IDENTIFYING BANK FRAUDS USING CRISP-DM AND DECISION TREES

Bruno Carneiro da Rocha[1,2] and Rafael Timóteo de Sousa Júnior[2]

[1]Bank of Brazil, Brasília-DF, Brazil
brunorocha_33@hotmail.com
[2] Network Engineering Laboratory, University of Brasilia (UnB), Brasilia-DF, Brazil
desousa@unb.br

ABSTRACT

This article aims to evaluate the use of techniques of decision trees, in conjunction with the management model CRISP-DM, to help in the prevention of bank fraud. This article offers a study on decision trees, an important concept in the field of artificial intelligence. The study is focused on discussing how these trees are able to assist in the decision making process of identifying frauds by the analysis of information regarding bank transactions. This information is captured with the use of techniques and the CRISP-DM management model of data mining in large operational databases logged from internet bank transactions.

KEYWORDS

Fraud detection, fraud prevention, decision taking, machine learning, decision trees, data mining.

1. INTRODUCTION

Banks have strong security systems aimed to protecting the access to internet banking services through the Internet, but can not guarantee the security of computers that customers use, and how they are used, to avoid problems of electronic fraud.

It is almost impossible to eradicate bank fraud. What can be done is to minimize frauds and prevent them. Quinlan, creator of the ID3 and C4.5 algorithms, described in his book C4.5: Programs for Machine Learning, published in 1993 [1], that many applications of artificial intelligence are based on a model of knowledge that is usually employed by a human specialist. In some cases, the data analyzed by the expert should be classified for better observation, or placed in certain categories or classes according to their main features. In this paper, studies of classifications and their results are used to help in the prevention of bank fraud. The method of study is called Decision Trees, which will be discussed in the sections below and will be implemented within a management model of data mining, called CRISP-DM.

This paper is organized as follows. Section 2 is aimed at discussing related work. Section 3 presents a review of CRISP-DM. In section 4 we discuss the main characteristics of Decision Trees and the methods to build a good decision tree based on information theory principles. The implementation of a decision tree for bank fraud detection is described, as well as the analysis of the results are presented on Section 5. Finally, we conclude our work in section 6.

2. RELATED WORK

There are several types of research works in the domain of fraud detection. They include fraud detection in credit cards, telecommunications, money laundering, and intrusion detection. Usually the proposed techniques use artificial intelligence in general, employing either

individually or conjointly solutions from artificial neural networks, statistical analysis, econometrics, expert systems, fuzzy logic, genetic algorithms, machine learning, pattern recognition, visualization and others.

Papers [2] and [3] present broad surveys and discussion of research regarding techniques for tackling various types of frauds. Paper [4] describes the tools available for statistical fraud detection and the areas in which fraud detection technologies are most used, pointing out the fundamental fact that seldom one can be certain, by statistical analysis alone, that a fraud has been perpetrated. Due to this uncertainty, in [5], the discussion is centered on how databases of customer transactions have to be submitted to several data mining techniques that search for patterns indicative of fraud, a process which represents a challenge in fraud detection given the need to find algorithms that can learn to recognize a great variety of fraud scenarios and adapt to identify and predict new scenarios.

This paper takes into account these studies and tries to bring an effective fraud detection solution based on decision trees and data mining, with tests on large databases logged from bank transactions.

3. CRISP-DM

The Cross Industry Standard Process for Data-Mining – CRISP-DM [6] [7] is a model of a data mining process used to solve problems by experts. The model identifies the different stages in implementing a data mining project, as described bellow.

3.1. Implementation of the CRISP-DM

CRISP-DM is based on the process flow showed in Figure 1. The model proposes the following steps:

1. Business Understanding – to understand the rules and business objectives of the company.

2. Understanding Data – to collect and describe data.

3. Data Preparation – to prepare data for import into the software.

4. Modelling – to select the modelling technique to be used.

5. Evaluation – to evaluate the process to see if the technique solves the problem of modelling and creation of rules.

6. Deployment – to deploy the system and train its users.

3.2. Business Understanding

The first phase of the CRISP-DM is the Business Understanding. For the sake of this paper this phase is aimed at defining the business objectives of the bank. The proposed goal is to detect fraud from a fraud history log. It should also be aware of the need to extracting data so as to obtain a better understanding of those transactions that may result in fraud. A good assessment of the current bank situation is also very important, especially in which regards losses that fraud is causing to customers and the bank itself. After the implementation of the model, the evaluation should check if these losses were minimized. Also at the Business Understanding phase risk assessment and a project plan must be developed with the next steps for implementing the CRISP-DM process.

