

Overpayment Detection in Health Insurance Data

Iordan Slavov
Optum
555 Kappock St
Bronx, New York
iordan.slavov@optum.com

ABSTRACT

Overpayment (OP) detection of medical claims is an important facet in the quest to contain the raising medical costs in US. According to National Health Care Anti-Fraud Association and Government/Law Enforcement Agencies 3–10% of all health care spending is lost each year in improper payments (e.g. [1], [2]).

We would like to use outlier detection methods to tackle the above problem but they usually are designed for **continuous** only data. In health care, **categorical** data (such as diagnoses codes) are imperative thus we employ a method mapping the categorical data into two continuous similarity measures.

Two methods, the Robust Multivariate Distance and the density based Local Outlier Factor, are compared on simulated **mixed** (continuous and categorical) data. In real data, the selection of outlier observations from the resulting ranking is not trivial so we suggest using “effect size”. This is illustrated with healthcare insurance data from Optum.

Implementation of this framework in statistical software such as R or SAS is discussed.

Keywords

Outlier Detection, Health Insurance, R, SAS

1. INTRODUCTION

Supervised predictive models can be quite successful in detecting claims with known sources of overpayment. But the behavior of the providers, as well as their contracts, change quite dynamically and could produce unseen before (and thus not appropriate for supervised modeling) reasons for OP. This is where the unsupervised and semi-supervised models are useful. One such class of models falls in the outlier (or anomaly) detection paradigm.

We start the Methods section by describing a transformation which converts the non-continuous variables into continuous. There are other approaches to do so [3] but this one is natural enough and helps us to lay down our framework.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD 2014 New York, New York USA
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

In the following section we compare, on simulated data, two outlier detection methods – the robust Multivariate Distance (rMVD) and the density based Local Outlier Factor (LOF) [7]. Two types of anomalies are illustrated – one cluster with well defined outliers and two clusters with different local density. It is confirmed that the additional continuous variables representing the categorical components in the data do not change the expected by design ability of the two considered methods to detect outliers.

In view of the high computational complexity of most of the outlier methods and large volumes of data, we conclude by discussing how to accelerate detection in real life data.

2. METHODS

2.1 Categorical Variables Mapping Method

The method in [5] can reduce as many categorical variables as needed to a fixed number of “similarity” measures. It requires a reference dataset (class).

We only give an idea how the method works and defer to the original publication for other details. One of the four measures defined in [5], d_m , reflects the intuition that an observation belonging to the same class as the reference class will, on average, have more matching values on its categorical attributes with the reference class than an instance belonging to a different class. For example, let’s consider the reference data set given in Table 1 and compute d_m for the observations x and y :

Table 1: Reference Set (Class)

A	B	C
a_1	b_1	c_1
a_1	b_1	c_2
a_1	b_1	c_3
a_2	b_1	c_4
a_2	b_1	c_5

$$x = (a_1, b_1, c_1) \implies d_m(x) = \frac{3 + 2 + 2 + 1 + 1}{5} = 1.8$$

$$y = (a_3, b_1, c_5) \implies d_m(y) = \frac{1 + 1 + 1 + 1 + 2}{5} = 1.2$$

Observation x is more similar and thus closer to the reference set.

We use two of the four measures studied in [5]. After transforming the categorical variables into the two continuous measures, we compare the two outlier detection methods described next.

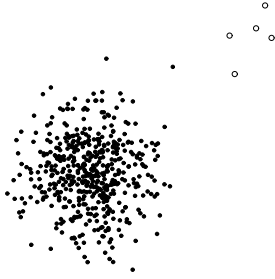


Figure 1: One well-defined cluster and outliers

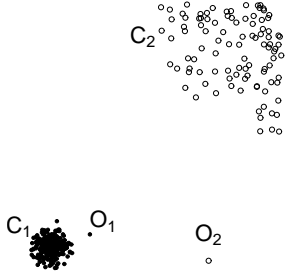


Figure 2: Two clusters with different local density

2.2 Robust Multivariate Distance

The idea to determine outliers as observations falling in the tail of a continuous distribution is generalized by using a robust Multivariate Distance (rMVD) between an observation x and the “center” of a sample:

$$(x - \mu)^T \Sigma^{-1} (x - \mu)$$

Here μ and Σ are robust versions of the mean vector and the sample covariance matrix defined in [6]. In addition to the existing R implementations, we implemented and tested this method as a SAS macro using [4].

We are interested in new outlier claims and a good approach is to compare the claims for the same provider over time. We can consider the claims filed in some initial time period as the “base” and compare them to claims from a later time. For the rMVD method, “effect size” defined as

$$effectsize = \sqrt{(\mu_{new} - \mu_{base})^T \Sigma_{base}^{-1} (\mu_{new} - \mu_{base})}$$

is used to detect providers for which the “center” varies unusually with time if compared to certain “base” sample of claims.

2.3 Local Outlier Factor

The rMVD method is performing well when there is only one large cluster of data as in Figure 1. But rMVD can’t identify outliers O_1 and O_2 in the situation depicted on Figure 2. Much better job in this two cluster situation is done by the LOF method introduced in [7].

Therefore we tested an R implementation (package DMwR described in [8]) on the same synthetic data.

3. NUMERIC EXPERIMENTS

We test the ability of the above outlier detection methods in two scenarios. Two 4-dimensional datasets (with two

continuous variables x_1, x_2 and two categorical a and b) are generated as follows:

Data 1 The (x_1, x_2) components of the $n = 510$ data rows are independent samples from the standard normal distribution $N(0, 1)$. The first 500 observations of the two categorical variables a (with levels a_1, a_2, a_3, a_4) and b (with levels b_1, b_2, b_3) are generated by weighted random sampling with replacement. The weights for the a values are $(0.1, 0.3, 0.3, 0.3)$ and for the b values – $(0.1, 0.5, 0.4)$.

