Proceedings of Machine Learning Research 71:114–120, 2017 KDD 2017: Workshop on Anomaly Detection in Finance

Deep Learning to Detect Medical Treatment Fraud

Daniel Lasaga DLASAGA@DELOITTE.COM
and
Prakash Santhana PSANTHANA@DELOITTE.COM
Deloitte

Abstract

Excessive treatment or testing of patients is considered one of the most ubiquitous and persistent forms of waste and abuse in healthcare. Some estimates show excessive treatment to be as high as 8% of all medical insurance provider expenditures. It is very difficult to identify an extraneous or unnecessary procedure or drug because there is such a wide variety of diagnoses and an equally large number of treatment options.

Our goal in this paper was to show how RBMs can be utilized effectively to ferret out abnormal treatments where the prescribed treatment for a given diagnosis is not strictly followed. To test our hypothesis we generated 200,000 different injuries and injected 10% of the injuries with unnecessary treatments to reflect estimated industry prevalence levels. Using testing and training sets we found that Restricted Boltzmann Machines (RBMs) were able to reach AUCs of .95, lifts at 9.5 and recalls at 50%. Implementing our approach on real-world client datasets have shown performances levels that approach simulation performances despite additional noise.

1. Introduction: Overutilization in healthcare

Waste and abuse from overtreatment (AKA overutilization) by medical providers is pervasive throughout the United States and many other nations. It persists despite concerted efforts by governments and insurance providers to catch this type of abuse. A report from the Institute of Medicine in 2010 estimated up to 30% of health insurance expenditures were due to waste and abuse. Overtreatment accounts for over a quarter of waste and abuse or 8% of total U.S. health expenditures.(Kliff, 2012)

Part of the problem in mitigating overutilization is the difficulty in parsing valid treatments from invalid/unnecessary treatments. The variety of diagnosis and potential drugs and treatments to address them create innumerable permutations and combinations. Additionally the validation of prescribed treatments for any given diagnosis require time and specialized medical knowledge from trained professionals. We offer a machine learning approach using Restricted Boltzmann Machines (RBMs) to model likely combinations of treatments and diagnosis enabling us to better highlight unlikely treatments where overutilization is most likely.

Our work in this paper with medical overutilization is motivated by a client engagement specifically related to occupational insurance where the client coverage only includes occupational injuries. Simulations of similar injury data sets revealed an ability for the RBMs to gain lift of 9 to 10 times given sample sizes of 200,000 with 10% fraud.

2. Data sources and data simulation

Due to underdevelopment of data collections from our motivating client, itemized treatments and diagnosis were predominantly in the form of transcribed descriptions. Lack of officially coded medical data necessitated the processing of descriptions from unstructured data in to structured pseudo medical codes so that we could conduct further modeling. We used common NLP techniques to prepare diagnosis and treatment descriptions by tokenizing, stemming, and removing stop words and common medical terms. To assign drug codes we conducted cosine similarity between the word frequency of described drug treatments and the published list of USDA drugs. On the remaining treatment and diagnosis descriptions we used a combination of topic modeling and hierarchical clustering to group similar treatments and similar diagnosis.

Ultimately the goal was a concise 800 by 200,000 matrix of treatment occurrences for each injury. Structuring the statements, diagnosis, drug prescriptions and treatments into codes allowed us to then transform the list of injuries in to an occurrence matrix of rows and columns. Each row was an injury and each column designated an injury claim was associated to assorted classes of diagnosis or treatments. An injury could contain multiple diagnoses and treatments.

In order to test the feasibility of capturing overutilization we simulated a dataset of 200,000 injuries and injected 10% of the injuries with fraud.[1] The simulation process started by defining 400 by 800 probability matrix. Each of 400 distinct diagnoses mapped 2 to 20 out of 800 possible treatments as high probability occurrences. Injuries with a particular diagnosis were not guaranteed to have that diagnosis treatment set but would likely select from the high probability mapped to that diagnosis. Likewise a given injury diagnosis might occasionally assign as a low probability treatment. Injuries were then simulated by picking a random diagnosis and probabilistically selecting a combination of mapped treatments. Each injury had a 10% chance of being selected for overutilization in which case a random extra treatment is added to the injury. The result was a 800 by 200,000 matrix of injuries with treatments as features and injuries as observations. All subsequent results were derived from training and testing on the simulated data. In order to stress test results we also ran a simulation of data with only 1% of injuries containing added treatments.