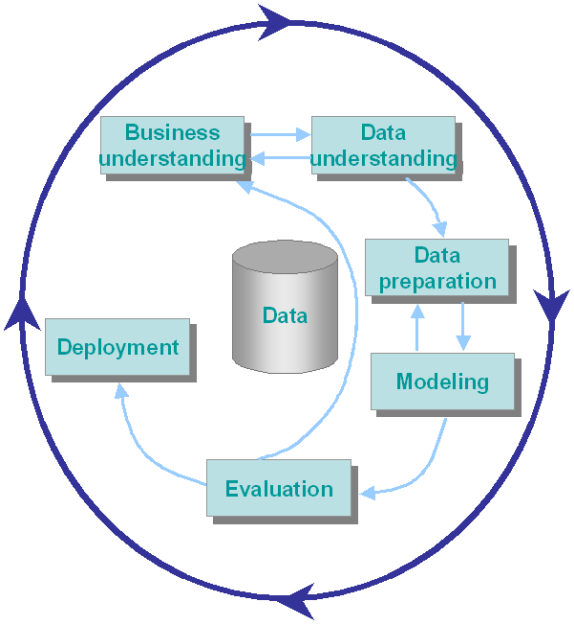

Figure 1. Phases of the CRISP-DM Process

3.3. Data Understanding

The second phase of the CRISP-DM is the Data Understanding. The initial data should be collected and a description of this data must be produced, as well as a verification of its quality. This is where the fraud history of the bank is synthesized, with the required attributes such as time of the fraud, the number of frauds, fraud types, and so on.

3.4. Data Preparation

The next step is aimed to prepare the data for import into fraud detection software, so this is the Data Preparation phase. In our case study, we are preparing data for use in the algorithms of decision trees. It is the phase to find calculated fields, incorporate external databases, perform a good data cleaning and classify the attributes as irrelevant, categorical and numerical.

3.5. Modelling

This phase uses modelling techniques on data that was prepared in the Data Preparation phase, so as to select, try and use an adequate modelling technique, such as neural networks. In our case study, we are using decision trees, using a database for training, validation and testing of bank frauds.

3.6. Evaluation

In this phase a checking procedure is performed to asses whether we have use the best tool for data mining and verifies that the data is really portraying the reality understood in the Business Understanding phase. If more processes are to be modelled, the process returns to the Business Understanding phase and reiterates the whole process.

3.7. Deployment

When we are ready with the design, the implementation is made in the Deployment phase. It requires that we must not forget to create artefacts in each preceding phase of the process so as to conduct training sessions with the users of the system.

With bank transaction logs being produced continuously and new frauds being forged in a rapid pace, a project of data mining does not last long and should always be updated. Information that is true today may not be tomorrow, since the data are very volatile and new types of fraud are always expected.

4. DECISION TREES

A decision tree is both a data representing structure and a method used for data mining and machine learning. This is the technique that is used in this paper for modelling frauds, during the CRISP-DM Modelling phase described in the previous section.

Let's assume a large amount of data and the need to classify them to find out answers on some subject. For this, we can use the concept of a decision tree as a model that maps the observations, taking into consideration a selected attribute as its starting point. The most difficult question here is to find the best attribute. The decision trees assist in this work of selecting the attribute that will develop a better performance in finding the required information.

The technique "Divide to Conquer" is used in decision trees, which consists in breaking the problem into simpler problems, and easier to solve [8]. Furthermore, strategies applied in a certain section of a tree can be applied recursively.

As illustrated in Figure 2, a decision tree is composed of the following parts [8],[11],[12],[13]:

1. Node – Contains a test of an attribute
2. Branch – Contains a response to each attribute
3. Leaf – Each leaf is associated with a class
4. Rule – Each route from the root to a leaf corresponds to a classification rule.

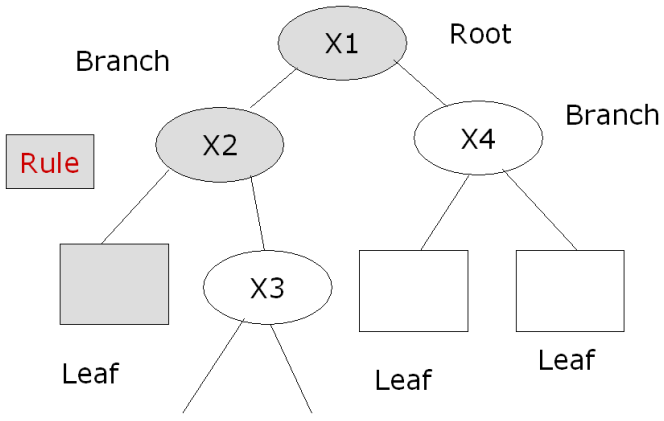

Figure 2. Structure of a Decision Tree

4.1. Methods of representation of a tree

The representation of a tree can be made as follows [8], [9], [10]:

1. Select an attribute.
2. Extend the tree by adding a branch for each value of this attribute.
3. Show the examples in the leaves.
4. For each leaf, if all examples are from the same class, add this class to the leaf. If not, repeat steps 1 to 4.

