33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Using Support Vector Machines to Detect Medical Fraud and Abuse

Charles Francis, Noah Pepper, Homer Strong

Abstract— **This paper examines the architecture and efficacy of Quash, an automated medical bill processing system capable of bill routing and abuse detection. Quash is designed to be used in conjunction with human auditors and a standard bill review software platform to provide a complete cost containment solution for medical claims. The primary contribution of Quash is to provide a real world speed up for medical fraud detection experts in their work. There will be a discussion of implementation details and preliminary experimental results. In this paper we are entirely focused on medical data and billing patterns that occur within the United States, though these results should be applicable to any financial transaction environment in which structured coding data can be mined.**

I. Introduction

With health costs rising across the United States and an abundance of codified time-series data available, medical bill review is amenable to the labors of data mining and analytics. Examining the state of the field, one finds that most companies rely on rigid rules-based systems and manual audits to identify fraud, waste and abuse.

This is an ineffective solution to the problem of detecting fraud: rules-based systems do not handle the medical domain well. In billing, the level of complexity and frequency of exceptions make the creation and maintenance of exhaustive rule-sets a terribly difficult and labor intensive task. Consider the geographic and temporal specificity of medical norms and the combinatorially large space of possible diagnosis and treatment combinations.

Employing humans to review bills introduces new problems. The time intensive nature of the work and necessary training means that scaling human audits is not economically viable. This inability to scale is exacerbated when most audited bills being are legitimate and thus produce little to no savings upon audit.

The focus of this paper is detailing and discussing a real-world integration of machine learning tools with a labor intensive knowledge task. The research and development of the engine discussed here is ongoing as we continually strive to increase the efficiency with which we utilize our human resources.

A. Medical Domain

Extreme data sanitation problems, the complexity of regulatory environments, and coding systems together create significant opportunities for error, abuse, and fraud.

Qmedtrix, Portland, OR 97214
francisc@reed.edu
peppernm@gmail.com
homer.strong@gmail.com

Medical bill review is performed by highly skilled workers who take a minimum of two to three years to train. For some tasks a medical degree and medical experience are prerequisites. Due to the limited rate at which an auditor can process bills, an expert system to augment manual reviews is highly attractive.

We consider any billing mistake on a medical claim an 'error', and patterns of these errors resulting in excess payments to providers are 'abuse'. The systematic and intentional abuse of medical billing procedures by individual providers is 'fraud'. Estimates of the losses from fraud and abuse in the United States range from $75 to $250 billion per year [1], [2]. While this amount is difficult to estimate with certainty, this is at minimum a very expensive problem.

The persistence of rampant abuse reflects the complexity of the practices and laws surrounding medical billing. While there are many potential applications for machine learning and automation in the health care cost containment industry, ultimately much of the work must be performed manually.

Prompt payment laws and other industry-specific regulations limit the time analysts have to detect error and fraud. This presents a non-trivial problem for companies that primarily rely on a manual process that depends on a large body of domain-experts to audit the most suspicious bills: given the high cost of training new experts, what to do when the volume of bills is too great for current staffing levels?

Ideally the entire process could be automated and executed by computers. A wholly automated system remains a long-term problem, given sparsity of reliable training data for crucial tasks performed by trained medical professionals. For example, the task of determining whether and why a bill is an example of abuse, let alone fraud, remains a difficult problem despite advances in the field.

The immediately crucial issue thus becomes how to leverage computer systems to greatly improve the efficiency of already existing domain-experts, ensuring that their time is spent only on tasks directly relevant to their expertise that have high likelihoods of producing reductions for a client and that subsequently generate valuable data for future attempts at automation. A central constraint in our experimentation has been engineering a solution that not only allows us to scale the valuable time of our domain experts but that can itself quickly adapt to a massive volume of data.

While overall accuracy is an important metric, we are especially concerned with minimizing false negatives. Quash is intended to be used in situations where the volume of potential claims to audit is greater than the human resources available. For example, in a training sample of several hundred thousand bills from inpatient and outpatient hospitals,

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

$\sim$22% were found to be abusive or require further review.