- Total injuries: 200,000
- Diagnosis classes: 400
- Treatment classes: 800
- Fraud simulated: 10%

3. Methodology

Finding treatment fraud can be framed as a categorical outlier problem. Diagnosis and treatments can be thought of as different categories. There are many situations in which one category often implies another: a bone fracture diagnosis often requires an x-ray treatment. There are also categories that rarely happen together. For example, a bone fracture diagnosis rarely coincides with a pupil dilation treatment. If we can effectively model the

[1] While certain estimates of total impact from abuse and fraud are higher, we chose a more conservative number of 10% and also experimented with 1% injection of fraud.

likelihood of different sets of categories in our categorical system (treatments and diagnosis), then we should be able to find injuries that stray from expectation.

There are a number of methods to mine categorical systems for outliers, such as Logistic regression. However, given the dimensionality of our matrices and the sparsity of data we found that this approach had difficulty converging. The logistic regression approach required training and maintenance of a separate model for each category. When working with high dimensional categorical variables this could result in management of hundreds or thousands of models.

We also investigated associated rule mining as a technique for categorical outlier detection.(Preetha and Radha, 2012) While associated rule mining can effectively find combinations of categories with good support it is not effective when data is sparse and general support levels are low. We found that most combinations we observed in our data did not offer high enough levels of support sufficient for us to differentiate categories that belonged together from those that did not.

RBMs are a limited graph model (or neural network) which has only one input layer and one hidden layer and no output. Unlike traditional neural networks they do not train based on a target, but instead try to converge based solely on input data – an unsupervised neural network. A further advantage of the RBM based approach is that it leverages Gibbs sampling which allows it to traverse the relatively high dimensionality of our data set to explore for all the potential interrelationships between the categories.(Hinton, 2002) RBMs are often compared to autoencoders in that they both have similar abilities to act as nonlinear dimensionality reducers. However, autoencoders typically use a deterministic back propagation approach whereas RBMs use Gibbs sampling which is a stochastic approach.

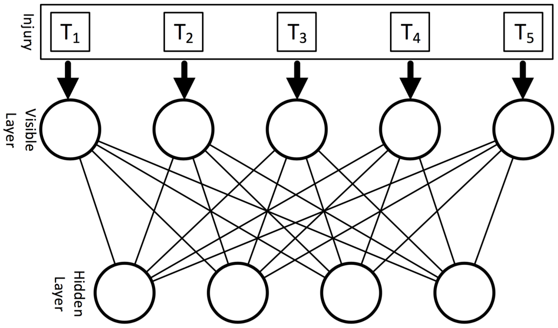


Figure 1: By definition an RBM is a two layered fully connected neural network. Above an Injury with 5 features train an RBM model of one visible and one hidden layer.

The resulting RBM model contains neuron weights that estimate the interrelationships of different features from the training data. We can then use the trained RBM to predict likely combinations of categories for new input data. If the trained RBM's prediction of categories matches the actual input data then the difference between the actual and predicted treatments will be small. If the actual input data contains unlikely combinations then the predicted treatments will not match and the difference between actual and predicted will be larger. Large differences represent high risk for treatment fraud small differences represent low risk.

We trained RBMs on an occurrence matrix of injuries where each column is a treatment. The resulting RBM model re-estimated the occurrence of given treatments. We compared the predicted treatments from the RBM with the actual treatments billed. We rolled up injury level risk using mean squared error (MSE):

$$injury\ risk = \frac{\sum_t^T (y_t - \hat{y}_t)^2}{T}$$

Where T is the total number of treatment categories, y_t is the actual occurrence of a treatment category, and $\hat{y}_t$ is the predicted occurrence of a treatment category. Using these metrics we could then estimate risk of overutilization and make suggestions to the medical team for further investigation into an injury claim.