Now the question is to determine which attribute is going to be the first chosen. We must choose first the ones that have the best information. Decision trees use the concept of entropy to test how much information has an attribute.

From the information theory field, we can define entropy as a measure of randomness of a variable. By means of this concept it is possible to measure whether an attribute is explicitly or not a good one. If there are *n* possible messages with equal probability then the probability *p* for each one is *1/n*, and so:

$$-\log(p) = \log(n).$$

Now, for a distribution of probabilities $P = (p1, p2... pn)$:

$$I(P) = -(p1 * \log(p1) + p2 * \log(p2) + ... + pn * \log(pn)).$$

For example:

- If *P* is (0.5, 0.5) then *I(P)* is 1.
- If *P* is (0.67, 0.33) then *I(P)* is 0.92.
- If *P* is (1, 0) then *I(P)* is 0.

In these examples, we can see that when the distribution of the probability is higher, then we have better information regarding this variable. This is the basic property at the time of selection of attributes in a decision tree.

The entropy is used to estimate the randomness of the variable to predict: the class. The gain of information measures the reduction of entropy caused by partitioning the examples according to the values of the chosen attribute. We define the gain of information as follows:

$$GAIN(attribute) = I(attribute) - I(specific\ attributes).$$

For example, an attribute can be wind and specific attributes sunny, rainy and cloudy. After the calculation of the wind attribute, one should also perform the calculation for the other attributes. The attribute that the GAIN operator shows to provide better information will be used first.

The gain of information is used to create small decision trees that can identify the answers with a few questions. The preference is given to the simplest answer according to the Occam's razor principle.

4.2. The Algorithms ID3 and C4.5

ID3 and C4.5 are algorithms introduced by Quinlan for induction of classification models (for decision trees). Usually, we use attributes that contain values only "true, false" or "success, failure". The difficulty is to know which attribute is used first so that the tree gets to the solution as quickly as possible, with the best performance in the time of search [8].

In ID3 and C4.5, each node corresponds to an attribute not categorized. And each leaf corresponds to each value of the attribute. It is the same concept of a decision tree. Each node is associated with an attribute not categorized with most of the information possible provided from the root. That is what we call a good decision tree.

Entropy and Information Gain are used to test how much information has a node. The algorithm works as follows [8], [9], [10]:

1. Check the cases from the database.
2. For each attribute, calculate the "Gain of Information", using the concept of entropy.
3. Create a decision node at the node that has the most information gain.
4. Use the resources of this node in other parts of the tree.

5. IMPLEMENTATION FOR BANK FRAUD PREVENTION

The tests were made using the software Weka (Waikato Environment for Knowledge Analysis) [14], which contain a lot of classification algorithms.

In these tests, the observed parameters were the duration of algorithm execution and the percentage of errors in the phase of training.

For testing, we used a database containing bank frauds and real transaction data from a financial institution. This data was obfuscated so that no legal status and no information are revealed that could compromise the security of any institution, being this data used only for academic purposes. Furthermore, the attributes of this database have their names camouflaged. The database has 17.753 records.

5.1 Using Decision Trees

For the application in a decision tree, the algorithm used was J48, which is an evolution of C4.5. Here, the training algorithm took only 0.34 seconds. The network showed a rate of 7.9536% errors and accuracy of 0.956 for the correct cases, where there is no fraud, as well as 0.568 for the cases where there is fraud.

```
Number of Leaves  :     191

Size of the tree :      240


Time taken to build model: 0.34 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       16341               92.0464 %
Incorrectly Classified Instances      1412                7.9536 %
Kappa statistic                          0.5263
K&B Relative Info Score             414123.9298 %
K&B Information Score                 2062.4857 bits      0.1162 bits/instance
Class complexity | order 0            8839.2672 bits      0.4979 bits/instance
Class complexity | scheme             4954.8602 bits      0.2791 bits/instance
Complexity improvement     (Sf)       3884.4071 bits      0.2188 bits/instance
Mean absolute error                      0.1176
Root mean squared error                  0.2425
Relative absolute error                 60.3721 %
Root relative squared error             77.7063 %
Total Number of Instances            17753

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.975     0.521      0.938     0.975      0.956    0
  0.479     0.025      0.699     0.479      0.568    1

=== Confusion Matrix ===

     a     b   <-- classified as
 15411   401 |     a = 0
  1011   930 |     b = 1
```

