Towards a Directory of Rare Disease Specialists: Identifying Experts from Publication History

Zihan Wang AVIN.WANG@MAIL.UTORONTO.CA
Division of Engineering Science
University of Toronto
Toronto, ON, Canada

Michael Brudno BRUDNO@CS.TORONTO.EDU
Centre for Computational Medicine
Hospital for Sick Children
Toronto, ON, Canada

Orion Buske BUSKE@CS.TORONTO.EDU
Centre for Computational Medicine
Hospital for Sick Children
Toronto, ON, Canada

Abstract

Accurate referral to a medical specialist is a challenging part of medical care, especially for patients with rare diseases. Because of the diversity of rare diseases, finding a specialist that has experience with the particular rare disease is important. This burden often falls on the patients and families, but they do not necessarily have the time or scientific expertise to evaluate the medical literature to identify experts. To help patients, families, and general practitioners find specialists in a particular rare disease, we trained machine learning models to predict the expertise of researchers in every rare disease based on their publication record. We compile a dataset of 209,110 disease-author associations from the literature and evaluate the performance of six machine learning methods, classifying known rare disease experts with 79.4% accuracy and predicting 41,129 disease-expert associations.

1. Introduction

Rare diseases affect over 350 million people worldwide, 50% of whom are children (Jacqueline, 2017; Mendlovic et al., 2016). Because each disease is rare, a general physician may have never seen a patient with a particular disease in their entire career. Symptoms can also vary considerably between individual cases, making correct diagnosis difficult (Singh et al., 2013). Misdiagnosis and incorrect treatment are extremely costly and potentially dangerous, so it is important for patients to be referred to specialists that have experience with their condition. For rare and undiagnosed conditions, the burden can often fall on the patient to find a specialist that has experience with their condition. The National Institutes of Health (NIH) website for rare diseases recommends patients to review the medical literatures themselves to find specialists.[1] However, it can be a daunting task for individ-

1. See https://rarediseases.info.nih.gov/guides/pages/25/how-to-find-a-disease-specialist

uals without a medical background to evaluate the literature to identify experts in related conditions.

Automated expert assessment and discovery systems have been applied to a number of problems, from companies identifying subject matter experts to hire (Maybury, 2006), to researchers identifying reviewers for submitted papers (Charlin et al., 2013). Some expert finding systems are based on a manual process, and use expert self-nomination and a knowledge directory system (Vivacqua et al., 1999). While manual approaches are difficult to scale, data mining can be used to identify expertise from large datasets in an automated fashion. The Internet enabled new research into expert finding systems that use email communications (Foner, 1997; Campbell et al., 2003), online bulletin board data (Swartz et al., 1993; Krulwich et al., 1996), and recently social network data (Xie et al., 2016) to identify experts. Several methods, including Referral Web (Kautz et al., 1997) and Autonomy[2], have used authors' publication history to determine their expertise, and recently machine learning methods have been applied to propose reviewers for academic papers (Charlin et al., 2013). Yet, to our knowledge, such expert finding systems have not yet been applied to finding specialists in particular disease areas.

To facilitate the specialist referral process, we trained machine learning models to predict the expertise of researchers in every rare disease based on their publication record. We compare the performance of six methods (a baseline method, logistic regression, SVM, random forest, Naive Bayes, and a neural network) on a dataset of 209,110 disease-author associations and are able to classify rare disease experts from GeneReviews with 79.4% accuracy and predict 41,129 new disease-expert associations. Automated specialist evaluation methods have immense potential to help patients, families, and general practitioners search for specialists for rare diseases.

2. Materials and Methods

2.1 Data Description

We compiled a data set of 2,160 known disease-expert associations and 206,950 unknown disease-author associations, based on the authorship of publications associated with rare diseases. Known disease-expert associations were obtained by downloading 664 GeneReviews chapters.[3] GeneReviews provides high-quality peer-reviewed summaries of a variety of inherited conditions (Pagon, 1993). GeneReviews chapters are written by one or more experts and focus on a specific condition or disease. Unknown disease-author associations were obtained using the API for OMIM.org, an online catalog of rare inherited diseases (Amberger et al., 2015). OMIM provides a list of primary source publications associated with each rare disease. The positive disease-expert associations from GeneReviews publications were combined with the unlabeled disease-author associations from OMIM to form the complete data set.