3.1. Validation

We took a stratified sample of data and submitted the results to the internal medical review team. The review team was then required to make blind evaluations of whether an injury was at high or low risk for fraud. The resulting sample was then used as a test set. For validation of model results on simulated data we created a 90/10 training/test split. Following results are based off of the simulated 10% test set from the 200K generated injuries.

	x-ray	dilation	CAT-scan	Antibiotics
Injury 1	0	0	1	0
Injury 2	1	0	0	0
Injury 3	0	1	0	1
Injury 4	0	0	1	0

Table 1: Treatment occurrence matrix

3.2. Picking hyper-parameters

Modeling was executed in R using the deepnet library.(Rong, 2015) Restricted Boltzmann Machines offers a number of potential hyper parameters to tune the RBM. Deepnet uses a sigmoid activation function for both visible and hidden layers. We used the default batch size (100) and learning rate (0.8) and focused our initial tuning efforts on finding the optimal number of hidden layer nodes and the number of iterations per sample. In general smaller numbers of hidden neurons are faster to compute and offer dimensionality reduction while higher numbers of hidden neurons allow greater ability to find nonlinear relationships between input features but risk overfitting. Epoch describes the total number of times the RBM iterated through weight estimations.

In order to find optimal parameters we iterated through different combinations of hidden layer and epoch combinations. For each combination we used the business validation results to calculate the Response Operator Characteristic curve and the area under the curve (AUC) as an approximation of relative model power. Using our generated data the calculation of the AUC leverages our a priori knowledge of which injuries were injected with fraud.[2] We then estimated the parameter combination that would result in the highest probable AUC for both mean over treatment and mean squared error. Poor performance near low values hidden nodes shows possible under-fitting.

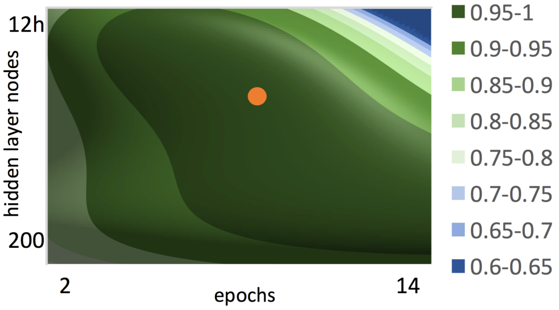


Figure 2: Mapped parameters with final selected hyper parameter at 8 epochs and 1000 hidden layer nodes.

[2] In practice, calculation of AUC on live data requires investigation and verification by users of fraud or not fraud on a stratified sampled data.

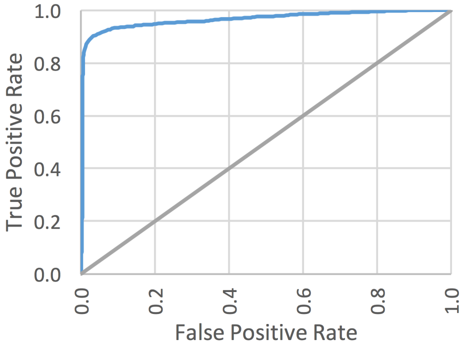
We expected a more degradation at high hidden node counts but instead found increasing epochs more quickly led to overfitting. The optimal combination of these hyper parameters was estimated to be 1000 hidden nodes and 8 epochs using the mean squared error calculation [see figure 2]. Subsequent results analysis is reported on this model version.

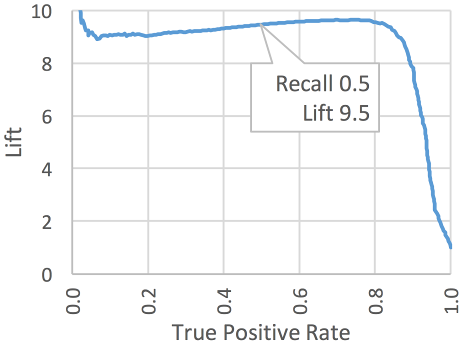
4. Results

The optimal hyper parameters resulted in an AUC of 0.96 (using the MSE calculation) and lift ratios of 9.5 at recalls of 50%.

