Predicting Corporate Influence Cascades In Health Care Communities

Shouzhong Shi, Chaudary Zeeshan Arif, Sarah Tran

December 11, 2015

Part A Introduction

The standard model of drug prescription choice made by a physician is highly dependent on the physician’s personal experience, for example speciality, professional ages and locations. At the same time, physicians also like to refer to other physician’s decision and leverage their knowledge, like learn from more experienced researchers or physicians. Thus, they also rely on other information they have obtained via the community and social interaction.

By applying machine learning methods, we attempted to explore how a physician’s personal experience at the individual level will drive them to make a prescription decision and if there is any influence from the various corporations making non-medical payments to the physicians. We have narrowed down some useful factors in the given dataset and proposed some future directions for a more concrete predictive model.

Part B Materials and Methods

1. Data Sources

- **General Payment Data - Detailed Dataset 2013 Reporting Year**: This data set contains Physician Financial Relationships. It has 685,296 healthcare providers, 1,554 healthcare companies, and 2,098 teaching hospitals, 14.8 million general payment records, 825,947 research related payments and 9,768 ownership related payment, aggregated 3.6 million payment records.
- **Medicare Provider Utilization and Payment Data: 2013 Part D Prescriber:** This data set contains line items of prescriptions doctors prescribed. It has 23,650,520 prescription types and 3,500 drug types.

2. Preprocessing and Feature Extraction

The feature extraction process had to go through extensive mining of large dataset to compose vectors that can provide meaningful input to learning algorithms.

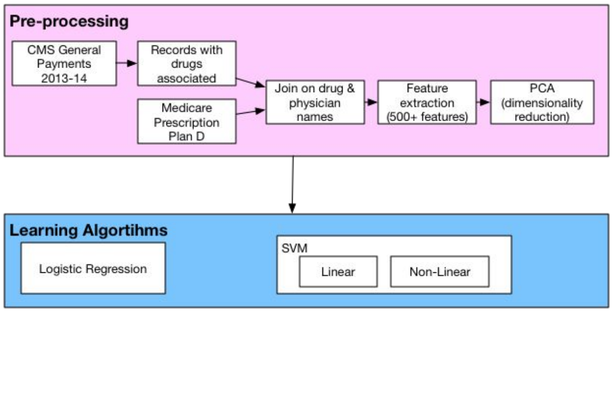
We first filtered out information that would not be useful. We started with filtering out records of payments that were made to an institution and not a physician. We also filtered out records where the payment did not involve any activity related to a drug manufactured by the company.

Next, we had to join the dataset between the payments and the prescription files. Initial attempts to join based on physician's profile did not work as the identifiers are different. After some review of the data, we joined the datasets using the drug name match. The difficulty is that there could be 0 to 5 drugs associated with a payment and this could match either the drug name or the generic name in the prescription record. Carefully matching these conditions, we’ve established two cases as follows:

1. A physician was paid by the company and also prescribed the same drug as was involved in the payment. This is represented by label=1.
2. A physician was paid by the company and did not prescribe the drug (in entire prescription data) that was involved in the payment. This is represented by label=0.

We initially started with Physician specialty vectors included in our feature set, about 540 of them. During the feature analysis stage, we discovered that these were having adverse effect in generalizing the problem, so we remove them. Essentially, it means that non-medical payment influence is not significantly dependent on the specialty of the physician.

The overall application flow can be depicted as follows:


Pre-processing
CMS General Payments 2013-14
Records with drugs associated
Medicare Prescription Plan D
Join on drug & physician names
Feature extraction (500+ features)
PCA (dimensionality reduction)
Learning Algortihms
Logistic Regression
SVM
Linear
Non-Linear

The features we have extracted for the learning algorithm are as follows:

Columns	Data
1	Same "business" state as company - If the company making the payment is in the same state as the physician is practicing in, its 1 else 0.
2-6	Same "license" state as company - there are 5 license states for a physician in data and this represents if the physician has license in the state which is same as the company making the payment.
6-12	Paid amount - we have categorized the payments in the range of >$100, >$500, >$1,000, >$10,000, >$25,000, >$50,000. If a physician received payment in excess of $50,000 all columns will be 1 for that physician.
13-16	Drug claim-count (per prescription)- the ranges captured here are >=100, >250, >500, >750 and the values are applied same as in the case of paid amount.
17-21	Drug day supply (per prescription)- the range captured here are >50, >100, >500, >1000, >1500 and the values are applied same as in the case of paid amount.
22-26	Drug cost (per prescription) - The range captured here are >$100, >$1000, >$10000, >$25000, >$50000 and the values are applied same as in the case of paid amount.
27	Physician ownership indicator - This is an indicator denoting if physician has some interest in ownership rights with respect to the payment being made.
28-42	Nature of Payment - There are 14 different types of payment categories (like food & beverages, travel, conference or consulting fee etc). For each payment transaction, we have captured its types and assign 0 or 1 value to it in the feature vector.

