Detecting Fraud in Health Insurance Data: Learning to Model Incomplete Benford's Law Distributions

Fletcher Lu[1] and J. Efrim Boritz[2]

[1] School of Computer Science, University of Waterloo &
Canadian Institute of Chartered Accountants,
66 Grace Street, Scarborough, Ontario, Canada, M1J 3K9
f2lu@cs.uwaterloo.ca
[2] School of Accountancy, University of Waterloo,
200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1
jeboritz@watarts.uwaterloo.ca

Abstract. Benford's Law [1] specifies the probabilistic distribution of digits for many commonly occurring phenomena, ideally when we have complete data of the phenomena. We enhance this digital analysis technique with an unsupervised learning method to handle situations where data is incomplete. We apply this method to the detection of fraud and abuse in health insurance claims using real health insurance data. We demonstrate improved precision over the traditional Benford approach in detecting anomalous data indicative of fraud and illustrate some of the challenges to the analysis of healthcare claims fraud.

1 Introduction

In this paper we explore a new approach to detecting fraud and abuse by using a digital analysis technique that utilizes an unsupervised learning approach to handle incomplete data. We apply the technique to the application area of healthcare insurance claims. We utilize real health insurance claims data, provided by Manulife Financial, to test our new technique and demonstrate improved precision for detecting possible fraudulent insurance claims.

A variety of techniques for detecting fraud have been developed. The most common are supervised learning methods, which train systems on known instances of fraud patterns to then detect these patterns in test data. A less common approach is to have a pattern for *non-fraudulent* data and then compare the test data to this pattern. Any data that deviates significantly from the *non-fraudulent* pattern could be indicative of possible fraud [2]. The difficulty with the latter approach for fraud detection is in obtaining a pattern that one is confident is free of fraud. Digital analysis is an approach which addresses this difficulty. Benford's Law [1] is one digital analysis technique that specifies a model of non-fraudulent data that test data may be compared against.

In 1938 Frank Benford demonstrated that for many naturally occurring phenomena, the frequency of occurrences of digits within recorded data follows a certain logarithmic probability distribution (a Benford distribution). This Benford's Law can only be

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, pp. 633–640, 2005.

applied to complete recorded data. However, incomplete records are very common. We introduce an algorithm to detect and adjust the distribution to take into account missing data. By doing so, we allow for true anomalies such as those due to fraud and abuse to be more accurately detected.

In this paper, we consider the situation where data is contiguously recorded so that the only missing data are due to cutoffs below and/or above some thresholds. Our algorithm, which we will call Adaptive Benford, adjusts its distribution of digit frequencies to account for any missing data cutoffs and produces a threshold cutoff for various ranges of digits. Our algorithm then uses those learned values to analyse test data. We return any digits exceeding a learned set of threshold bounds.

We apply our Adaptive Benford algorithm to the analysis of real healthcare insurance claims data provided by Manulife Financial. The data is a list of health, dental and drug insurance reimbursement claims for a three year period covering a single company's group benefits plan with all personal information removed.

2 Background

2.1 Benford's Law and Fraud Detection

As Benford's Law is a probability distribution with strong relevance to accounting fraud, much of the research on Benford's Law has been in areas of statistics [3, 4] as well as auditing [5, 2]. The first machine learning related implementation was done by Bruce Busta & Randy Weinberg [6].

The significant advantage of using the digital analysis approach over previous supervised learning methods for fraud detection is that we are not restricted to already known instances of fraud [7, 8]. By looking for anomalies that deviate from the expected Benford distribution that a data set should follow, we may discover possible *new* fraud cases.

2.2 Digit Probabilities

Benford's Law is a mathematical formula that specifies the probability of leading digit sequences appearing in a set of data. What we mean by *leading digit sequences* is best illustrated through an example. Consider the set of data

$$S = \{231, 432, 1, 23, 634, 23, 1, 634, 2, 23, 34, 1232\}.$$

There are twelve data entries in set S. The digit sequence '23' appears as a leading digit sequence (i.e. in the first and second position) 4 times. Therefore, the probability of the first two digits being '23' is $\frac{4}{9} \approx 0.44$. The probability is computed out of 9 because only 9 entries have at least 2 digit positions. Entries with less than the number of digits being analysed are not included in the probability computation.

The actual mathematical formula of Benford's law is:

$$P(D = d) = \log_{10}(1 + \frac{1}{d}), \tag{1}$$

where $P(D = d)$ is the probability of observing the digit sequence d in the first 'y' digits and where d is a sequence of 'y' digits. For instance, Benford's Law would state that the probability that the first digit in a data set is '3' would be $\log_{10}(1+\frac{1}{3})$. Similarly, the probability that the first 3 digits of the data set are '238', would be $\log_{10}(1+\frac{1}{238})$. The numbers '238' and '23885' would be instances of the first three digits being '238'. However this probability would not include the occurrence '3238', as '238' is not the *first* three digits in this instance.