Figure 3: Response operator characteristic curves from mean squared error risk score

Figure 4: Lift compared to recall for RBM mean squared error risk score





To understand how sample size affects the performance we generated a learning curve [see figure 5]. For benchmarking purposes, we used a smaller number of hidden layer of 300 nodes and 9 epochs. It shows that under differing circumstances that the RBMs stabilize at similar rates. Our main simulation set with 10% fraud and another set with 1% fraud both tend to stabilize with sample sizes above 40,000. This shows the RBMs can be more sensitive to smaller sample sizes, and less sensitive to variation in the prevalence of fraud. In real data with more noise the performance requires more observations for similar dimensionalities.

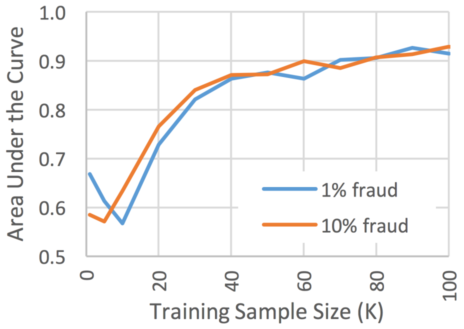


Figure 5: Test set based learning curve from our simulation of 200K injuries with a 10% fraud rate and 100K injuries with a 1% fraud rate

We calculated the risk scores for the population by taking the mean squared error found for each injury in the test set. Because our data set had 800 dimensions and most treatments for a given injury are zero (not present), the mean squared errors range from near 0 to 0.003. For display purposes we scaled the test set scores from zero to one and power transformed the scores to have a median of 0.5 resulting in the following equation:

$$rebalanced\_score = \left(\frac{mse - \min(mse)}{\max(mse) \;\; - \min(mse)} \right)^{\frac{1}{9.85}}$$

We can see the distribution of scores in figure 6 with injuries generated with fraud distinguished from those with no fraud. We can see clearly that fraud injuries are represented among the top scores. Using this score we can effectively isolate most of the injuries where we injected extra unnecessary treatments.

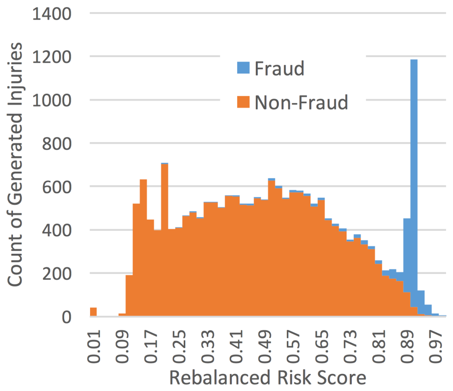


Figure 6: Distribution of fraud and non-fraud risk scores from the test set after rebalancing scores to a median of 0.5.

5. Conclusions

Real world data is generally not as clean as simulation data. However, we've seen the RBM performs well with lifts in the range of 2 to 4 for recalls of 50%. Real client data has shown more sensitivity to sample size than simulated data with large improvements from sample increases. This is likely due to greater levels of noise in real data. Mandating the use of standardized medical coding like ICD10 into the invoicing would reduce noise in the formation of the injury occurrence matrix. In lieu of coded data, matching claim descriptions against integrated diagnosis dictionaries and procedure dictionaries may offer better medical coding than purely unsupervised methods.

Though we do not suggest RBMs are the only method in this space, (Bayesian networks have been used extensively) RBMs have proven to have utility in identifying treatment overutilization. These results show RBMs can effectively identify high risk outliers despite the noisiness of our data and the dimensionality of the feature space. We recommend consideration of RBMs not just within insurance fraud, but in any situation where we are faced with high dimensional categorical outlier detection.[3]

References

Geoffrey Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1774 – 1776, 2002.

Sarah Kliff. We spend $750 billion on unnecessary health care. two charts explain why. *Washington Post*, September 2012.

[3] This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Preetha and Radha. Enhanced outlier detection method using association rule mining technique. *International Journal of Computer Applications*, 42(7), March 2012.

Xiao Rong. *Deep Learning Toolkit in R*. CRAN-R, February 2015.