A fractal statistical fraud detection analysis in databases

P V Hripunov[1]

[1]Pension Fund of the Russian Federation, Shabolovka str. 4, Moscow, Russia, 119991

Abstract. The paper proposes a method for fraud detection analysis in databases. The main idea of the method is to use fractal analysis of numerical information in databases to identify anomalies caused by substitution or distortion of initial data.In this paper, a criterion based on fractal analysis is used. The results of experiments on simulated data are presented.

1. Introduction

The task of identifying fraud in corporate and government databases is one and the most important in data mining. In recent years, a large number of methods for solving this problem have been developed [1]. The main approaches to fraud detection today are: neural networks [2,3], logistic model [4], support vector machine [4], decision trees [5], genetic algorithm [5], text mining [6], Bayesian belief network [7], featureless approach [12] and others. In this paper, a criterion based on fractal analysis is used. This is a fairly new approach for the problem of searching for fraudulent operations in databases. The effectiveness and prospects of this approach is shown in [8,9]. At the same time, the use of fractal methods in other areas, for example, for pattern recognition on images [10], to detect intentional distortions [11], is well developed. In this paper, the methods of fractal analysis for image recognition are adapted to work with large databases.

2. Iterated function systems for database analysis

Classiciterated function systems (IFS) algorithm for images searches the best affine transformation from domain to range block for every range block[10]. As a result, an input image is coded by several affine transformations:

$$\begin{aligned} \mathbf{I}^* &= F\left(\mathbf{I}\right) = \mathbf{C}_{1,4}\mathbf{I} + \mathbf{c}_{5,6}, \\ u^*_{i,j} &= c_7 \cdot u_{i,j} + c_8, \end{aligned} \tag{1}$$

where $\mathbf{I}^* = (i^*, j^*)^T$, $\mathbf{I} = (i, j)^T -$) is the coordinates of pixel from domain and range block accordingly, $\mathbf{C}_{1,4} = \begin{bmatrix} c_1 & c_2 \\ c_3 & c_4 \end{bmatrix}$, $\mathbf{c}_{5,6} = \begin{bmatrix} c_5 \\ c_6 \end{bmatrix}$ – transformation coefficients, $u^*_{i,j}$, $u_{i,j}$ – is the pixel brightness from range and domain area, a c_7, c_8 – contrast and brightness shift parameter.

The transformation is conducted in a class of contraction mapping to obtain a unique and stable fractal image. Parameters of transformations $c_1 - c_8$ are computed by IFS fractal compression algorithm: $c_1 - c_4$ are selected from the possible sets, c_5, c_6 are calculated in the process of searching the best affine transformation from domain to range block, c_7, c_8 – are calculated on the average brightness of domain and range blocks. Specificity analysis of numerical and textual data requires

adaptation of this approach. The main element for the analysis was selected rows of database tables. First, the source data in the database tables often contain heterogeneous information: text, numbers, images, binary data, etc. In our experiments, only text and numeric information was used, and in the preprocessing phase, the text was converted into numbers in accordance with the character encoding table. As a result of preprocessing, each row of the table was represented as a one-dimensional array of numbers. After that, each line of the database is divided into one-dimensional range and domain areas, and by analogy with the formula (1), self-similar data sections are searched. As a result, at the training stage, we select from the database knowingly genuine and correct rows, and form a set of corresponding transformations. At the stage of recognition of fraudulent records in the database, using sets of received transformations, we find the distance:

$$D_i = \frac{d(F_i^* I^*, I^*)}{I_s^*}, \quad (2)$$

where I^* – initial database row, I_s^* – size of initial database row, F_i^* – set of transformations for correct rows, d - Euclidean norm. A distance value significantly greater than the average value for a particular table in the database means that the current row can be fraudulent.

3. Results and discussion

To test the approach described above, we used test database tables describing pension contributions in the corporate enterprise database.An example of the initial data is shown in Figure 1.