2.3 Benford's Law Requirements

In order to apply equation 1 as a test for a data set's digit frequencies, Benford's Law requires that:

1. The entries in a data set should record values of similar phenomena. In other words, the recorded data cannot include entries from two different phenomena such as both census population records and dental measurements.
2. There should be no built-in minimum or maximum values in the data set. In other words, the records for the phenomena must be complete, with no artificial start value or ending cutoff value.
3. The data set should not be made up of assigned numbers, such as phone numbers.
4. The data set should have more small value entries than large value entries.

Further details on these rules may be found in [9]. Under these conditions, Benford noted that the data for such sets, when placed in ascending order, often follows a geometric growth pattern.[1] Under such a situation, equation 1 specifies the probability of observing specific leading digit sequences for such a data set.

The intuitive reasoning behind the geometric growth of Benford's Law is based on the notion that for low values it takes more time for some event to increase by 100% from '1' to '2' than it does to increase by 50% from '2' to '3'. Thus, when recording numerical information at regular intervals, one often observes low digits much more frequently than higher digits, usually decreasing geometrically.

3 Adaptive Benford

As section 2.3 specifies, one of the requirements to be able to apply Benford's law is that there are 'no built-in minimum or maximum values.' However, often data is only partially observed, such as when only a single month of expenses are reported. Adaptive Benford's Law has been designed to handle such missing data situations.

3.1 Missing Data Inflating

The problem with traditional Benford's Law and incomplete data is that the frequency of the digits that are observed become inflated when computed as a probability. For instance, Benford's Law states that in a data set, a first digit of '4' should occur with probability $\log_{10}(1+\frac{1}{4}) \approx 0.0969$. Suppose with complete data, out of 100 observations, 4

[1] Note: The actual data does *not* have to be recorded in ascending order. This ordering is merely an illustrative tool to understand the intuitive reasoning for Benford's law.

appeared as a first digit 10 times, which closely approximates the Benford probability. However if the data set is incomplete with only 50 observations recorded, but all 10 occurrences of first digit 4 are still recorded, then we get a probability of $10/50 = 0.20$, essentially inflating the probability of digits that are observed higher due to the missing digits not being included in the total count for the probability computation.

3.2 Algorithm

Under the condition that we are aware that the observed data follows a Benford distribution and is contiguous, if we are missing data only above or below an observed cutoff, we can use this knowledge to artificially build the missing data. First, let

- d be a leading digit sequence of length i.
- $f_{d,\texttt{observed}}$ be the frequency that the leading digit sequence d occurs in the data set, and
- $P(D_i)$ be the Benford probability for digit sequence D_i, where D_i is any digit sequence of length i.

Now, consider that to compare the actual frequency of occurrence of a leading digit sequence 'd' to the actual Benford probability we would compute the ratio:

$$\frac{f_{d,\texttt{observed}}}{\sum_{D_i} f_{D_i,\texttt{observed}}} \simeq P(D_i = d), \tag{2}$$

where the denominator is summed over all digit sequences of the same length as digit sequence 'd', in other words over all digits of length i. Let

$$C_i = \sum_{D_i} f_{D_i,\texttt{observed}}. \tag{3}$$

Then equation 2 can be rearranged to

$$\frac{f_{d,\texttt{observed}}}{P(D_i = d)} \simeq \sum_{D_i} f_{D_i,\texttt{observed}} = C_i, \tag{4}$$

C_i is essentially a constant scaling factor for *all* digit sequences of the same length i. If there are missing digit sequences in our observed data due to cutoff thresholds, we can compute C_i using the observed digit sequences, since those sequences that do still appear should still follow the Benford's Law probabilities.

In order to produce a best fit for the missing data, we average over all possible C_i values for a given digit sequence length i. Therefore, let

$$C_i = \frac{f_{d,\texttt{observed}}/P(D_i = d)}{|\texttt{digit sequences of length i}|}. \tag{5}$$

This scaling factor C_i will be used to 'fill-in' the missing data of our Benford distribution.

As an example, C_2 would be the averaged constant scaling factor over all first two digit frequencies. We use C_i to multiply our Benford probabilities for the digit sequences of length i and use that as a benchmark to compare the frequencies of the actual observed data against.

