Data Mining for Detecting Carelessness or Mala Fide Intention

Rajesh Kumar
Assistant Professor, Department. of ECS
Dronacharya College of Engineering
Gurgaon, India.

ABSTRACT

Fraud is one of the greatest challenges for the organizations, it needs a machine to be equipped with data mining algorithms , so that it can detect a crime pattern before it takes place. This paper will explore the data mining and Knowledge discovery in data base and later one of the most effective data mining techniques called Benford's law for detecting the fake entries in medical insurance claims, electricity bills, water bills etc will be discussed. Applications of Benford's law with limitations will be discussed so that machines exhibits some intelligence in its domain and later proposed to embed the Benford law in software to identify all the entries made by carelessness or with a mala fide intention .

Keywords

Benford law, Data mining, KDD, Preprocessing.

I. INTRODUCTION

The information processing capability increased at a very fast pace. It was very difficult for the IT experts to develop a safe and secure systems . Every systems was vulnerable to attack and every attack brought defame to the organization so the need of developing software having some intelligence emerged. Paper discusses the Benford law and its application, limitations and analysis of data.

Paper is organized as follows

Section 2 deals with data mining and knowledge discovery in data bases.

Section 3 covers Benford's law.

Section 4 covers Benford's law applications and limitations.

Section 5 covers the data analysis.

Section 6 covers conclusion.

II. DATAMINING

Data mining is the discovery of new information in terms of pattern or rules from the huge amount of data or can be stated as the process of employing one or more computer learning techniques to automatically analyze and extract the knowledge from the data. Whereas knowledge discovery in data bases (KDD) field is concerned with the development of methods and techniques for making sense of data.[2] . In general view, KDD refers to the overall process of discovering useful knowledge from data and data mining refers to a particular step in the KDD process. Data mining is the application of specific algorithms for extracting patterns from the huge data[2] . Steps like data preparation, data selection, data cleaning ,application of Apriori algorithms, and proper interpretation of the results of mining, are essential to ensure that useful and relevant knowledge is derived from the data. KDD has matured from the various interdisciplinary fields like machine learning, artificial intelligence, neural networks and statistics to extract some useful knowledge from the data. It is necessary to acquaint with the steps of KDD process. Following are the steps of KDD process as displayed in Figure 4.

A. Sampling

Huge amount of data cannot be analyzed so some sampling methods are applied to make task easier. Various methods of sampling are in use for business purpose. Sampling can be problastic (random sampling) or non probabilistic (sampling depending on expert judgment)

B. Preprocessing

Humane is error. Generally in haste people leave some forms item blank, spelling mistakes, noisy data and inconsistency. This leads to a detected pattern which is not so useful. Certain data mining algorithm requires pre-processing for better performance, for instance Neural Networks is incapable of performing on string data type. Various authors has proposed to measure the central tendency and to replace the unfilled data with the central tendency. Outlier detection is essential to find the noisy data to extract relevant information.

C. Transformation

Transformation or normalization is the rescaling of data into suitable range. This increases the memory utilization, which further leads to high speed and simplicity. Various transformation methods are being used.

- Decimal scaling divides each value by same power of 10.
- Linear transformation of original input to the new range. Ex. All number between 1-100 are transformed over the scale0-1. It can be found out by the formula.

$$Y' = \frac{(\text{actual data} - \min)}{(\text{Max-min})} * [\max' - \min'] + \min' \quad (1)$$

- Z score transformation can be done to transform where minimum and maximum are not known.

Z score can be calculated by the formula.

$$Z = \frac{(\text{Actual data} - \text{mean Value})}{\text{standard deviation.}} \quad (2)$$

Z score will transform in the range 0 to 1. Whereas Z score greater than 3 is considered as outlier.

D. Data mining

Data mining is the searching for patterns of interest in a particular representational form or a set of such representations by using a particular algorithm. It includes rules or trees, regression and

clustering. The user can significantly improve the data-mining method by correctly performing the above steps [2].