To go through the entirety of this collection, looking for that rare bill with great savings is impractical. The challenge is to automatically eliminate as many bills as possible, while remaining confident that we are not throwing out bills with potential savings. A false negative will likely never be examined, while false positives are tolerable. Hence a central priority in engineering Quash is that false negatives must be minimized before maximizing overall accuracy.

B. Related Work

Related work has been done in [3] and [4] but they differ from our own research both in terms of the origin of data and the learning problem under consideration. We focus on workers compensation bills, which are often audited more aggressively than group health bills.[1] Workers compensation, like Medicare and Medicaid, often has more stringent rules about proper payment. Many of these rules are enforceable by payers. This means that there is likely more actionable overpayment in our set of medical bills than in the other researchers bills. Additionally the distribution and nature of such overpayment is likely quite different.

The work in [4] discusses finding overpayment in medical bills. They focus on price fluctuation and anomaly detection, while we are interested in discovering a broader category of abuse. In particular, we are especially interested in types of abuse that involve the misuse of codes.

For example, unbundling is a type of abuse in which a procedure that is supposed to be charged as a unit is broken down into its component procedures which can often increase the amount that can be charged for a bill. Such abuse is difficult to detect except by a trained medical expert, since the amounts charged will perfectly match the billed procedures and the relationships between procedural codes is opaque to the untrained eye.

In [3] Kumar et al. are also specifically searching for medical payment errors instead of fraud and abuse. While they include overpayments as a type of error, they are not creating a system restricted to finding abuse and fraud nor do they differentiate the nature of the error or type of bill in their engine.

Furthermore, they claim that errors exist in 2-5% of medical claims. In our experience the abuse rate is as much as 5 times higher, counting only overpayment errors. We believe there are two reasons for such a discrepancy between our data. First, this could be accounted for by the aforementioned differences in our domain of interest (group health versus workers compensation). Second, their data came directly from two large insurance carriers and can only be acted upon by the latter groups. Large carriers, especially in the group market, typically do not focus on rooting out fraud and abuse for fear of litigation

[1]Group health bills represent the vast majority of medical claims, but because the patient is not indemnified in group health, payers are less likely to challenge suspicious billing than in workers compensation.

II. Infrastructure

The two most crucial infrastructure problems that we have addressed are coping with the uncleanliness of received data and scaling. The visualization and the user interface to our system are also crucial concerns but are outside the scope of this paper. For examples of our research in this domain consult [5].

A. Data Cleansing

Medical bills in the United States are required to be encoded according to various standards: notably ICD9[2] for diagnoses and CPT[3] for procedures. While standards promise ease of processing across billing platforms, this convenience is mitigated by the disparate software and human solutions employed by billing specialists for internal storage and analysis of bills. Such tools often have varying standards for formatting bills. Some examples: are decimals to be included in ICD9s or not? Are trailing 0s to be included? While not overly problematic for human analysis, data must be standardized to be useful for machine learning tools.

In our data management infrastructure we have a robust set of automated tools to flag and often correct coding errors. This allows our machine learning engine to be used with a focus on abuse and fraud, leaving the more trivial coding errors to a separate system. There are a significant number of bills with coding errors that can not be automatically corrected, which must be fixed by human auditors.

B. Scalability

Architecturally, we have been concerned with our systems ability to handle large loads of bills. While this is not the focus of the paper, we feel that a brief account of our experience will be of use to other teams pursuing a similar path. Central in the architecture of this system was the decision to engineer it as a sequence of decoupled, distributed agents atop a message queue.

In particular, our data-flow is a stream with occasional bursts, where it is most essential that we minimize the average time necessary to hand actionable bills for humans. Given that each individual task can be rapidly executed in a sequential manner, we recognized that the necessary efficiency could best be obtained by running multiple instances of each task in parallel instead of parallelizing the execution of any individual process.

This allows us to most efficiently handle our day-to-day load, while providing us with the capacity to greatly increase the number of running processes, and thereby our capacity, when presented with a large set of data from a client.

III. Learning

With bills standardized and elementary errors resolved, we utilize our data in the resolution learning problems we face. The primary issues which we address are the assignment of adjudication types to bills and the detection of abuse.