The last 10 values for a and b belong to the **test** set. They include the 2 “normal” values (a_4, b_2) and 8 anomaly values which use the new levels a_5 and b_4 : (a) 3 pairs (a_5, b_2) ; (b) 3 pairs (a_4, b_4) and (c) 2 pairs (a_5, b_4) .

Data 2 The second set resembles the set from Figure 2. It has $n = 252$ observations with the first 150 generated as above – the (x_1, x_2) are coming from uncorrelated standard normal distributions. The next 100 are drawn from a uniform distribution on a circle segment (the lower left quarter of the circle). The categorical variables are with 4 and 3 levels and probability weights for the simulated levels $(0.2, 0.3, 0.3, 0.2)$ and $c(0.3, 0.35, 0.35)$, correspondingly.

Only 2 clear outliers (the last 2 of all 252 observations) are introduced and they are outliers *only* in the (x_1, x_2) plane.

The results of applying our approach in terms of the proportion of known outliers in data identified by the corresponding method are summarized in Table 2. In the first line, the (non-robust) MVD method results are listed for comparison.

There are 8 known (forced in) outliers in the first scenario (“Data 1”) so that 0.75 for rMVD means 6/8 outliers were successfully detected. In the LOF algorithm we used a range

Table 2: Methods Performance

Method	Data 1	Data 2
MVD	0.625	0.5
rMVD	0.75	0
LOF	0.25–0.325	0.5–1

of values for the number of neighbors k . Thus we report in the last line the range of the correctly identified proportions for the k range 1–15. LOF detects only 2 – 3 outliers but this is not surprising given that only the two instances with values (a_5, b_4) for the last two coordinates (those in case (c) for Data 1) are truly “novel”.

There are only 2 known outliers in “Data 2” which LOF, in contrast with rMVD, captures correctly.

4. REAL DATA

The rMVD method was applied to a large set of Optum medical claims with the following characteristics

- A mix of variables was used for outlier detection
 - 10 continuous characteristics – claim amount paid, claim amount charged, length of service, etc.

- 5 categorical variables – demographics, diagnosis code, procedure performed code, etc.

- The claims for a given provider were fewer than 1000 at each time point.

For outlier selection we use the effect size introduced above to first narrow down our search by selecting relatively homogeneous groups of observations (claims belonging to the same provider).

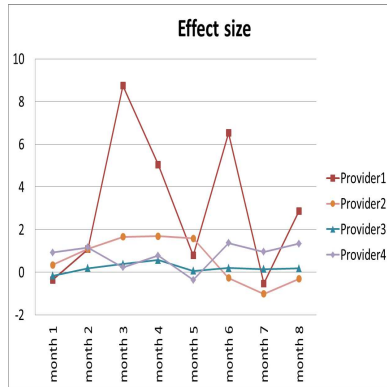


Figure 3: Outliers at the provider level

In Figure 3, the “base” sample used to define the effect size, includes the claims for a given provider from the 3 month period prior to “month 1”. The graph refers to a small subsample of the initial data set. The effect size clearly indicates that there are too many deviations from “baseline” for Provider 1 (and maybe Provider 2) in the months following the base period.

5. DISCUSSION

The bottleneck in the discussed approach for mixed data is computing the similarity measures (mapping categorical to continuous variables). Complexity is $O(n^2)$ since every observation is compared to all others for the categorical variables.

One way to speed up the overall detection is to break down data in blocks allowing independent parallel computing. In our real world example such convenient choice is dividing data by provider – this will improve performance when many (large) providers are considered. In addition, using the “effect size” we can also focus only on providers with suspicious (historically) payment patterns.

The rMVD method can be used for further individual claim detection (inside a provider) but the distribution of the rMVD is not clear due also to the use of the categorical variables mapping. It requires investigating proper cut-off values on the rMVD. Alternatively, graphical procedures or other methods, e.g. the discussed here LOF, could be used.

6. ACKNOWLEDGMENTS

The author would like to thank Jan Rowland and Joe Asta from the Advanced Analytics Lab at Optum for their support of his desire to research and apply new methods in his day to day work as a modeling professional.

7. REFERENCES

- [1] Department of Health and Human Services (HHS): Medicaid, September 2014
<https://www.paymentaccuracy.gov/programs/medicaid>
- [2] HHS: Health Care Fraud and Abuse Control Program Annual Report for Fiscal Year 2013, February 2014
<http://oig.hhs.gov/publications/docs/hcfac/FY2013-hcfac.pdf>
- [3] C. C. Aggarwal. *Outlier Analysis*. Springer, 2013.
- [4] C. Chen. Robust Regression and Outlier Detection with the ROBUSTREG Procedure. In *SUGI 27 Conference Proceedings*, paper 265-27, April 2002.
- [5] V. Chandola, S. Boriah, and V. Kumar. A Framework for Exploring Categorical Data. In *Proceedings of 2009 SIAM Data Mining Conference*, pages 187–198, April 2009.
- [6] P. J. Rousseeuw, and K. Van Driessen. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41:212–223, 1999.
- [7] M. M. Breunig, H.- P. Kriegel, R. T. Ng, J. Sander. LOF: Identifying Density-based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*
- [8] Luis Torgo. *Data Mining with R, learning with case studies*. CRC, 2010.