	Регистр...	Но...	Акти...	Период	Физлицо	Организация	Период ...
−	Списан...	1	✔	04.06.2010...	Краснов Васи...	ИнвестЗаказ...	01.01.20...
−	Списан...	2	✔	04.06.2010...	Женилов Серг...	ИнвестЗаказ...	01.01.20...
−	Списан...	3	✔	04.06.2010...	Иванов Иван ...	ИнвестЗаказ...	01.01.20...
−	Списан...	4	✔	04.06.2010...	Сидоров Петр ...	ИнвестЗаказ...	01.01.20...
+	Начисл...	1	✔	30.06.2010...	Краснов Васи...	ИнвестЗаказ...	01.06.20...
+	Начисл...	2	✔	30.06.2010...	Женилов Серг...	ИнвестЗаказ...	01.06.20...
+	Начисл...	1	✔	30.06.2010...	Петров Петр И...	ИнвестЗаказ...	01.06.20...
+	Начисл...	1	✔	30.06.2010...	Иванов Иван ...	ИнвестЗаказ...	01.06.20...
+	Начисл...	1	✔	30.06.2010...	Сидоров Петр ...	ИнвестЗаказ...	01.06.20...

Figure 1. Fragment of the database table.

To train the method, 10,000 rows were used from the database.Then, we changed the database, in the first version we added 100 new lines with the correct information, in the second version we added 100 lines with false and fraud information.And counted the distance by formula (2) for both versions.The results of the distribution of distances are shown in Figure 2.

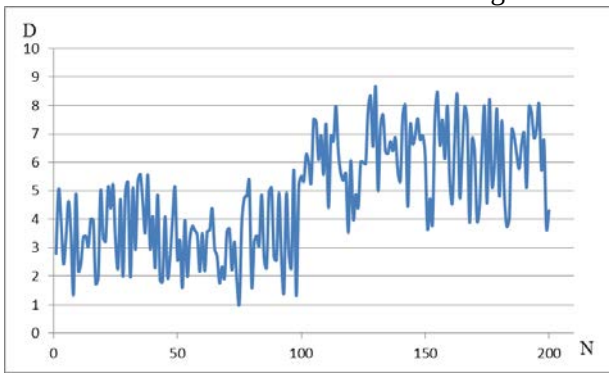

Figure 2. Distribution of distances, N=1…100 – correct rows, N=101…200 – false rows.

As a result of the experiment, it was found that for 73% of false rows, the distance was significantly larger than the correct ones.In future work planned to find the area of applicability of this approach and to conduct experiments on a large sample of data.

4. References

[1] West J and Bhattacharya M 2016 Intelligent financial fraud detection: a comprehensive review *Computers and security* **57** 47-66
[2] Ngai E, Hu Y, Wong Y, Chen Y and Sun X 2011 The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature *Decision Support Systems* **50** 559-569
[3] Zhang D and Zhou L 2004 Discovering golden nuggets: data mining in financial application *Systems, Man, and Cybernetics, Part C: Applications and Reviews IEEE Transactions* **34** 513-522
[4] Bhattacharyya S, Jha S, Tharakunnel K and Westland J C 2011 Data mining for credit card fraud: a comparative study *Decision Support Systems* **50** 602-613
[5] Ravisankar P, Ravi V, Raghava Rao G and Bose I 2011 Detection of financial statement fraud and feature selection using data mining techniques *Decision Support Systems* **50** 491-500
[6] Humpherys S L, Moffitt K C, Burns M B, Burgoon J K and Felix W F 2011 Identification of fraudulent financial statements using linguistic credibility analysis *Decision Support Systems* **50** 585-594
[7] Kirkos E, Spathis C and Manolopoulos Y 2007 Data mining techniques for the detection of fraudulent financial statements *Expert Systems with Applications* **32** 995-1003
[8] Padua R N and Borres M S 2017 From Fractal Geometry to Statistical Fractal *Recoletos Multidisciplinary Research Journal* **1(1)**
[9] Uy K J D and Zanoria M L E 2017 A Fractal Statistical Analysis of Enron Stock Prices *Recoletos Multidisciplinary Research Journal* **2(2)**
[10] Minaev E Y and Nikonorov A V 2012 Object detection and recognition in the driver assistance system based on the fractal analysis *Computer Optics* **36(1)** 124-130
[11] Ozawa K 2008 Dual fractals *Image and Vision Computing* **26(5)** 622-631
[12] Yumaganov A S and Myasnikov V V 2017 A method of searching for similar code sequences in executable binary files using a featureless approach *Computer Optics* **41(5)** 756-764