[2]International Classification of Diseases version 9

[3]Current Procedural Terminology

Predicted:	ASC	PRO	AMB	IPH	OPH	DME	ER
ASC	99.5%	.27%	0%	0%	.23%	0%	0%
PRO	.36%	99.21%	0%	0%	.33%	.1%	0%
AMB	0%	0%	99%	0%	0%	0%	0%
IPH	1.5%	0%	0%	98%	0%	0%	.5%
OPH	.25%	.5%	0%	.25%	98%	0%	1%
DME	0%	.15%	0%	0%	.15%	99.7%	0%
ER	0%	0%	0%	0%	2%	0%	98%

TABLE I
ADJUDICATION CLASSIFICATION CONFUSION MATRIX.

For our learning tasks we use the liblinear implementation of linear SVM [6]. Linear SVM is well regarded for scaling of both number of features and observations.

We have experimented with several other learning algorithms and received weaker results, in terms of training and classification run-time and also with respect to accuracy.

A. Encoding

Many critical features are medical codes. As features, we represent the codes with a sparse binary encoding. The motivation for this choice is that no assumptions can be safely made about which codes can co-occur. Since any combination of medical codes might appear on a bill, a sparse encoding is the natural choice for a large vector space in which to embed a bill.

The binary features are indicators for the occurrences of a code in a bill. Binary parity allows both training and testing speed to be inelastic under the addition of further binary features. There are about 75,000 sparse features. For example, ICD9 codes account for around 8,500 of these binary features.

B. Feature Selection

We have determined a set of features that have provided optimal performance in our experiments. Our features were initially chosen via consultation with bill reviewers, who recommended the most indicative features. From this starting point, we refined the feature set through a series of experiments.

The features we have determined to be most performant are:

1) occurrences of codes: whether ICD9, CPT, HCPC, or modifier codes appear in a bill
2) the Tax Identification Number (henceforth TIN) of the billing hospital
3) the number of lines in the bill (encoded categorically)
4) the number of days of service (encoded categorically)
5) the total duration of the bill: how many days between the first and last date of service, including days when no services were being performed (encoded categorically)

C. Adjudication Type

In bill review, an essential first step is to accurately determine the type of facility from which the bill originated. The type of facility is called the adjudication type. Depending on whether a bill is from an inpatient hospital, an ambulance, an emergency room, etc, there are different kinds of expected or possible procedures and practices and different acceptable payment practices. Before analysis can be performed on a bill, its adjudication type must be determined so that an analyst can reason about reasonable payments. Reflecting on the state of medical billing practices, many bills report their adjudication type erroneously or not at all.

For years, Qmedtrix has used a rules-based classifier in production which has generated a tremendous number of labeled bills. This system is problematic, as it relies on a set of simplified, human- discovered indicators of adjudication type. Many bills cannot be classified and must be handled by humans, causing delays and taking up valuable auditor time. Even then, the accuracy of this system is unreliable, around $\sim$85%. Hence, even though we have a large set of labeled data, it is in general unreliable.

Note that we consider an additional categorical variable: the presence of keywords (e.g. hospital, ambulance) in the name of the hospital. The name of the facility is often missing or abbreviated, so while this feature is helpful, contributing an average additional 1% accuracy, it is no panacea. It is not possible to use only hospital names because there is not a bijection between hospitals and adjudications.

For this task we trained using 120,000 of our $182,809$ sample set of bills and tested using the rest. Examining table I note that we are presently obtaining, on average, 99% accuracy with a minimum of 98% across all of the adjudication types.

D. Abuse Detection

Traditionally, auditors process bills sequentially, quickly scanning for discrepancies or reviewing reductions made by rules-based pricing engines. The former manual process is too slow, while the latter rules-based system only catches abuse from known errors, and is inadequate when dealing with a rapidly changing landscape of abuse.

Predicted:	Abuse	Not Abuse
Abuse	72%	28%
Not Abuse	7%	93%

TABLE II
ABUSE CLASSIFICATION CONFUSION MATRIX.

Predicted:	Abuse	Not Abuse	On-hold
Abuse	70%	15%	15%
Not Abuse	3%	91%	6%

TABLE III
ABUSE CLASSIFICATION RESULTS WITH HETEROGENEOUS BAGS AND 90% THRESHOLD.