2. http://www.ttivanguard.com/sfreconn/autonomy.pdf
3. Code available at https://github.com/AvinWangZH/RareDiseaseExpertIdentification

2.2 Processing

We preprocess the data in two main steps: name standardization and mapping disease-author associations. GeneReviews provides full author names, but the OMIM reference lists use IEEE citation format (e.g., Smith, J.). In order to compare across these datasets, we first convert author names from GeneReviews into the IEEE format (e.g. James Smith as Smith, J..) to match those in OMIM. Next, we map disease-author associations from GeneReviews to those from the OMIM API using the OMIM associations provided by most GeneReviews chapters. We filtered out 35 GeneReviews articles that did not provide any corresponding OMIM identifier. For GeneReviews articles that listed multiple OMIM IDs, each was evaluate separately. The 2,160 positive disease-author associations were taken as the set of 6,555 GeneReviews disease-author associations that also appeared in OMIM. Table 1 shows an example of the processed data.

Table 1: Snapshot of processed GeneReviews dataset and corresponding OMIM publications

GeneReviews disease name	GeneReviews authors (considered experts)	Related OMIM ID	Number of publications linked from OMIM	Number of unique authors
Cohen Syndrome	Heng Wang, Marni J. Falk, Christine Wensel, Elias I Traboulsi	216550	57	291
		607817	14	157
GLB1-Related Disorders	Debra S Regier, Cynthia J Tifft	230500	36	138
		230600	14	59
		230650	25	84
		253010	15	77
		611458	47	196

2.3 Feature Design

We annotated each author-disease association with 18 features across five categories, shown in Table 2: (i) Author Publications; (ii) Disease Publications; (iii) Publication Contribution; (iv) Year of Publication; (v) Publication venues. Author Publications are quantitative measurements of the number of publications by the researcher, overall and for the specific disease. Disease Publications are metrics across publications for the disease, to capture differences in publication volume between different diseases. Publication Contribution measures the impact of the researchers placement in authorship lists, such as first author or last author. Expertise may change over time, so features were added to capture temporal publication metrics. Not all publications are equally impactful, so we added features to count publications in top journals separately. The top 3, 5, and 10 journals were identified based on h5-index in Google Scholar for the fields of Genetics and Health and Medical Science.

2.4 Algorithm Comparison and Selection

We evaluated five different machine-learning algorithms to predict if a person is an expert, and compared them to a baseline measure using logistic regression on the number of publications the author has for the disease (Feature 1 in Table 2):

1. Logistic regression (LogR): used logistic function to predict the score of each person expertise (Freedman, 2009).
2. Supported vector machine (SVM): used the SVR package from sci-kit learn (Varoquaux et al., 2015; Smola et al., 2004), version 0.18.1.
3. Nave Bayes (NB): implemented using sci-kit learn, version 0.18.1, with an assumption of independence between features (Zhang, 2004).
4. Random Forest (Forest): a random forest with 100 trees, implemented with the RandomForestClassifier package from sci-kit learn (Breiman, 2001), version 0.18.1.
5. Neural Network (NNet): implemented as a single fully connected hidden layer with 300 hidden units using Tensorflow (Abadi, 2016).

For all these methods, we used 5-fold cross-validation over all positive examples and a random subset of the unknown examples as negatives. Because there is a large number of unknown author-disease associations, it is better to balance the positive data and negative data in both training set and test set.

Table 2: The list of features annotated for each disease-author association.

Feature Category	Feature
i) Author Publications	1) # of Publications of an Author on the OMIM ID
	2) Normalized Feature 1
	3) # of OMIM ID the Author Published on
ii) Disease Publications	4) # of Authors Published on the OMIM ID
	5) Total Number of Publications on the OMIM ID
	6) Feature 1/Feature 5
iii) Publication Contribution	7) # of Publications on the OMIM ID as First Author
	8) # of Publications on the OMIM ID as Last Author
iv) Year of Publication	9) # of Publications on the OMIM ID in 3 years
	10) # of Publications on the OMIM ID in 5 years
	11) # of Publications on the OMIM ID in 10 years
v) Publication Venue	12) # of Publications in the Top 3 Health and Medical Science Venues
	13) # of Publications in the Top 5 Health and Medical Science Venues
	14) # of Publications in the Top 10 Health and Medical Science Venues
	15) # of Publications in the Top 3 Genetic Venues
	16) # of Publications in the Top 5 Genetic Venues
	17) # of Publications in the Top 10 Genetic Venues
	18) # of Publications in Nature or Science

3. Results

The performance of the 5 machine learning methods and the baseline measure are shown in Figure 1. The Random Forest, Logistic Regression, and Neural Network models all performed similarly, with the Random Forest having the best performance with a ROC AUC of 0.88, with the baseline measure of logistic regression on the number of disease publications having a ROC AUC of 0.69. With a cutoff score of 0.5, the Random Forest method achieved an accuracy of 79.5%, with 80% precision and 78% recall.