Appendix A illustrates the Adaptive Benford algorithm. The major steps of the Adaptive Benford algorithm are:

1. Compute the C_i constant values for various leading digit sequence lengths.
2. Compute artificial Benford frequencies for the digit sequence lengths.
3. Compute a standard deviation for each of the sequence lengths.
4. Flag any digit sequences in the recorded data that deviate more than an upper bound number of standard deviations from the artificial Benford frequencies.

We compute the artificial Benford frequencies as follows:

$$f_{d,expected} = C_i \times \log_{10}(1 + \frac{1}{d}). \tag{6}$$

We scale up to actual frequencies in contrast to dividing by a sum total of observed instances that would produce probabilities. By doing so, we avoid the inflating effect we noted in section 3.1. We may compute a variance against observed data by:

$$\sigma^2_{expected,i} = \frac{1}{n_i} \sum_{D_i} (f_{D_i,observed} - f_{D_i,expected})^2, \tag{7}$$

where n_i is the number of different digit sequences of size i. We compute an upper bound U_i based on a number of standard deviations from the artificial Benford frequencies. We use this upper bound to determine if the observed data deviates enough to be considered anomalous and potentially indicative of fraud or abuse.

4 Experiments

For the purposes of our experiments, we will analyse up to the first 3 leading digit sequences. The choice of digit sequence lengths to be analysed is dependent on the data set's entries (the digit lengths of the data set's elements as well as the number of elements in the data set). For a further discussion on choice of digit sequence length see [10].

4.1 Census Data Tests

As an initial test of our system we use as a test database the year 1990 population census data for municipalities in the United States, which has been analysed previously by Nigrini [9] and been verified to follow a Benford Law distribution.

Table 1 records the amount of conformity for complete and incomplete census data whereby we measure conformity as the percentage of digit sequences that fall within ± 2 standard deviations of the Benford estimate value out of the total number of digit

Table 1. Census Data: Percentage of Digits within ± two standard deviations of Benford and Adaptive Benford distributions.

Range of Values x 10^4	Data Size	Classic Benford	Adaptive Benford
Complete Census	3141	96.2%	94.9%
100,000 - 1,100,000	431	85.0%	89.4%
200,000 - 1,200,000	220	85.6%	96.9%
300,000 - 1,300,000	144	49.5%	94.4%
400,000 - 1,400,000	106	30.8%	91.7%
500,000 - 1,500,000	84	32.0%	87.2%
600,000 - 1,600,000	65	34.0%	84.0%
700,000 - 1,700,000	48	33.8%	82.4%
800,000 - 1,800,000	36	19.2%	82.7%
900,000 - 1,900,000	24	11.8%	73.5%

sequences that had non-zero frequency.[2] We modified the 1990 census data to include sets of various population ranges of municipalities. Notice that for the range 100,000-1,100,000 all leading digit sequences may start with any of 1,2,...,9. However, for the range 900,000-1,900,000, the only possible leading digits start with 9 or 1. With fewer possible leading digit sequences, the inflating effect mentioned in section 3.1 becomes more likely, resulting in lower conformity as the population range shifts higher. The Adaptive Benford, which compensates for the cutoff data, produces higher conformity values, ranging from 73.5% to 96.9%.

4.2 Health Insurance Data Tests

We now analyse health insurance claims data covering general health, dental and drug claims for financial reimbursement to Manulife Financial covering a single company's group benefits plan for its employees from 2003 to 2005. With recorded data before 2003 cutoff, we expect 'inflated' percentages of *anomalous* digits with traditional Benford compared with our Adaptive Benford method.

Our goal is to detect anomalies in our data sets that may be indicative of fraud activity. Table 2 reports the percentage of anomalous digit sequences for the insurance database for various data sets. As expected, by handling missing data ranges due to cutoff levels, Adaptive Benford can be more precise, reporting fewer actual anomalous digit sequences then traditional Benford. We used a 95% upper bound confidence interval as our anomaly threshold.

The main advantage of our Adaptive Benford algorithm over traditional Benford for fraud detection is its improved precision for detecting anomalies. The goal, once anomalous digit sequences have been identified, is then to determine the data entries that are causing the high amounts of anomalous digit sequences. A forensic auditor may then,

[2] The two standard deviations should cover approximately 95% of the digit sequences if the data conforms to Benford's distributions. The standard deviation used for all tests of table 1 are computed using the complete 1990 census data.

Table 2. Health Insurance Fraud Detection: Comparing Traditional and Adaptive Benford against percentage of anomalous digit sequences.

