**PatientKind Expert: A Novel Method for Rare Disease Expert Classification**

**Introduction:**

Rare diseases have different definitions from countries. In Europe, a disease or disorder has less than 1 affection in 2,000 people called rare. In USA, a disease affects less than 200,000 Americans defined as rare [1]. Unlike influence of a single rare disease, the rare diseases in a whole affect around 6%-8% of the general population and the 50% of patients are children [1, 2]. Based on the rarity of these diseases, a general physician may not likely have a single patient in his/her whole career [3]. Furthermore, the diagnosis of these diseases are relatively hard, because the symptoms often vary between individual causes [4]. Therefore, it is important to refer patients to an appropriate healthcare expert who are familiar with the specific disease and symptom set. In order to create a good match between rare disease patients and experts, we created a novel machine learning application by implementing SVM regression on OMIM publication data [5, 6]. By using this machine learning approach, we distinguished 2,1224 experts over 209,110 people with publications in 1,292 diseases.

**Approach:**

In order to evaluate if a person is an expert or not, we have to establish a metrics to measure expertise first. In our project, the metrics are based on people’s publications. On OMIM, which is an online catalog of rare diseases and disorders, each rare disease or symptom has a specific OMIM ID [6]. Under each OMIM ID, related research or clinical publications are included in the Reference Section. We scraped the publication data by using OMIM API and stored them into Python dictionaries [6, 7].

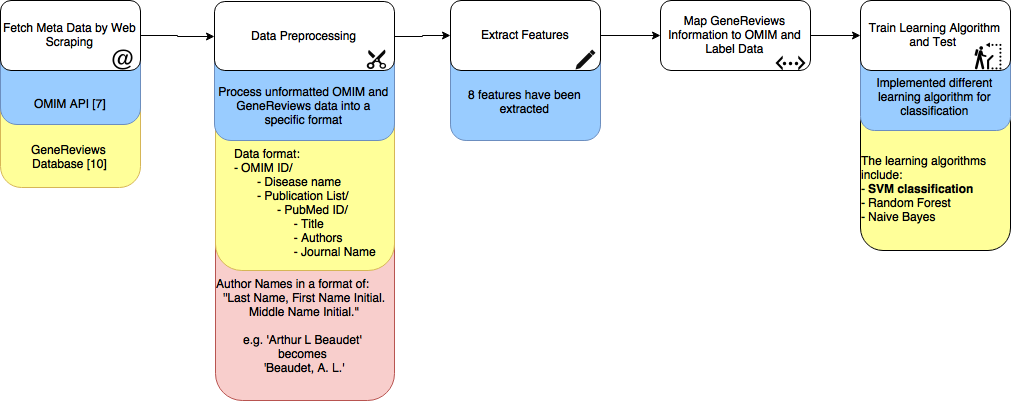


Figure 1: Flow of Experiment Process

The original OMIM data are not formatted. In order to use OMIM data, we have to reorganized them into a data structure which is a dictionary in this case and extract useful information. The useful information includes OMIM ID, disease name, publication list. Each publication contains a unique PubMed ID and we stored publication information, such as paper title, author names and journal name, under each PubMed ID. In the part of author name processing, we found GeneReviews author names had a different format from OMIM author names. Therefore, GeneReviews author names had been parsed into the same format as OMIM author names which only include full last name, first name initial and middle name initial.

Features are the keys for drawing classification boundaries or fitting regression lines [8, 9]. In our algorithm, there were 8 features had been chosen: number of publications on the disease1, normalized number of publications on the disease2, number of diseases that the author has publication on3, number of publications as first author on the disease4, number of publications as last author on the disease5 and the number of publications on the disease in 3 years6, 5 years7 or 10 years8. Feature 1 represents the person’s research ability on the disease in general. Feature 2 shows a relative research ability compared with other researchers on the same disease. Feature 3 reveals if a person’s research interest focus on one specific disease or for many diseases in general. Feature 4 and 5 provide weights to the impact from authorship. The last three features represent the effect from time of publications.

After extracting features from the meta data, mapping from GeneReviews to OMIM ID has been made. GeneReviews is an authorized resource which provides relative information for inherited conditions. The articles are written by one or more expert on the specific condition or disease [10]. We used GeneReviews expert data as our positive dataset. We filtered people with no publications, because our analysis was based on publications. After mapping, there were 2,160 positive data points and 206,950 unknown data points. We randomly marked 2,160 unknown points as negative. Then we used both positive and negative data to train different learning algorithms, which include SVM classification [11], random forest [12] and naïve Bayes [13], and compared the results. We used 10 times tenfold cross-validation to make the results more accurate [14].