III. BENFORD LAW

Initially it seems that digits are equally likely to distribute in a number that forms an observations like electricity bill , but this conception was wrong and demystified by Benford law. Benford law and its aid in analytical procedures by distribution of numbers on various digits positions is given in paper "The use of benford law as an aid in analytical procedure ".[6] . Benford law states that the probability of any digit D from 1 to 9 being the first digit is where distribution is not uniform is given by

$$\text{Log}_{10}(1+1/D) \quad (3)$$

Whereas probability at 2nd digit can be given by

$$\sum_{D_1=1}^{9} \log_{10}(1+1/D_1D_2) \quad (4)$$

$D_2=\{0,1\text{----}9\}$

And probability of combination of 1st digit and 2nd digit can be given by the formula.

$$P(D1D2)=\log_{10}(1+1/D1D2) \quad (5)$$

Whereas D1D2={10,11------,99)

Table 1. Frequencies based on Benford's Law[5]

Digit	1st Place	2nd	3rd Place	4th Place
0		0.11968	0.10178	0.10018
1	0.30103	0.11389	0.10138	0.10014
2	0.17609	0.19882	0.10097	0.1001
3	0.12494	0.10433	0.10057	0.10006
4	0.09691	0.10031	0.10018	0.10002
5	0.07918	0.09668	0.09979	0.09998
6	0.06695	0.09337	0.0994	0.09994
7	0.05799	0.0935	0.09902	0.0999
8	0.05115	0.08757	0.09864	0.09986
9	0.04576	0.085	0.09827	0.09982

Benford law is most suitable for the following applications[4]

- Data comes from the two distributions , example electricity bill is composed of no of units (KWH) and rate(KWH *rate)..
- No sampling of data is done in order to extract information.
- It gives better result on the large data set.

IV. BENFORD LAW APPLICATIONS AND LIMITATIONS

As we have talked earlier, Benford law is used in finding the fake bills. We can find out the fake bills by finding the occurrence of digits in bills under study. We can first go for the first digit test followed by the second ,third digit and combination of first, second digits and so on. A bill is scrutinized manually after finding deviation from the Benford proposed distributions. Following authors has worked on Benford's law to give mesmerizing results.

Image forensics is discussed by Ding Dong et al.[7].

Rigging of Iranian election was understood by Benford law [1].

Identifying the persons involved in tax evasions was proposed by Benford law[5].

Benford law was used to find the naturally occurring prices on Ebay.[3].

Benford law can also be used in finding the fake scientific data and it can be used in all applications where probability of occurrence of an event is not same.

Following are the limitations where Benford law cannot be used[4].

- Data sets constituted by the assigned numbers. Example Zip number, Pin code etc.
- Numbers that are influenced by the human thoughts. Example prices set at psychological threshold like Rs. 99, Rs. 999 etc.
- Accounts with large number of firm specific numbers or default values. Example incase door is closed minimum reading policy adopted by electricity departments.
- Accounts with a built in minimum or maximum.

V. DATA ANALYSIS.

In Figure 1. data is taken from the bills of a firm, analysis shows that there is wide disparity of number 2 at first digit. This can interpreted as bifurcation of few bills greater than 25000 and less than 50000, It was done to avoid the paper work as it was mandatory to fill the declaration for all bills above 25,000.

In the second case as displayed in Figure 2. number one repetition was too much than the Benford analysis, it may be interpreted as company purchased most of the items whose price was between Rs.1000 to Rs.1999, as the item was frequently used.

In the third case as represented in Figure 3., electricity bills of customers were collected, wide disparity of occurrence of digit 9 was found, it was due carelessness of electrician by not taking the readings and just manipulating himself.

VI. CONCLUSION

It is very common in the information age that most of the data is filled online and it is possible that data entry is done incorrectly, to avoid this Benford's law can be embedded in software to provide artificial intelligence in software. This method can be used in all companies where data entry is done online and it is difficult to identify the entries done in carelessness or with a mala fide intention.

