Detecting Anomalous Data in Health-Care Service Records

Alvaro Riascos & Sergio Camelo

Quantil | Matemáticas Aplicadas

Universidad de los Andes

alvaro.riascos@quantil.com.co
sa.camelo38@uniandes.edu.co

Introduction

In Colombia's healthcare system, most of the health services are provided institutions that hold some contract with insurers, who then report and charge the government with a portion of the total costs. These charges are numerous and a number of opportunities for anomalies may arise. In this article we propose a technique that looks for anomalous data in the healthcare databases through parametric calculations of Cross-Entropy. We draw important ideas from Major (2002) We test the technique with a small portion of the 2010 Colombian dataset and find a number of interesting anomalous reports.

Data

We use a dataset that shows the health-care services provided to a random sample of one million Colombians during 2010. This dataset is completely anonymous in patients, insurers and service providers. The variables that will be used in this study are:

1. Risk factors: Gender, Age Group and a list of 33 dummy variables that indicate whether the patient suffers or not from 33 different chronic high-cost diseases.
2. Insurer, from a list of 20 different insurers
3. Total spending
4. Number of claimed services
5. Number of first-time appointments at general medicine
6. Number of control appointments at general medicine
7. Number of first-time appointments at specialised medicine
8. Number of control appointments at specialised medicine

We refer to first-time appointments as those times the patient gets consultation about new health conditions. We expect the number of first-time appointments to be small for every patient and high frequency should be considered anomalous behaviour.

Methodology

We divide the population in different risk groups and look for anomalous behaviour inside each of these groups by comparing the distribution of the population in one particular variable of interest with the distribution of one health insurer. A risk group is characterized by a unique combination of gender, age group and medical diagnosis from the list of 33 diseases.

Inside a risk group, patients are insured by different companies and have information of variables (3)-(7). Within a specific risk group, we say that an insurer is anomalous if its patients show anomalous behaviour in any of those 5 variables explored.

To determine if the data from a particular insurer (say insurer I) is anomalous for a certain variable (say total spending), we compare total spending for Insurer 1 with respect to the total spending by patients from the other insurers.

We report the highest discrepancies found and explore the dataset to see the reasons for these differences.

Results

The next table shows the six reports with the highest relative discrepancies. A row corresponds to the data from a specific insurer and risk group, and about a specific numeric variable.

Age Group	Gender	Disease	Indiv.	Insurer	Variable	Entropy
19-44	F	Asthma	27	A	Spending	36544
5-14	M	Epilepsy	10	E	Services	20038
19-44	M	Osteoarthritis	6	H	1stTime-General	24698
50-54	F	Renal-Other	2	H	Control-General	9926
75-	M	NonChronic	3	AF	1stTime-Specialized	9206
45-49	M	Non-Chronic	36	O	Control-Specialized	9969

This information is not useful unless we look for the causes for the high entropy within each provider and risk group. To do this, we investigate the corresponding data and see that the reason for the high entropy is usually a sole observation with an anomalous report. The next table shows, for each of these 6 anomalous reports, the anomalous observation and compares it with the mean value of the variable within that risk group. We also report spending for the anomalous observation and mean spending for the risk group.

Activity	Anomaly	Mean	Spending	Spending Mean
Spending	US 470.163	US 810	US 470.163	US 810
Services	1354	28	US 23.197	US 1.187
1stTime-General	49	1.9	US 829	US 1.000
Control-General	35	2.64	US 11.791	US 1.891
1stTime-Specialized	11	0.44	US 1.660	US 927
Control-Specialized	12	0.17	US 1.956	US 460