Figure 3. Results of Training a Decision Tree for Bank Fraud Detection

3. CONCLUSIONS

In today's world, the decisions taken by experts and practitioners from many different branches of activity must be fast, accurate and with the possible lowest level of problems caused by these decisions. Notwithstanding this fact, due to the complexity of factors and methods, specialists are prone to making incorrect conclusions in their work. For example, a bank fraud may go unnoticed by a specialist. So the need arises to help the specialist with tools and methods that give some level of certainty to the decisions that should be taken.

A decision tree offers the capacity to make classifications, with the help of mathematical concepts, specifically of entropy, studied in information theory, allowing an algorithm to mathematically calculate the randomness of variable regarding the possible choices, thus reducing the difficulty of precisely attaining the goal decision. But, we should be able to create small decision trees so that the goal is reached in a few questions and as quickly as possible.

Specifically when the bases are numeric, they can generate huge trees that have a difficult analysis. Scenarios that require quick responses, like bank fraud logs, can not be used with applications that have a high delay. Based on the implementation of our proposed decision tree, and the test results on a sample real database, we conclude that the decision trees with a criteria for data mining help in decision making, especially in the handling of large data.

REFERENCES

[1] QUINLAN, J. R. (1988). C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA.

[2] Kou, Y., Lu, C., Sirwongwattana, S., & Huang, Y. (2004). Survey of Fraud Detection Techniques. Proc. of the 2004 International Conference on Networking, Sensing, and Control, 749-754.

[3] Phua, C., Lee, V., Smith-Miles, K. and Gayler, R. (2005). A Comprehensive Survey of Data Mining-based Fraud Detection Research. Clayton School of Information Technology, Monash University.

[4] Bolton, R. & Hand, D. (2002). Statistical Fraud Detection: A Review (With Discussion). Statistical Science 17(3): 235-255.

[5] Weatherford, M. (2002). Mining for Fraud. IEEE Intelligent Systems July/August: 4-6.

[6] CRISP-DM. Available at http://www.crisp-dm.org . Accessed in 1 July 2010.

[7] CUNHA, R. The CRISP-DM Process Model. Available in http://www.cin.ufpe.br/~compint/ aulas-IAS/kdd-042/AulaCRISP-DM-OK.ppt. Accessed in 25 July 2010.

[8] INGARGIOLA, G. Building Classification Models: ID3 and C4.5. Available in: http://www.cis. temple.edu/~ingargio/cis587/readings/id3-c45.html. Accessed in 20 may 2010.

[9] QUINLAN, J. R.: C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[10] C4.5 Algorithm - Disponible in http://www.cse.unsw.edu.au/~quinlan . Accessed in 25 July 2010.

[11] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[12] BERNARDES, R. M. C4.5: Um recurso para geração de arvores de decisão. Available in http://www.cnptia.embrapa.br/files/INSTRTECNICAS7int.pdf . Accessed in 20 May 2010.

[13] GAMA, J. Arvores de Decisão. Available in http://www.liaad.up.pt/~jgama/Bdc/arv.pdf. Accessed in 20 May 2010.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

Authors

B. C. da Rocha was born in Brasília – DF, Brazil, on July 03, 1983. He graduated in Computer Science, Centro Universitário de Brasília, Brasília – DF, Brazil, 2005, and got his Specialist Degree in Technology Information Management, University of Brasília, Brasília – DF, Brazil, 2008. He is following the Master Degree in Electrical Engineering at the University of Brasília, Brasília – DF, Brazil. He works in Banco do Brasil as a systems analysis. His field of study is Network and Information Security. His fields of interest and research include Artificial Intelligence, Network Security and Cryptography.

R. T. de Sousa, Jr., was born in Campina Grande – PB, Brazil, on June 24, 1961. He graduated in Electrical Engineering, Federal University of Paraíba – UFPB, Campina Grande – PB, Brazil, 1984, and got his Doctorate Degree in Telecommunications, University of Rennes 1, Rennes, France, 1988. His field of study is Network Engineering, Management and Security. His professional experience includes technological consulting for private organizations and the Brazilian Federal Government. He is a Network-Engineering Professor at the Electrical Engineering Department, University of Brasília, Brasília – DF 70910-900 Brazil, and his current research interest is trust and security in information systems and networks.