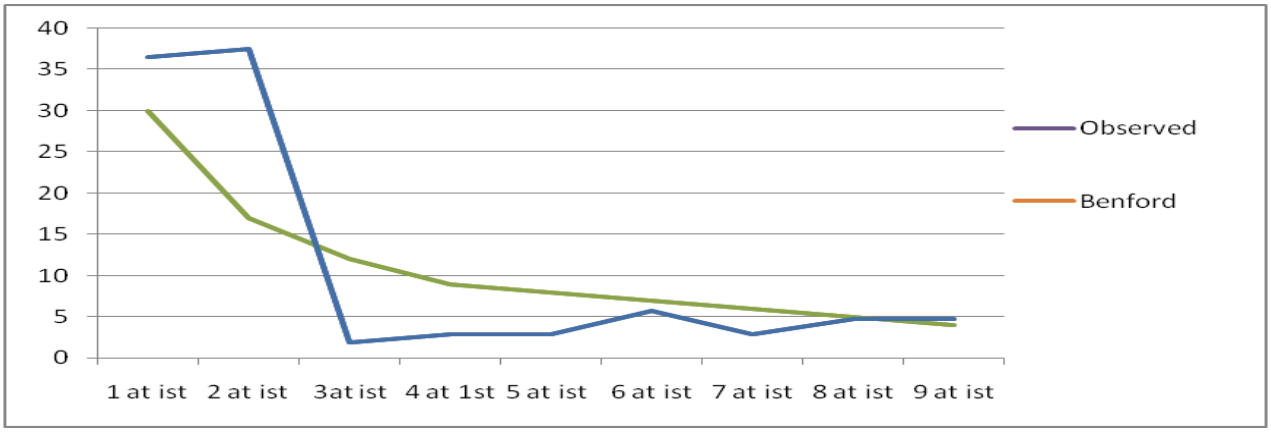

Figure 1.Comparison of Benford law with observed pattern in Bills Where Y axis is % of Occurrence and X axis represents digit at 1st position .

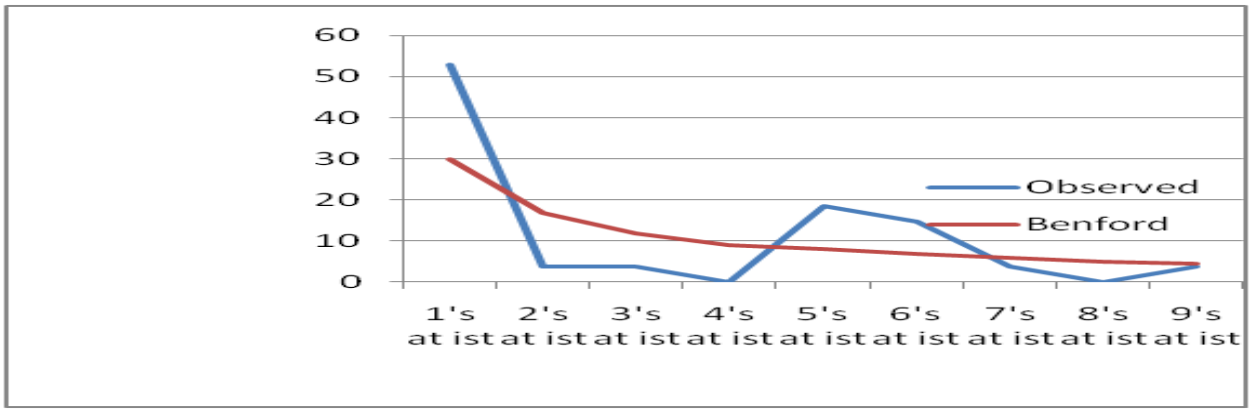

Figure 2.Comparison of Benford law with the bills of utility firm where Y axis is % of occurrence and X axis represents digits at first position