Data Set	Description	Data Size	Classic Benford	Adaptive Benford
1	Misc. Dental Charges	589	21.05%	11.48%
2	Expenses Submitted by Provided	31,693	3.77%	3.44%
3	Submitted Drug Costs	6,149	16.80%	10.40%
4	Submitted Dispensing Fee	7,644	18.18%	9.09%
5	Excluded Expenses	1,167	10.39%	5.19%
6	Expenses after deductibles	29,215	3.65%	3.13%
7	Coinsurance reductions	3,871	8.85%	6.19%
8	Benefit Coordination Reductions	286	29.29%	6.06%
9	Net Amount Reimbursed	28,132	4.40%	3.70%

for instance, decide whether these entries are likely cases of fraud or abuse. Making such decisions is often a qualitative judgement call dependent often on factors related to the specific application area. Anomalies, in some cases, may be due to odd accounting or data entry practices that are not actual instances of fraud. We therefore have avoided here labeling our reported anomalies as actual fraud. Instead, we emphasize that these digit sequence anomalies are to be used as a tool to indicate possible fraud.

5 Discussion and Conclusions

In contrast to typical supervised learning methods which will train on known fraud instances, our Adaptive Benford algorithm models the data to an expected non-fraudulent Benford data pattern and any large anomalies are reported as possible fraud.[3] Our Adaptive Benford algorithm allows us to analyse data even when the data is partially incomplete. We made such an analysis with incomplete health insurance data, which only included data for a three year period. Our Adaptive Benford algorithm reports fewer anomalous digit sequences, avoiding the transient effect due to artificial cutoff start and end points for recorded data. This produces a more precise set of anomalous leading digit sequences than traditional Benford for forensic auditors to analyse for fraud. In effect, our Adaptive Benford algorithm removes requirement 2 of the rules specified in section 2.3 needed for Benford's Law to be applied. Our Adaptive Benford algorithm therefore expands the areas where Benford's Law may be applied.

Acknowledgments

We would like to thank Manulife Financial for providing the insurance data and Mark Nigrini for providing the census data. We would also like to thank the Canadian Institute of Chartered Accountants and the Natural Sciences and Engineer Research Council (NSERC) for providing funding.

[3] This modeling is under the pre-condition that, if we had a complete data set without fraud activity, it should follow a Benford distribution. (i.e. The complete, non-fraudulent, data set satisfies the requirements of section 2.3.

A Adaptive Benford Algorithm

Let S be the observed testing data
  Let U_i be an upper bound on the number of standard deviations
  For digit sequences d = 1,2,3,...,Upperbound:
    $f_{d,observed}$ = number of times digit sequence d appears as a leading digit sequence in S
  For i = 1...Upperbound on digit length:
    Let n_i be number of digits of length i that appeared at least once in S
    Compute over all digit sequences D_i of length i:
      $\hat{C}_i = \frac{1}{n_i} \sum_{D_i} \frac{f_{D_i,observed}}{\log_{10}(1+\frac{1}{D_i})}$
    Compute for each digit sequence d of length i:
      $f_{d,expected} = \hat{C}_i \times \log_{10}(1 + \frac{1}{d})$
    Compute over all digit sequences D_i of length i:
      $\hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{D_i} (f_{D_i,observed} - f_{D_i,expected})^2$
    Compute for each digit sequence d of length i:
      if $U_i < \frac{f_{d,observed} - f_{d,expected}}{\sigma_i}$ then store d as anomalous

References

1. Benford, F.: The Law of Anomalous Numbers. In: Proceedings of the American Philosophical Society. (1938) 551–571
2. Crowder, N.: Fraud Detection Techniques. Internal Auditor **April** (1997) 17–20
3. Pinkham, R.S.: On the Distribution of First Significant Digits. Annals of Mathematical Statistics **32** (1961) 1223–1230
4. Hill, T.P.: A Statistical Derivation of the Significant-Digit Law. Statistical Science **4** (1996) 354–363
5. Carslaw, C.A.: Anomalies in Income Numbers: Evidence of Goal Oriented Behaviour. The Accounting Review **63** (1988) 321–327
6. Busta, B., Weinberg, R.: Using Benford's Law and neural networks as a review procedure. In: Managerial Auditing Journal. (1998) 356–266
7. Fawcett, T.: AI Approaches to Fraud Detection & Risk Management. Technical Report WS-97-07, AAAI Workshop: Technical Report (1997)
8. Bolton, R.J., Hand, D.J.: Statistical Fraud Detection: A Review. Statistical Science **17(3)** (1999) 235–255
9. Nigrini, M.J.: Digital Analysis Using Benford's Law. Global Audit Publications, Vancouver, B.C., Canada (2000)
10. Nigrini, M.J., Mittermaier, L.J.: The Use of Benford's Law as an Aid in Analytical Procedures. In: Auditing: A Journal of Practice and Theory. Volume 16(2). (1997) 52–67