3. Related Work

ProPublica[3] has attempted to investigate many aspects of publicly released data files by Medicare. However, the investigation has mostly been on the reporting perspective and not so predictive aspects.

In [4] and related areas, work has focused on a particular drug prescription based on patient symptoms and other drug facts.

4. Technology

We have utilized a 5-node Spark Cluster in AWS to perform feature extraction over the large dataset. We performed PCA using the dataset using Spark APIs. We have used Weka for classifier algorithms and used Matlab for classifier algorithms and to generate plots for bias and variance analysis.

5. Classification Methods

For this project, we are using two machine learning algorithms for evaluation. The first is logistic regression and the second one is support vector machine. These two methods will briefly be discussed in this section.

Logistic regression

We chose to use this because in the data we saw category outcome variable which violates the assumption of linearity in normal regression. The dependent variable in our datasets is limited (0 not prescribe / 1 prescribe), and logistic regression is a type of analysis where dependent variable is dummy variable.

The formula for binary logistic regression: $P(Y=1| X= x) = 1 / (1 + e^{-(w0 + w1*x1 + w2*x2 + ...)}) = 1 / (1 + e^{-z})$

Log likelihood is: $l(w) = \Sigma (y_i * \log P(x_i; w) + (1-y_i) * \log(1-P(x_i,w)))$

Using gradient ascent method can achieve the Maximum Likelihood Estimation.

Support Vection Machine

SVM is a state of the art classification algorithm. It is explicitly based on theoretical model of learning and has a lot of advantages, like not affected by local minima, does not suffer from the curse of dimensionality. It maximize the margin around the separating hyperplane, and it only specified by a subset of training samples, the support vectors. By adding the regularization term, it is quite tolerant to errors in prediction.

Optimization for SVM is given as, $min\ ||w||^2 + C * \Sigma(\&)$ subject to, $y_i * (w\ t * X + b) >= 1 - \&$

Part C Results and Analysis

We are using Weka to train the dataset and generate the learning models. We have a total of 77,599 examples in the dataset and split it into a training set with 55699 examples (71.78%) and a test set with 21900 examples (28.22%). We also used Matlab to generate the bias/variance plots for data and algorithm analysis. Using our extracted features set and the two classifier algorithms, we trained the data to get the following hypothesis performance.

Logistic Regression Analysis

It took 172.99 seconds to training the Logistic Regression model on the full training set. The model correctly classified 18845 test examples, which is 86.0502%, and incorrectly classified 3055 examples, which is 13.9498%.

Detailed Accuracy By Class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
0	0.766	0.05	0.936	0.766	0.843	0.898
1	0.95	0.234	0.811	0.95	0.875	0.898
Weighted Avg.	0.861	0.144	0.872	0.861	0.859	0.898

Confusion Matrix

	Predict Negative	Predict Positive
Actual Negative	8176	2492
Actual Positive	563	10669

SVM Analysis

It took 172.99 seconds to training the SVM classifier model on the full training set. The model correctly classified 18736 test examples, which is 85.5525%, and incorrectly classified 3164 examples, which is 14.4475%.

Detailed Accuracy By Class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
0	0.736	0.031	0.958	0.736	0.832	0.853
1	0.969	0.264	0.794	0.969	0.873	0.853
Weighted Avg.	0.856	0.15	0.874	0.856	0.853	0.853

Confusion Matrix

	Predict Negative	Predict Positive
Actual Negative	7852	2816
Actual Positive	348	10884

Bias Variance Analysis:

Using SVM with a linear kernel achieved a 14% error rate. However, we wanted to see if we could make predictions with better accuracy. Based on our bias/variance analysis, we observed we had a bias problem as depicted in Fig-1, and we attempted a set of approaches to reduce the error rate.

To address the bias issue, we tried

1. Better modeling the dataset with a more complex model like a non-linear kernel to encapsulate the interactions between the features.
2. Collecting more attributes for the dataset that would be more indicative of the likelihood of prescriptions.
3. Improving the features set by removing noise by using PCA.