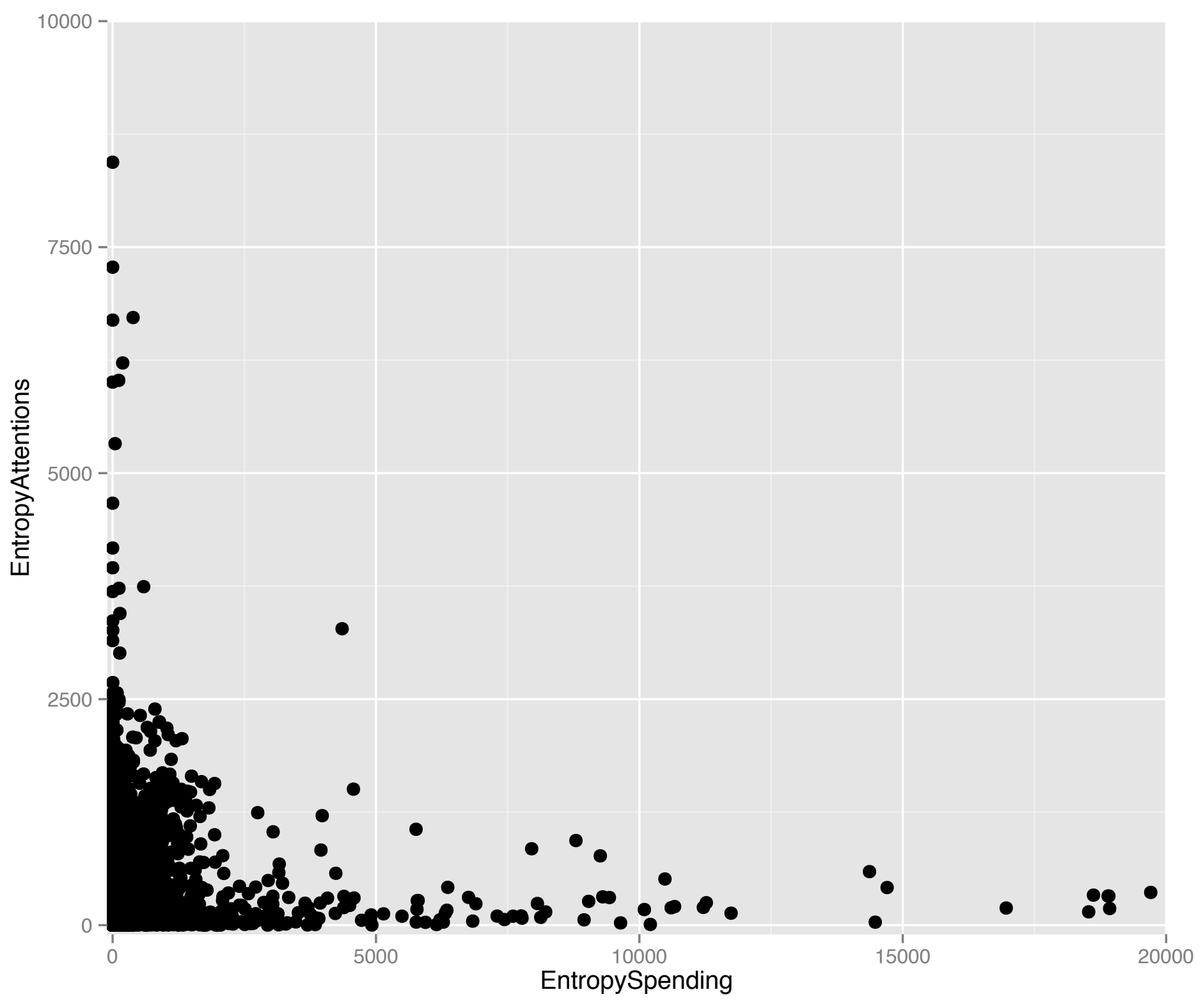
The results are interesting. The first row of both tables show an individual who spent US 470.163, even though the average expenditure of women from age group 5 that suffer from asthma is US 810. The second row shows an epileptic man from age group 3 that reports 1354 health requests, when the average for the people from this risk group is 28. The table also reports the spending for this man, which is US 23.197, in comparison with a mean spending of US 1.187.

Row 3 is really interesting. In Colombia, when a patient reports an untreated medical condition, he has to go through general medicine, and this is reported as a First-Time General Medicine appointment. This situation happens few times a year for each individual, but row 3 shows a patient who went through first-time general medicine 49 times, compared with an average of 1.9 times for individuals from the same risk group. This was probably a problem registering data, because the patient's spending (US 1000) does not differ much from the average within the risk group (US 829).

Rows 4, 5 and 6 show similar results. A non-chronic man, for example, who goes 11 times to First-Time Specialized Medicine in a year, when the risk-group average is 0.44. This man reports spending of US 11.791, but the mean spending inside of the risk group is of US 1.891.

Visualization

Usually anomalous observations come from problems when filling forms. This is not a serious issue and we would like to focus our tool in detecting anomalous situations that are accompanied by anomalous spending. The next plot shows, for all the reports from the database, the relationship between the discrepancy in spending and the discrepancy in the number of medical attentions.



The idea is to focus on maximal points on this graph near the diagonal. In this graph, such points will be highly anomalous both in spending and in the number of attentions. A similar analysis is done for these points as in the previous section. We do not include the results here because of space.

Conclusions

- We propose a technique to check the presence of anomalous data in healthcare datasets using relative cross-entropy.
- The idea is to divide the population in risk groups. Within each risk group we assume that the distribution of the variables under study must be the same independent of the insurer, because of similar health conditions. If an insurer shows a distribution that differs considerably from the risk group, we say its reports are anomalous.
- The technique shows good performance. Interesting anomalous reports are found on the datasets: patients with anomalous spending, high number of claimed services and high number of medical appointments relative to the population from the same risk group.
- A similar study in conducted, this time looking only for cases of anomalous behaviour that are also accompanied by anomalous spending. The results are again very interesting and show the tool is working properly.

References

[1] Kullback, S. 1959. Information Theory and Statistics. New York: Wiley.

[2] Major, J., Riedinger, D. 2002. EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud. Journal of Risk and Insurance Volume 69, Issue 3. Pages 309324.

[3] Morris, C., 1983. Parametric Empirical Bayes Inference: Theory and Applications. Journal of the American Statistical Association, Volume 78. Pages 47-55.

[4] Wen, W., 1990. Information and Probabilistic Reasoning. Annals of Mathematics and Artificial Intelligence, Volume 2. Pages 367-381.