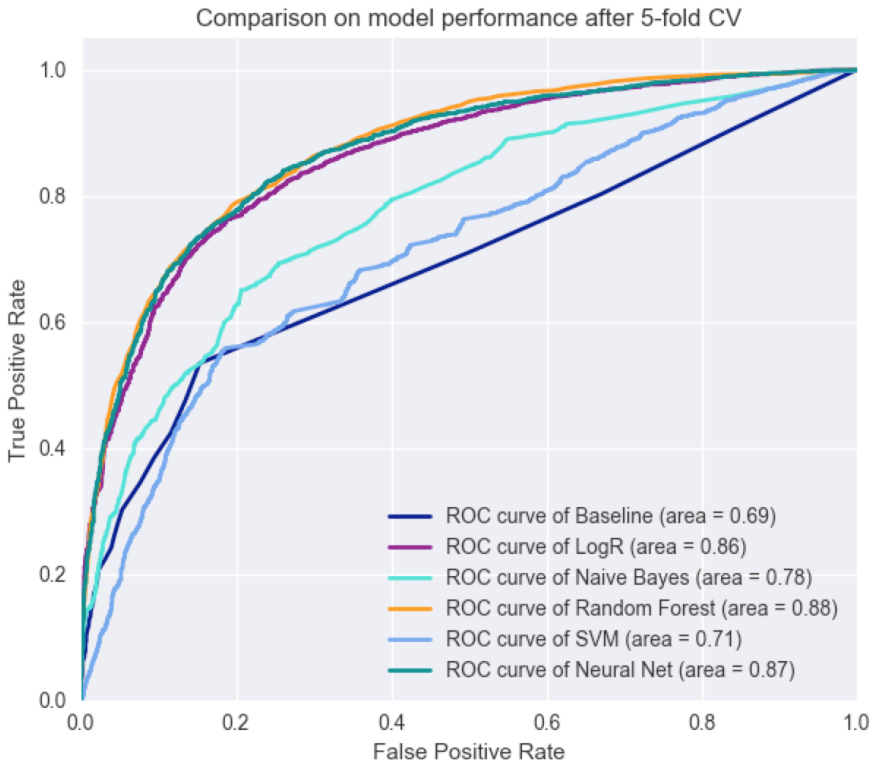

Figure 1: ROC curve comparing the performance of each machine learning model on held out data during 5-fold cross validation.

Using the Random Forest method, we compared the importance of each feature across the folds of cross-validation, with the results shown in Figure 2. Author Publication and Disease Publication features were weighted the most highly, with publication year and venue being the least important features. We also measured the distribution of scores (posterior probabilities) from the Random Forest model for the positive and negative test datasets, shown in Figure 3.

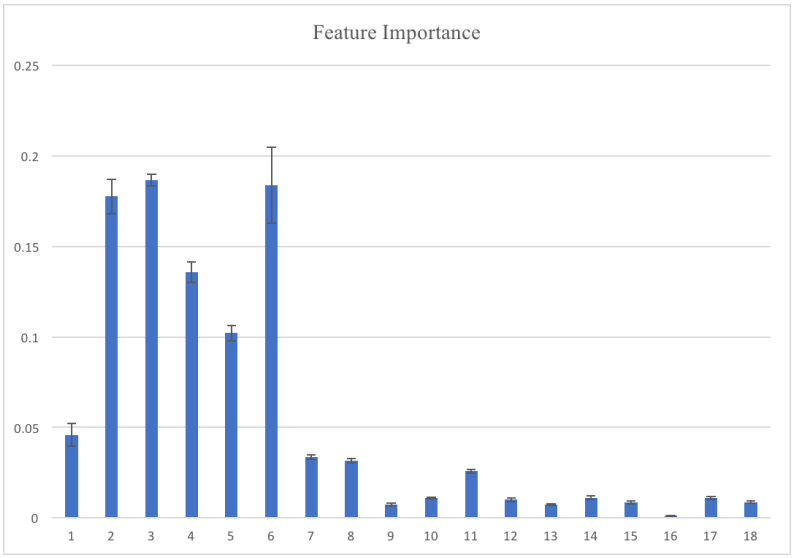

Figure 2: The relative importance (scored on the y axis) of each feature (along the x axis) as evaluated by the sci-kit learn Random Forest model, with the mean and 95% confidence intervals shown in error bars.