**Analysis:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 |
| SVM Classification | **0.7569** | **0.7560** | **0.7590** | **0.7569** |
| Random Forest | 0.7375 | 0.7206 | 0.7766 | 0.7469 |
| Naïve Bayes | 0.7190 | 0.7543 | 0.6494 | 0.6974 |

Table 1: Comparison between different learning algorithm

As shown in Table 1, SVM Classification algorithm has the best performance on accuracy, precision and F1 score. The relatively high performance shows our classifier working properly.

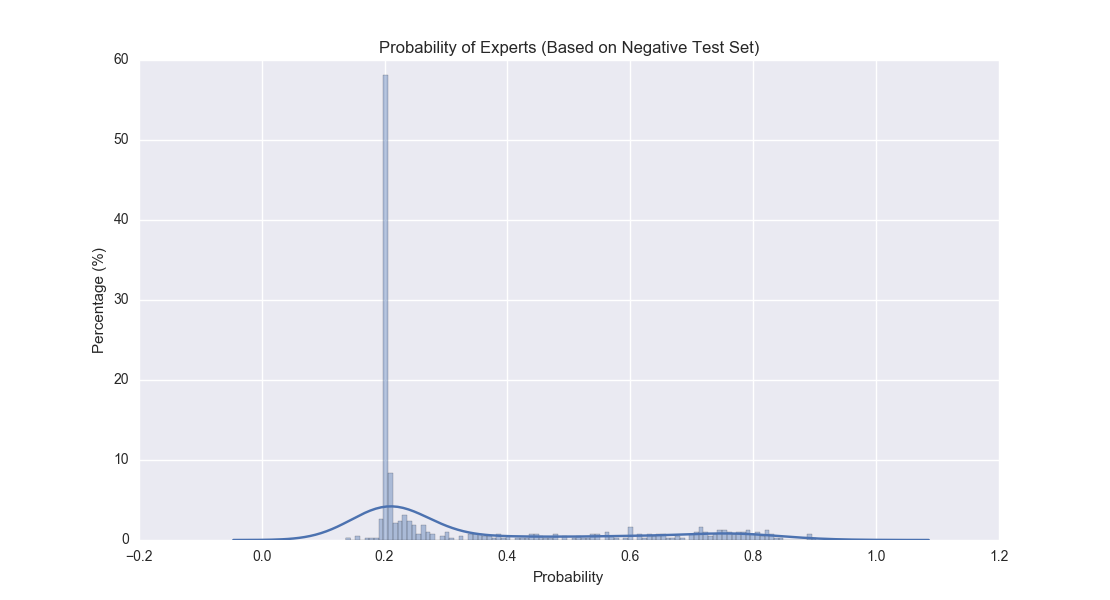


Figure 2: Probability of Experts on Negative Test Dataset

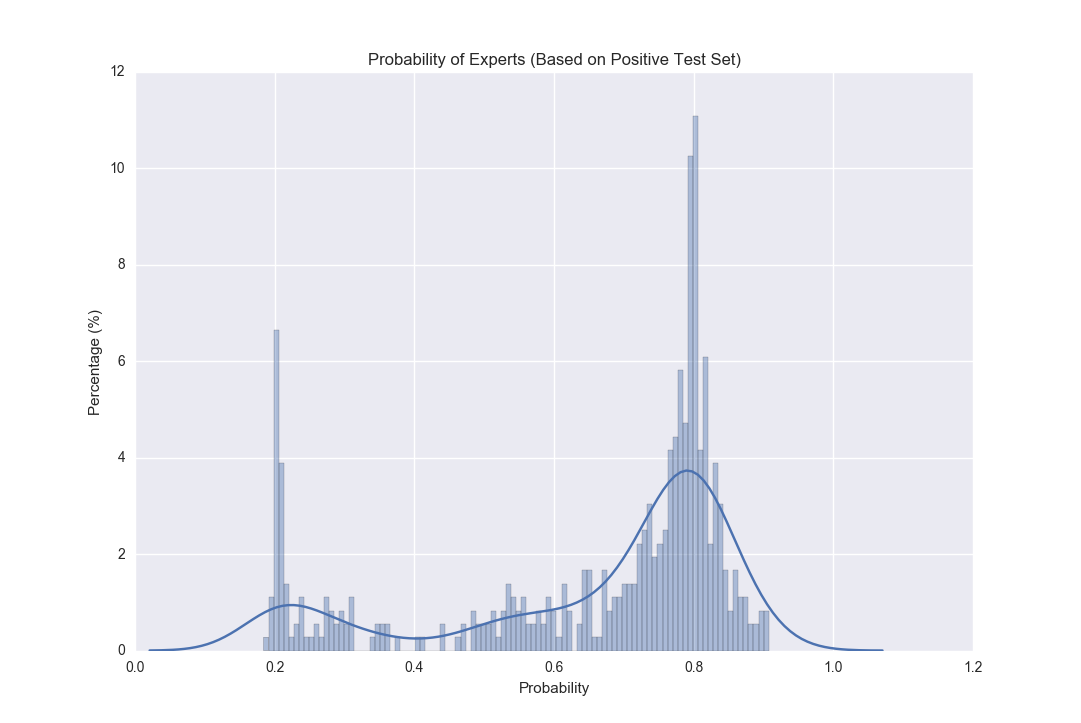


Figure 3: Probability of Experts on Positive Test Dataset

Figure 2 and Figure 3 shows the probability predictions on negative and positive dataset. As shown in Figure 2, the negative data are mainly predicted below 0.5 and most of the points had a score of 0.2. There were also some people having a high score around 0.8 and 0.9. It can be explained by the pollution of our test data. In our test data, we marked unknown data as negative. It is based on the assumption that most people are not experts. However, there are also a portion of people who are the experts but mislabelled. In Figure 3, most positive data points were predicted above 0.5 and clustered at 0.8. Interestingly, there is a significant portion of people who are at 0.2 probability, which is abnormal. Further investigation should be conducted on the abnormal data prediction.

**Conclusion:**

In conclusion, this research discovered the relationship between people’s expertise and their publications. We analyzed people’s publications, extracted useful features and utilized machine learning techniques to classify if a person is an expert on a specific disease. We had a surprisingly good performance on expert classification. By implementing SVM classification on OMIM data, we can make a prediction with a relatively low false positive rate. Moreover, the similar strategy can be conducted into other expert classification problems.

There are also some improvements can be made in the future. Firstly, the testing data can be human labelled to be more accurate. Secondly, more features can be selected to prevent underfitting. Finally, more machine learning algorithms should be implemented and compared. Future study will be focused on implementing these three improvements and to extend expert classification into other fields.

**References:**

[1] "Rare Disease Day 2016 - 29 Feb - Article", Rare Disease Day - 29 Feb 2016, 2016. [Online]. Available: http://www.rarediseaseday.org/article/what-is-a-rare-disease. [Accessed: 22- Aug- 2016].