Predicted:	Abuse	Not Abuse	On-hold
Abuse	79%	19%	2%
Not Abuse	5%	94%	1%

TABLE IV
ABUSE CLASSIFICATION RESULTS WITH 50 $2/3$ OF TRAINING SET BAGS AND 90% THRESHOLD.

As both unsatisfactory processes have been ongoing for numerous years with their results validated we possess a reliable set of abusive bills. This learning task constitutes generating a model that predicts whether we would reduce the bill, indicating whether an auditor or rules-based system has in the past found any systematic billing errors which resulted in an excess payment, i.e. abuse.

Working with our validated training set of $182,809$ bills, we have a total of $13,173$ bills that were not determined to be abusive and $51,056$ that were. For building our models we took a sample of $120,000$ bills.

Examining table II, one notes that it is easier to predict impactedness than abuse. Abusive bills are classified with an unacceptably high false positive rate.

While this remains a valuable tool for an auditor, since a preliminary negative result is a reason to move on to a subsequent bill, improvement is necessary.

E. Minimizing False Positives

In working to overcome the high false positive rate indicated in table II, we have found the introduction of bagging schemes useful. Recall that bagging, a.k.a. bootstrap aggregating, is the practice of obtaining classifications via the consultation of many models trained on small sets of data sampled with replacement from a larger training set.

Generally, such a technique is recommended in cases where data is known to be randomly mislabelled, where models trained on smaller sets of data frequently return divergent results.

In this experiment there are two principal parameters: the number of bags and the number of bills per bag. We have made attempts to optimize these parameters and have found two optima, one more accurate and the other minimizing false negatives.

Our first optima is found when bagging fifty models each containing 2/3 of the original training set. Neither lowering the total number of models nor lowering the count of bills per model positively effected our results. With these parameters, our accuracy was highest and we received a significant decrease in our false negative rate (see table IV).

Our second optima was found using a heterogeneous bagging scheme with 20 models each containing $2/3$ of the sample, 20 models containing $1/4$ of the sample, and another 10 models each containing $1/2$ of the sample. With this combination our accuracy decreased slightly, but our false negative rate also went down non-trivially.

Across both bagging schemes we train on the original $182,809$ bills. To determine any given bills classification we use a simple majority vote with optional thresholding on the percent of models that have to agree.

Tables III and IV show the results of our experimentation.

In a production setting we can use heuristics to further cut down on false negatives. This approach is inelegant and has the standard drawbacks associated with rules-based solutions to this problem. There is ongoing research to find better machine learning solutions to further lower the false negative rate.

IV. CONCLUSION

Overall we have been pleased with our results. This satisfaction is especially high with respect to our results in bill adjudication: the level of accuracy obtained is acceptable and the model is already useful in a production environment.

The abuse detection engine is accurate enough to be of non-trivial use to auditors but we believe that there is research to be done. We are currently working with human auditors to understand what data will be important to collect. We plan to build an active learning system for abuse detection that can predict the presence of various types of fraud.

We believe that giving the classifier more detailed information on auditor behavior is an important step towards engineering a more robust automated solution.

As our tools are rolled into production, we find the application of machine learning to improve our efficiency as a company. We are also interested in the exploration of more sophisticated voting and bagging schemes as this has produced the greatest marginal returns to effort.

REFERENCES

[1] S. Rosenbaum, N. Lopez, and S. Stifler, Health insurance fraud: An overview, George Washington University Medical Center School of Public Health and Health Services, Department of Health Policy, Tech. Rep., 2009.

[2] L. Morris, Combating fraud in health care: An essential component of any cost containment strategy, *Health Affairs*, vol. 28, no. 5, pp. 13511356, 2009.

[3] M. Kumar, R. Ghani, and Z.-S. Mei, Data mining to predict and prevent errors in health insurance claims processing, in *KDD 10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.* New York, NY, USA: ACM, 2010, pp. 6574.

[4] A. Anand and D. Khots, A data mining framework for identifying claim overpayments for the health insurance industry. in *INFORMS Workshop on Data Mining and Health Informatics*, 2008.

[5] N. Pepper, H. Strong, and K. Lynagh, Qyz: A platform for visual analysis of error, abuse and fraud in medical bills, in *IEEE VisWeek Workshop on Visual Analytics in Health Care*, 2010.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research*, vol. 9, pp. 1871 1874, 2008.