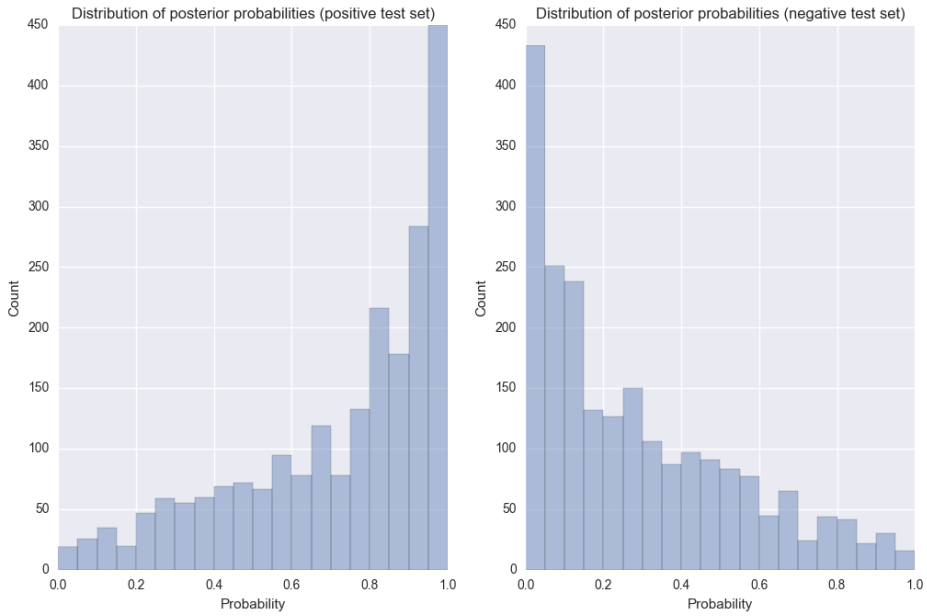

Figure 3: Posterior probabilities for the Random Forest model on the held-out examples during cross-validation.

We then used the Random Forest model, to score the full set of 206,950 unlabeled disease-author associations after training on the full set of positive associations and a random set of unlabeled associations. Table 3 shows the top predictions for a random set of OMIM IDs. The highest predicted score is 100%, the fourth quartile is from 44% to 100%, and the third quartile is 20%-44%. The median is at 20%.

In the unknown data set, the classifier predicted 41,129 disease-expert association as positive out of 206,950 unlabeled associations (∼20%) with 0.5 cut-off score.

Table 3: Statistics and the top-scoring experts for five random OMIM diseases using the trained Random Forest model.

OMIM ID	Disease Name	Number of publications	Number of unique authors	Number of "experts" classified	Top 5 ranked authors
146920	ADENOSINE DEAMINASE, RNA-SPECIFIC; ADAR	21	121	34	Rice, G. I. Weier, H.-U. G. Tomita, Y. Livingston, J. H. Kondo, T.
201750	ANTLEY-BIXLER SYNDROME WITH GENITAL ANOMALIES AND DISORDERED STEROIDOGENESIS; ABS1	10	72	32	Jabs, E. W. Miller, W. L. Pandey, A. V. Fluck, C. E. Arlt, W.
203800	ALSTROM SYNDROME; ALMS	31	150	25	Nishina, P. M. Naggert, J. K. Collin, G. B. Wilson, D. I. Martin, M.
609310	COLORECTAL CANCER, HEREDITARY NON-POLYPOSIS, TYPE 2; HNPCC2	18	164	25	Lindblom, A. Thibodeau, S. N. Nordenskjold, M. Gallinger, S. Peltomaki, P.
600275	NOTCH, DROSOPHILA, HOMOLOG OF, 2; NOTCH2	21	134	56	Oakey, R. J. Gridley, T. Kaplan, P. Robertson, S. P. Majewski, J.

4. Discussion

In this experiment, we compared five different learning algorithms in their ability to predict expertise in a particular rare disease based on their publication history. We find that many methods perform well, achieving an ROC AUC of above 85% discriminating verifiable experts (GeneReviews authors) from other researchers that have published on the same disease. When we compared the relative importance of the various features, both author-specific and disease-specific features were highly important, suggesting that an author's publication history is most informative when interpreted in the the context of the literature for each particular disease. Interestingly, features related to the recency of publications or the publication venue were not as important in predicting expertise.