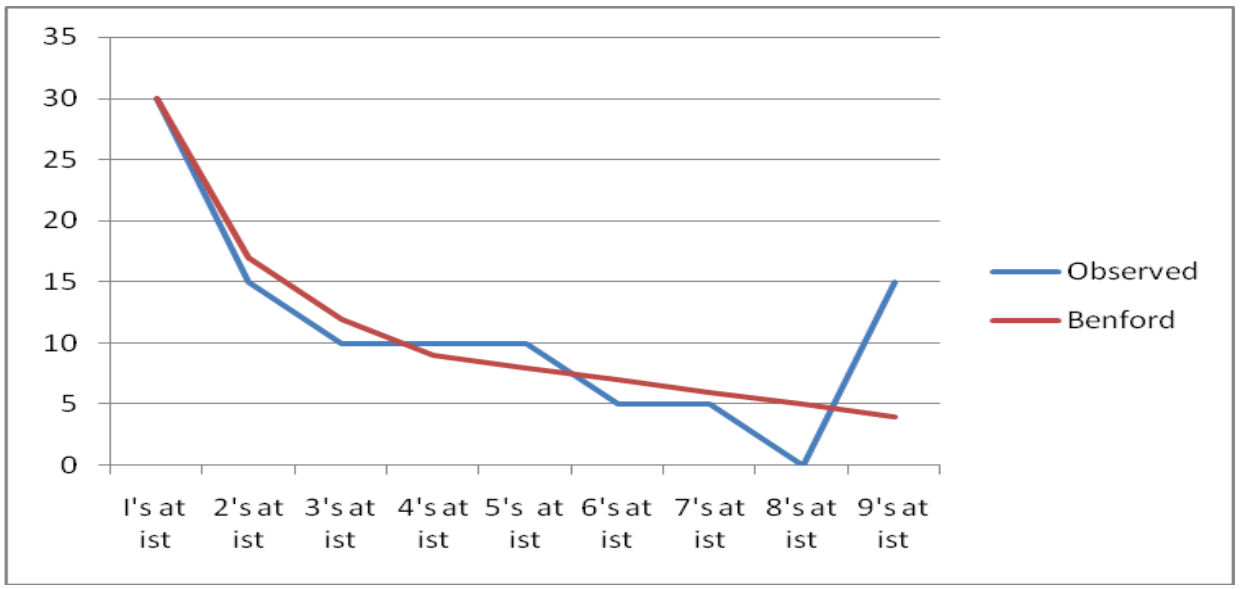

Figure 3. Comparison of Benford law with observed pattern in electricity bills where Y axis is %of occurrence and X axis represents digits at first position.

