by at least one term in the explanatory set. This process involves optimisation of several variables, and can therefore can be considered as a multiple objective optimisation problem, considering the scoring heuristics as objective functions. The ε-constraints solution retains one objective function, and transforms the rest into a set of constraints between which the remaining objective function can be optimised [9]. Our method is based upon this solution, retaining overall inclusivity as the objective function. However, instead of optimising this within a set of static constraints, it steps down through upper constraint boundaries in a priority order, to optimise overall inclusivity while also identifying large values of the other measures. A full characterisation of the measures and method is available in the supplementary material.

3 Use Case: Pulmonary Embolism

Pulmonary embolism, a condition associated with considerable mortality rates, presents in ways that render the conditions difficult to diagnose when associated with other comorbidities, such as COPD, and typically shares symptoms with other more common conditions, such as pneumonia and acute bronchitis [10]. The critical time dependence of treatment on diagnosis makes it important to identify combinations of discriminating symptoms as rapidly as possible [11]. To demonstrate Klarigi's functionality, and to gain insight into the phenotypic presentations associated with pulmonary embolism and pneumonia, we created and evaluated text-derived phenotype profiles for characterising terms.

We identified 337 admissions in MIMIC-III [12] whose primary coded diagnosis was pulmonary embolism (ICD-9:41519), and 704 with pneumonia (ICD-9:486), for a total of 1,041 admissions. We then used Komenti [13] to perform concept recognition on the discharge letters for the admissions with the Human Phenotype Ontology (HPO), identifying 43,597 terms in total. We then excluded negated and uncertain terms, using Komenti, producing a set of phenotype profiles for the

Table 1. Explanatory sets for text phenotypes derived for admissions whose primary diagnosis was pulmonary embolism or pneumonia.

pulmonary embolism (337 members)	**Exclusion**	**Inclusion**	**IC**
Chest pain (HP:0100749)	0.71	0.39	1.0
Hypertension (HP:0000822)	0.52	0.5	0.89
Dyspnea (HP:0002094)	0.46	0.43	0.82
Increased blood pressure (HP:0032263)	0.52	0.5	0.82
Abnormal breath sound (HP:0030829)	0.38	0.41	0.8
Abnormal systemic blood pressure (HP:0030972)	0.38	0.63	0.76
Edema (HP:0000969)	0.4	0.58	0.67
Abnormality of fluid regulation (HP:0011032)	0.37	0.6	0.67
Overall	0.86	0.96	-
pneumonia (704 members)	**Exclusion**	**Inclusion**	**IC**
Hypertension (HP:0000822)	0.5	0.48	0.89
Cough (HP:0012735)	0.73	0.63	0.87
Dyspnea (HP:0002094)	0.57	0.54	0.82
Increased blood pressure (HP:0032263)	0.5	0.48	0.82
Fever (HP:0001945)	0.74	0.47	0.82
Abnormal breath sound (HP:0030829)	0.59	0.62	0.8
Overall	0.7	0.97	-

admissions consisting of all positive concept mentions in their discharge letters. This constitutes grouped data with which Klarigi can derive characteristic explanations, shown in Table 1.

Our findings almost precisely mirror those reported by [10], although we do not have imaging and clinical pathology data available. Particularly, that there is a strong cross-over in the characteristic phenotypes associated with the two diseases. Many phenotypes, such as chest pain, have exclusion and inclusion values that add up to around one, indicating low discriminatory power. Several individual phenotypes show greater discriminatory power, with cough and fever being more strongly and exclusively associated with pneumonia. Moreover, overall inclusivity and exclusivity values show that both explanatory sets, are discriminatory (though many individual terms are not). We also find that edema, not considered by [10], is a discriminator when it appears with other pulmonary embolism-associated phenotypes.

4 Conclusion

Klarigi enables researchers to create semantic explanations for any entity groups associated with ontology terms. As such, it presents a contribution to the reduction of unexplainability in semantic analysis.

Ethical approval

This work makes use of the MIMIC-III dataset, which was approved for construction, de-identification, and sharing by the BIDMC and MIT institutional review boards (IRBs). Further details on MIMIC-III ethics are available from its original publication (DOI:10.1038/sdata.2016.35). Work was undertaken in accordance with the MIMIC-III guidelines.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

GVG and LTS acknowledge support from support from the NIHR Birmingham ECMC, NIHR Birmingham SRMRC, Nanocommons H2020-EU (731032) and the NIHR Birmingham Biomedical Research Centre and the MRC HDR UK (HDRUK/CFC/01), an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health. RH, PNS and GVG were supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/3790-01-01. AK was supported by by the Medical Research Council (MR/S003991/1) and the MRC HDR UK (HDRUK/CFC/01). PNS and GVG acknowledge the support of the Alan Turing Institute, UK.

References

1. Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: A Whole-Phenome Approach to Disease Gene Discovery. Nucleic Acids Research. 2011 Oct;39(18):e119–e119.
2. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies. The American Journal of Human Genetics. 2009 Oct;85(4):457–464.
3. Slater LT, Karwath A, Williams JA, Russell S, Makepeace S, Carberry A, et al. Towards Similarity-Based Differential Diagnostics for Common Diseases. Computers in Biology and Medicine. 2021 Jun;133:104360.
4. Veatch OJ, Veenstra-Vanderweele J, Potter M, Pericak-Vance MA, Haines JL. Genetically Meaningful Phenotypic Subgroups in Autism Spectrum Disorders. Genes, Brain, and Behavior. 2014 Mar;13(3):276–285.
5. Kulmanov M, Smaili FZ, Gao X, Hoehndorf R. Machine Learning with Biomedical Ontologies. bioRxiv. 2020 May:2020.05.07.082164.
6. Greene D, Richardson S, Turro E. Phenotype Similarity Regression for Identifying the Genetic Determinants of Rare Diseases. American Journal of Human Genetics. 2016 Mar;98(3):490–499.
7. Gentleman R, Morgan M, Huber W. Gene set enrichment analysis. In: Bioconductor Case Studies. Springer; 2008. p. 193–205.
8. Slater LT, Williams JA, Karwath A, Fanning H, Ball S, Schofield P, et al. Multi-Faceted Semantic Clustering With Text-Derived Phenotypes. medRxiv. 2021 May:2021.05.26.21257830.
9. Haimes Y. On a bicriterion formulation of the problems of integrated system identification and system optimization. IEEE transactions on systems, man, and cybernetics. 1971;1(3):296–297.
10. Paparoupa M, Spineli L, Framke T, Ho H, Schuppert F, Gillissen A. Pulmonary Embolism in Pneumonia: Still a Diagnostic Challenge? Results of a Case-Control Study in 100 Patients. Disease Markers. 2016;2016:1–8. Available from: `https://doi.org/10.1155/2016/8682506`.
11. Bělohlávek J, Dytrych V, Linhart A. Pulmonary embolism, part I: Epidemiology, risk factors and risk stratification, pathophysiology, clinical presentation, diagnosis and nonthrombotic pulmonary embolism [Journal Article]. Exp Clin Cardiol. 2013;18(2):129–38.

12. Johnson AEW, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, et al. MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data. 2016 May;3(1):1–9.

13. Slater LT, Bradlow W, Hoehndorf R, Motti DF, Ball S, Gkoutos GV. Komenti: A Semantic Text Mining Framework. bioRxiv. 2020 Aug:2020.08.04.233049.