As the set of real experts is much larger that just the GeneReviews authors, we expect many of the false positives to represent real expertise. This absence of true negative examples is a limitation of the performance evaluation. To obtain a more accurate measure of performance, domain experts should manually evaluate the appropriateness of the predicted experts. Another limitation of this work is the potential for ambiguous author names to affect the results. While the designed features enable classifiers to avoid ambiguous author names, this prevents the model from identifying real experts with common names. However, common names also present challenges for patients and specialists. Without affiliation information, unambiguous names are the most useful to recommend.

These results provides a proof-of-concept for predicting disease expertise from publication history, and demonstrate the feasibility of helping patients find relevant specialists without needing to manually review the medical literature.

References

Martin Abadi. TensorFlow: learning functions at scale. ACM SIGPLAN Notices, vol. 51, no. 9, pp. 1-1, 2016.

Joanna S. Amberger, Carol A. Bocchini, Francois Schiettecatte, Alan F. Scott, and Ada Hamosh. OMIM.org: Online Mendelian Inheritance in Man (OMIMR), anonline catalog of human genes and genetic disorders. Nucleic acids research, 43(D1):D789?D798, 2015.

Leo Breiman, Random Forests. Machine Learning, 45(1), 5-32, 2001.

Christopher Campbell, Paul Maglio, Alex Cozzi, and Byron Dom. Expertise Identification using Email Communications. In Proceedings of ACM Conference on Information and Knowledge Management CIKM. New Orleans, LA. 528-531, 2003.

Laurent Charlin and Richard Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system. in Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, W&CP volume 28. 2013

Leonard N. Foner. Yenta: A Multi-Agent Referral System for Matchmaking System. Proceedings of The First International Conference on Autonomous Agents, Marina Del Ray, CA, 1997 .

David A. Freedman. Statistical Models: Theory and Practice. Cambridge University Press. p. 128. 2009.

GlobalGenes. RARE Diseases: Facts and Statistics, Global Genes, 2017. URL: https://globalgenes.org/rare-diseases-facts-statistics/.

Henry Kautz, Bart Selman, Mehul Shah. Referral Web: combining social networks and collaborative filtering. Communications of the ACM. 40(3): 63-65, 1997.

Bruce Krulwich and Chad Burkey. The ContactFinder agent: Answering bulletin board questions with referrals. In AAAI-96, 1996.

Mark T. Maybury. Expert Finding Systems, 1st ed. Bedford: The MITRE Corporation, pp. 7-8, 2006.

Joseph Mendlovic, Hila Barash, Hadar Yardeni, Yonit Banet-Levi, Hagith Yonath, and Annick Raas-Rothschild. Rare diseases DTC: Diagnosis, treatment and care. Harefuah,155(4):241?253, 2016.

Roberta A. Pagon. GeneReviews. University of Washington, 1993.

Hardeep Singh, Traber D. Giardina, Ashley N.D. Meyer, Samuel N. Forjuoh, Michael D. Reis, and Eric J. Thomas. Types and origins of diagnostic errors in primary care settings. JAMA Internal Medicine, 173(6):418, 2013.

Alex J. Smola and Bernard Schlkopf. A Tutorial on Support Vector Regression. Statistics and Computing, vol. 14, no. 3, pp. 199-222, 2004.

Michael F. Swartz and D. C. M. Wood. Discovering shared interests using graph analysis. Communications of the ACM. 36(8): 78-89, 1993.

G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa and A. Mueller. Scikit-learn: Machine Learning Without Learning the Machinery. GetMobile: Mobile Computing and Communications, vol. 19, no. 1, pp. 29-33, 2015.

Adriana S. Vivacqua. Agents for Expertise Location. In Proceedings 1999 AAAI Spring Symposium on Intelligent Agents in Cyberspace. Technical Report SS-99-03. Stanford, CA, USA, March 1999.

Xiaoqin Xie, Yijia Li, Zhiqiang Zhang, Haiwei Pan, Shuai Han. A Topic-Specific Contextual Expert Finding Method in Social Network. In: Li F., Shim K., Zheng K., Liu G. (eds) Web Technologies and Applications. APWeb 2016. Lecture Notes in Computer Science, vol 9931. Springer, Cham, 2016.

Harry Zhang. The Optimality of Naive Bayes. in FLAIRS Conference, V. Barr and Z. Markov, Eds. AAAI Press, 2004.