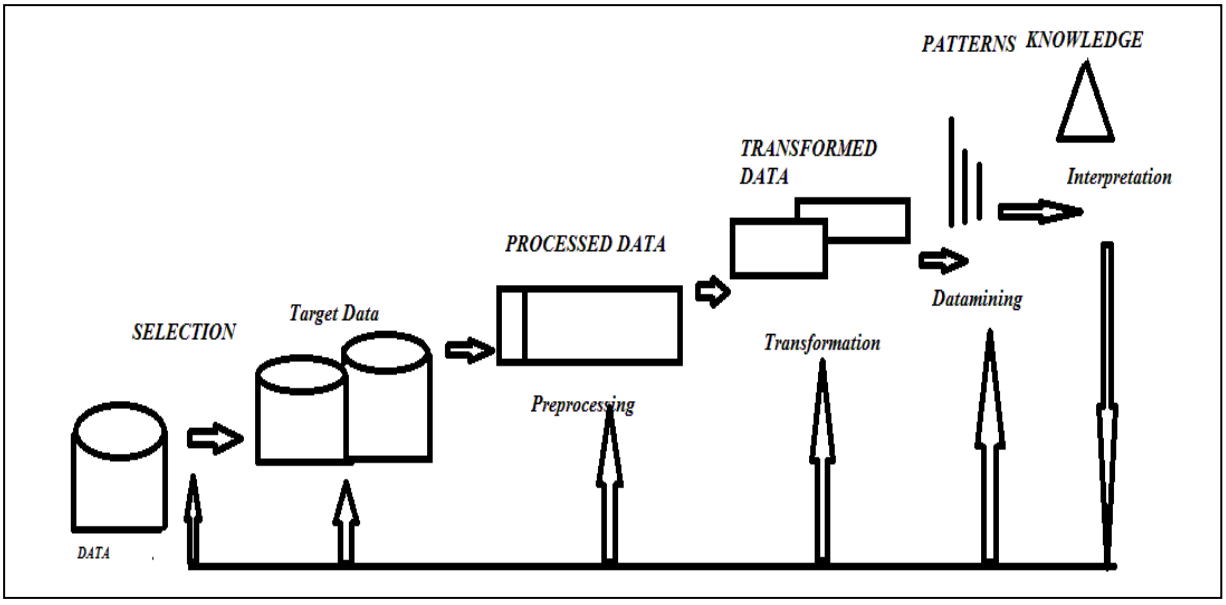

Figure 4. Process of Knowledge discovery in database

VII.REFRENCES.

[1] Stephen Battersby "Rigging of electoral polls in Iranian elections. Statistics hint at fraud in Iranian election", New scientist june24,2009.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth ,"To Knowledge Discovery in Databases", 6, American Association for Artificial Intelligence. AI Magazine Volume 17 Number 3 (1996)

[3] Giles," Benford's law and naturally occurring prices in certain e bay auctions". Econometrics Working Paper EWP0505, University of Victoria, Department of Economics. Forthcoming in Applied Economics Letters.,2006.

[4] Durtschi, Cindy and William Hillison and Carl Pachini. "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data", Journal of Forensic Accounting 1524-5586/Vol. V(2004): 17-34.

[5] Nigrini, M. J. (1997). Digital Analysis Tests and Statistics. Allen, Texas: The Nigrini Institute, Inc. Mark_Nigrini@classic.msn.com,1997

[6] Mark, j . Nigrini and Linda ,"The use of benford law as an aid in analytical procedure ",Auditing a journal of practice and theory ,vol16,no2,fall1997

[7] Dongdong Fu*a, Yun Q. Shi*a, Wei Sub ,"A generalized Benford's law for JPEG co efficient and its applications in image forensics"

[8] Agrawal, R., and Psaila, G. Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), 3–8. Menlo Park, Calif.: American Association for Artificial Intelligence,1995.

[9] Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.;and Verkamo, I. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, eds,1996.

[10] Smyth, and R. Uthurusamy,." Detection of Abrupt Changes: Theory and Application,. Englewood Cliffs, N.J.: Prentice Hall, 514–560. Menlo Park, Calif.: AAAI Press. Basseville , M., and Nikiforov, I. V. 1993.

[11] Brachman, R., and Anand, T.. "The Process of Knowledge Discovery in Databases: A Human-Centered Approach". In Advances in Knowledge Discovery and Data Mining, 37–58, Edition, 1996.

[12] Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. Classification and Regression Trees,1984.

[13] Hall, J.; Mani, G.; and Barr, D. "Applying Computational Intelligence to the Investment Process" ,In Proceedings of CIFER-96: Computational Intelligence in Financial Engineering. Washington ,D.C.: IEEE Computer Society,1996.

[14] Langley, P., and Simon, H. A. " Applications of Machine Learning and Rule Induction",. Communications of the ACM 38:55–64,1995.

[15] Djoko, S.; Cook, D.; and Holder, L." Analyzing the Benefits of Domain Knowledge in Substructure Discovery" In Proceedings of KDD-95: First International Conference on Knowledge Discovery and Data Mining, 75–80. Menlo Park, Calif.: American Association for Artificial Intelligence,1995.

[16] Dzeroski, S.. Inductive Logic Programming for Knowledge Discovery in Databases. In Advances in Knowledge Discovery and Data Mining, eds,1996

[17] Etzioni, O. 1996. The World Wide Web: Quagmire or Gold Mine? Communications of the ACM (Special Issue on Data Mining). November 1996.