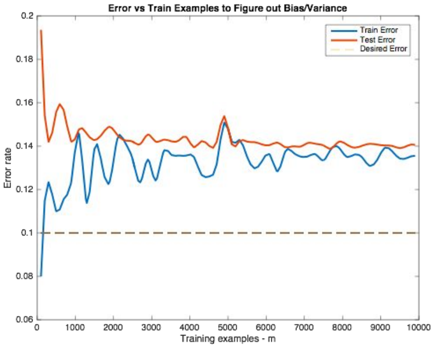


Fig-1 Bias-Variance analysis for Linear Kernel

Non-linear Kernel

Although, the train set was better modeled with convergence of 12% error, using a nonlinear kernel did not improve the generalization/test error, still at ~14%. On analysis, we found it having variance problem and the model is overfitting the training set.

Detailed Accuracy By Class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
0	0.745	0.033	0.956	0.745	0.837	0.856
1	0.967	0.255	0.8	0.967	0.876	0.856
Weighted Avg.	0.859	0.147	0.876	0.859	0.857	0.856

Confusion Matrix

	Predict Negative	Predict Positive
Actual Negative	7952	2716
Actual Positive	370	10862

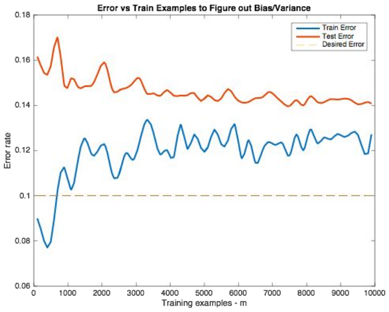


Fig-2 Bias-Variance analysis for Polynomial Kernel of Degree 2

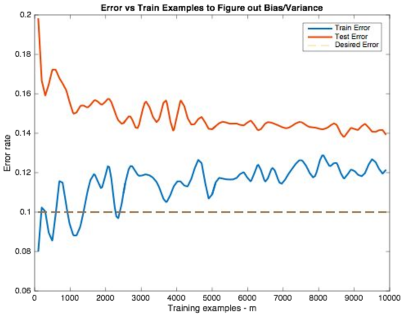


Fig-3 Bias-Variance analysis for Polynomial Kernel of Degree 3

PCA Analysis

We used PCA and obtained the most meaningful 68 attributes from the feature set and removed the other noisy features. The 500+ column physician specialty features were found to be redundant and negatively impacting the hypothesis. After the PCA analysis, we ran SVM using a linear kernel. Removing these resulted in improvement in reducing the test error by ~1% (from 14.4475% to 13.4715 %).

Detailed Accuracy By Class

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
0	0.761	0.035	0.955	0.761	0.847	0.863
1	0.965	0.239	0.808	0.965	0.88	0.863
Weighted Avg.	0.865	0.139	0.88	0.865	0.864	0.863

Confusion Matrix

	Predict Negative	Predict Positive
Actual Negative	8145	2562
Actual Positive	388	10803

Part D Conclusion and Future Work

Dealing with two such large datasets, there was significant pre-processing and feature extraction. We originally had many features, but not all the features were useful. Using PCA, we removed some noisy features.

To test our hypothesis, we used two classification algorithms, logistic regression and SVM. These two algorithms achieved similar results in our experiments using a linear model, although SVM takes more computation time. For further analysis, we used the SVM model.

The experimental results indicate the our original hypothesis holds. Non-medical payments seemingly have an influence in the physician community, especially based on geo-location information. However, with a 14% error rate, it is difficult to definitively say this.

In order to gain better prediction quality and to mitigate our bias issue, we will need to collect more features possibly in the form of detailed drug facts and patient symptom records. We believe physicians tend to recommend well-established and known drugs more often. Also, we think patient symptoms and preferences can play a significant role in prescription decision. These facts are currently not represented in our dataset.

Finally, based on our bias/variance analysis, our models represents the current dataset well and using a more complex model will cause overfitting for training data.

Bibliography

1. https://openpaymentsdata.cms.gov/
2. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html
3. http://projects.propublica.org/checkup/
4. DOI: 10.1007/978-3-642-15431-7_16 Conference: Artificial Intelligence: Methodology, Systems, and Applications, 14th International Conference, AIMSA 2010, Varna, Bulgaria, September 8-10. 2010. Proceedings