[2] J. Mendlovic, H. Barash, H. Yardeni, Y. Banet-Levi, H. Yonath and A. Raas-Rothschild, "RARE DISEASES DTC: DIAGNOSIS, TREATMENT AND CARE", Harefuah, vol. 155, no. 4, pp. 241-253, 2016.

[3] "Rare diseases strategy - Publications - GOV.UK", Gov.uk, 2016. [Online]. Available: https://www.gov.uk/government/publications/rare-diseases-strategy. [Accessed: 22- Aug- 2016].

[4] H. Singh, T. Giardina, A. Meyer, S. Forjuoh, M. Reis and E. Thomas, "Types and Origins of Diagnostic Errors in Primary Care Settings", JAMA Internal Medicine, vol. 173, no. 6, p. 418, 2013.

[5] A. Smola and B. Schölkopf, "A tutorial on support vector regression", Statistics and Computing, vol. 14, no. 3, pp. 199-222, 2004.

[6] J. Amberger, C. Bocchini, F. Schiettecatte, A. Scott and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders", Nucleic Acids Research, vol. 43, no. 1, pp. D789-D798, 2014.

[7] "OMIM API", Omim.org, 2016. [Online]. Available: http://www.omim.org/help/api. [Accessed: 22- Aug- 2016].

[8] E. Alpaydin, Introduction to machine learning. Cambridge, Mass.: MIT Press, 2010.

[9] K. Murphy, Machine learning. Cambridge, Mass.: MIT Press, 2012.

[10] GeneReviews. Seattle, WA: University of Washington, 2005.

[11] T. Wu, C. Lin and R. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling", The Journal of Machine Learning Research (JMLR), vol. 5, pp. 975-1005, 2004.

[12] L. Breiman, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[13] "EXPLORING CONDITIONS FOR THE OPTIMALITY OF NAÏVE BAYES", International Journal of Pattern Recognition and Artificial Intelligence, 2016.

[14] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", in International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1143.

[1] <http://www.rarediseaseday.org/article/what-is-a-rare-disease>

[2] <http://www.ncbi.nlm.nih.gov/pubmed/27323543>

[3]<https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/260562/UK_Strategy_for_Rare_Diseases.pdf>

[4] <http://www.ncbi.nlm.nih.gov/pubmed/?term=23440149>

[5] <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.4288>

[6] <http://www.ncbi.nlm.nih.gov/pubmed/25428349>

[7] <http://www.omim.org/help/api>

[8] <http://alex.smola.org/drafts/thebook.pdf>

[9] <https://www.cs.ubc.ca/~murphyk/MLbook/pml-intro-22may12.pdf>

[10] <http://www.ncbi.nlm.nih.gov/books/NBK1116/>

[11] <http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf>

[12] Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[13] <http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf>

[14] <http://web.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf>