Prescription-Based Predictive Modeling of Healthcare Provider Specialty

Sijia Zhang

ABSTRACT

The goal of this project is to build a model to predict the specialty of healthcare providers based on the prescription history in the 2013 Medicare Part D Prescriber dataset. The original dataset contains over 200 specialties with substantially imbalanced sample counts. The specialties with over 10,000 occurrences were undersampled and the specialties associated with a small number of physicians were combined into one category, forming a classification problem of 51 classes. The Medicare Prescriber dataset was transformed into a sparse frequency data matrix containing the total claim count of all generic drugs for each healthcare provider. Singular value decomposition (truncated) was used to reduce the dimension of the feature space and a Random Forest classifier was built for predictive modeling. The model performance using 3-fold cross validation was better than a baseline classifier in terms of accuracy, Matthews Correlation Coefficient and confusion matrix.

WORKFLOW & ANALYTIC INFRASTRUCTURE

The original dataset was first examined to observe data organization, check for missing values and obtain descriptive statistics. Next, a cohort of healthcare providers corrected for large class imbalance was constructed. Features used for predictive modeling were constructed by data aggregation, normalization and transformation. Finally, supervised learning models were trained and tested.

This project was implemented in Python 3.6. Numpy and Pandas libraries were used for data restructuring and analysis. Dimensionality reduction and machine learning were performed using scikit-learn (0.19.1). The code was ran on a 64-bit Windows 10 machine with 18 GB memory.

METHODS

Data Exploration & Descriptive Statistics

The 2013 Medicare Part D Prescriber dataset contains 807973 unique healthcare providers, most of which are individual prescribers (less than 20 organizations). Among the 21 columns/fields, “bene_count” and those associated with beneficiaries age 65 and older have 40-60% missing values. The original data is organized in the long format, with repeated row entries for each healthcare provider. There are 202 specialties with largely unbalanced counts. As shown in Table 1, half of the specialties are associated with fewer than 42 doctors; such small sample size is not sufficient for predictive modeling. The 75-th percentile count is 1081, which is about $^{1}/_{100}$ of the counts of the most common specialty (Internal Medicine).

Class Balancing & Cohort Construction

To address the large class imbalance, the lower ¾ of the specialties were combined into a new specialty category called “others”. This procedure reduced the number of classes to be predicted from 202 to 51. The remaining specialties for predictive modeling are listed in the Appendix. The specialties with over 10,000 occurrence in the original dataset was randomly downsampled to have exactly 10,000 data points. As a result, the sample ratios of each pair of classes lie between 1-to-10 and 1-to-1. Pre-processing of the raw data not only reduced the class imbalance (Table 1), but also decreased the computational complexity by simplifying the muli-classification problem to have fewer classes and reducing the amount of training data to less than 50% of the original cohort size.

Table 1. Descriptive statistics of specialty counts

		Original	Processed
Number of unique specialties		202	51
Count of specialty occurrences	Mean	4000	6358
	Standard deviation	13858	3561
	Minimum	1	1127
	First quartile	4	2701
	Median	41	7349
	Third quartile	1081	10000
	Maximum	104712	10000

Data Reorganization & Feature Construction

The National Provider Identifiers (“npi” column) were used as unique identifiers of healthcare providers; last name, first name, provider city and state fields were dropped since they cannot be used for unique indexing or as informative features for predictive modeling. Drug names and beneficiary/claim counts are descriptors of a doctor’s prescription history. Compared to the “drug_name” filed that includes both brand and generic drug names, the “generic_name” column is less noisy and more indicative of the chemical ingredient which is supposed to associate with the drug targets. Therefore, the generic drugs listed in the original dataset are used as features for specialty prediction. Beneficiary count and total claim count within the year of 2013 can both be used to represent how frequent a certain drug is being prescribed by each healthcare provider. However, the “bene_count” field has over 60% missing data because counts fewer than 11 were suppressed. Therefore, the “total_claim_count” column, which includes no blank entries was used for feature construction.

The dataset was transformed from long format to wide format, with each row representing a unique healthcare provider and each column representing a generic drug. The values in the new data matrix are the sum of total claim count for each combination of drug and prescriber. Since only a few drugs out of 1573 generic drugs were prescribed by each doctor, the feature matrix is sparse. The sparse matrix was stored using the Compressed Sparse Row algorithm in Scipy for efficient machine learning. Values of each feature was scaled to [0, 1] to normalize different variables while preserving sparsity.

Since the number of features is greater than the number of samples of the smallest class, dimensionality reduction was performed to prevent overfitting. Truncated singular value decomposition (SVD) was used, because it may reveal latent correlations between the doctors and prescribed drugs and effectively handles sparse matrices with no need to center the data. The top 300 components were kept, which together explain approximately 75% of the variance.

Supervised Machine Learning

A Random Forest Classifier with 30 estimators were built to predict prescriber's specialty. The Random Forest model with bootstrap was chosen due to the following reasons: (1) it is an ensemble method, which combines several models to reduce variance and overfitting and improves predictive performance; (2) it can handle a large amount of features; (3) it has been shown as an effective model for disease prediction [1-3]. Pilot parametric studies with a sub-sample of the original dataset were performed to determine parameters of the model. The minimum number of samples at a leaf node was set to 5 and the fraction of features used for each split was set to 0.2; the other parameters were set to the default values in the scikit-learn package.

The model was trained and tested using 3-fold cross validation. The predicted classes of the entire dataset was compared to the true class labels for model evaluation. Accuracy (range [0,1]), the average Matthews Correlation Coefficient (MCC) (range [-1,1]) from 1-versus-all binary confusion matrix and multi-class confusion matrix were computed. A baseline classifier with randomly shuffled class labels were built for comparison purpose.

RESULTS

The accuracy of the Random Forest model with 30 estimators is 0.61, which is substantially higher than the 0.03 accuracy obtained from the baseline classifier or simply guessing the majority class label for all samples (Figure 1). Similarly, the binary MCC, the geometric mean of true and false positives and negatives, is 0.51±0.30; the baseline MCC is 0 for random guess and 0.00±0.002 from the baseline classifier (Figure 1). Both the accuracy and MCC increase with the number of estimators used for the Random Forest Classifier (Figure 1). The maximum number of estimators tested was 30 due to time and computational power constraints and relatively small performance improvement observed in Figure 1. The Random Forest model also outperforms the baseline classifier in terms of the multi-class confusion matrix (Figure 2). Most of the diagonal elements of the normalized confusion matrix have values that are much closer to 1 as compared to the rest of the matrix (Figure 2), meaning that the correct class is most often predicted for a majority of the specialties. In contrast, the confusion matrix of the baseline model have values that are all close to 0 (Figure 2).

	Accuracy [0, 1]	MCC [-1, +1]
Original labels	0.61	0.51
Shuffle labels	0.03	0.00
Random guess	0.03	0

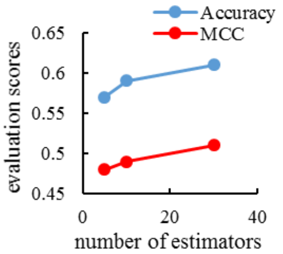

Figure 1. Accuracy and MCC obtained from the Random Forest Classifier, baseline classifier and random guess.

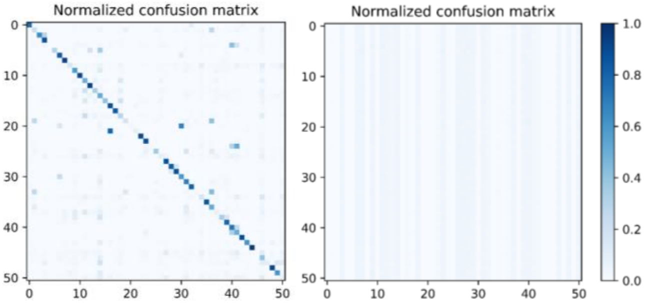

Figure 2. Normalized confusion matrix obtained from the Random Forest Classifier and the baseline classifier. See Appendix for the list of numbered specialties.

DISCUSSION & CONCLUSIONS

Many aspects of the data processing and modeling methods need to be further investigated. For example, the problem formulation may be revisited by exploring deeper on the similarity and hierarchy among the medical specialties to be predicted. It would be interesting to examine why certain specialties are consistently misclassified as another and whether grouping specialties into subtypes would enhance the prediction performance.

Feature construction also requires further consideration. A preliminary test using the beneficiary count as feature values did not increase the prediction accuracy, but other informative descriptors of the prescription history may exist in the original dataset. Other transformation methods, such as classification-driven linear discriminant analysis may be tested, but requires first correcting for the highly skewed data distribution. CUR decomposition may be used instead of SVD to preserve sparsity after decomposition. Feature selection by instance filters and attributor evaluators has been shown to improve the performance of Random Forest models [1] and should be tested for this project.

Future effort is needed towards optimizing the Random Forest Classifier via comprehensive parametric studies. Evaluation of different sampling methods, splitting criteria and maximum tree depths are requisite. Other types of multi-class predictive models, such as distance-weighted K-nearest neighbor and a combination of binary classifiers, should be tested and compared to the current model.

Nevertheless, the current Random Forest model built in the SVD-reduced drug feature space for this project is able to predict the correct class label for most of the 51 specialties the majority of the times. It substantially outperforms the baseline classifier in terms of accuracy and MCC.

REFERENCES

1. Chaudhary A, Kolhe S, Kamal R. 2016. An improved random forest classifier for multi-class classification. *Information Processing in Agriculture*. 3(4): 215-222.
2. Azar AT, Elshazly HI, Hassanien AE, Elkoran AM. 2013. A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*. 113(2): 256-473.
3. Özçift A. 2017. Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. *Computers in Biology and Medicine*. 41(5): 265-271.

APPENDIX

1. Allergy/Immunology; 2. Anesthesiology; 3. Cardiac Electrophysiology; 4. Cardiology; 5. Certified Clinical Nurse Specialist;
6. Colorectal Surgery (formerly proctology); 7. Dentist; 8. Dermatology; 9. Diagnostic Radiology; 10. Emergency Medicine;
11. Endocrinology; 12. Family Practice; 13. Gastroenterology; 14. General Practice; 15. General Surgery; 16. Geriatric Medicine
17. Hematology/Oncology; 18. Infectious Disease; 19. Internal Medicine; 20. Interventional Pain Management; 21. Maxillofacial
Surgery; 22. Medical Oncology; 23. Nephrology; 24. Neurology; 25. Neuropsychiatry; 26. Neurosurgery; 27. Nurse Practitioner;
28. Obstetrics/Gynecology; 29. Ophthalmology; 30. Optometry; 31. Oral Surgery (dentists only); 32. Orthopedic Surgery;
33. Otolaryngology; 34. Pain Management; 35. Pediatric Medicine; 36. Pharmacist; 37. Physical Medicine and Rehabilitation;
38. Physician Assistant; 39. Plastic and Reconstructive Surgery; 40. Podiatry; 41. Psychiatry; 42. Psychiatry & Neurology;
43. Pulmonary Disease; 44. Radiation Oncology; 45. Rheumatology; 46. Specialist; 47. Student in an Organized Health Care
Education/Training Program; 48. Thoracic Surgery; 49. Urology; 50. Vascular Surgery; 51. others