

FRAUD DETECTION IN ELECTRIC ENERGY USING DIFFERENTIAL EVOLUTION

ANGELO DARCY MOLIN*, JOÃO ONOFRE PEREIRA PINTO*, ALEXANDRA MARIA ALMEIRA CARVALHO PINTO*, LEANDRO SAUER*, EVANDO COLMAN†

**Federal University of Mato Grosso do Sul,
Campo Grande, MS, Brasil,
79074-460, Campo Grande, Mato Grosso do Sul, Brasil.*

† *ENERSUL - Mato Grosso do Sul Electrical Energy Distribution Company
Campo Grande, MS, Brasil.*

Emails: `angelo@batlab.ufms.br`

Abstract— This work proposes the use of differential evolution algorithm to find the parameters of a data mining system used to pre-select electrical energy consumers with suspect of fraud. A pattern recognition system was built in order to identify suspicious behavior of electrical energy consumers. However, the system only indicates such clients, and the frauds must be confirmed through in locus inspection. For that reason, it is important that true alarms be high to justify the trade-off of the in-locus inspection. Therefore, the parameter of the pattern recognition system must be well tuned, and that can be modeled as an optimization problem using the available training data. This work describes the pattern recognition system in details, and shows the algorithm modeling as an optimization problem. The differential algorithm will be described and results will be show. Results confirm that this approach is feasible.

Keywords— Differential Evolution, Fraud Detection, Electrical Energy Consumers

1 Introduction

Electric energy frauds bring high financial losses to the concessionary. With this problem in mind, the study of profile identification of fraudulent consumers was made. Data mining techniques were used to identify these profiles. The objective of this work is to identify the optimal point of step/ramp. A step is a suddenly fall in electric energy consumption, while a ramp is a soft fall. Both, step and ramp, change the consumer profile, bringing suspicion, the reason for suspicion is because before the fraud it is known that there is a change of behavior, and throughout step and ramp it is possible to control accuracy, hit probability, and coverage, which is the percentage of the quantity of fraudsters consumers indicated for the system, so we can maximize the rate of return of the concessionary.

The fast growth of databases creates a need and an opportunity to extract knowledge of them. In the 80's, a new branch of computing was created, KDD - Knowledge Discovery in Database, with the major objective of finding one methodology to exploit these databases, and to recognize existing profiles using modeling of real-world phenomena.

Fayyad created steps to knowledge extraction: definition of the problem, selection of attributes, cleaning and pre-treatment, processing of data, data mining, tests and analysis.

The step that calls more attention is the datamining because it is in this step that the technique is applied, using statistical methods, artificial intelligence, and others. This work will use differential evolution technique, which will be

responsible to determine step and ramp optimal points, reconciling accuracy and coverage.

The use of differential evolution as data mining technique decreases computing cost, because this algorithm evolves from generation to generation, until the optimal solution, or near optima is found, even do not all the existents possibilities have being covered.

Differential evolution is a good tool for problems that have large quantities of data, time being one variable that is considered.

2 KDD and Differential Evolution

2.1 KDD - Knowledge Discovery in Database

Fayyad created one methodology based in steps for knowledge discovery in database (Fayyad et al., 1996), The steps are:

Definition of Problem: It has the objective that is wanted to achieve. It's necessary to limit the other steps.

Select of data: The task here is to determine which are the attributes that influence the attributes of the output. Here, help from the specialist of the area in question is required.

Clean and pre-treatment: The objective of this step is to eliminate redundancy and prepare the data for the data mining step, because to do the profile identification, each algorithm requires data availability in an specific form. This is the step that needs a longer period, because the knowledge extraction depends of how these data are organized and representing the real world. It is necessary to guarantee the consistency, so the search can reach satisfactory results that represent the

real world through out rules that will be find in the step of mining.

Processing of data: In this step, new conditional attributes can be created from existing ones, such as: medium, summations, statistic, etc.

Data mining: This step calls more attention, because it is where the algorithm will be applied to find the rules that applies to the conditional attributes that maps the decision attributes.

Testis and analysis: This step is needed to qualify the search gain, and to present the results that were found, it is here that it is possible to monitor the effectiveness. If the objective is not reached, it is required to review all the steps before, to see where the error can occur, or to conclude that it isn't possible to extract knowledge of the database, because many reasons: inability of the researchers, problems that don't have solution, data that don't represent the real world, etc.

2.2 Differential Evolution

This algorithm works with the idea that it is possible to arrive at the optimal or near optimal solution through the evolution of one randomly chosen initial population. It's based on Darwin's natural evolution, where he identified that beings more suited to the medium have more surviving probability, which he called natural selection. See that beings inherit features of the parents, if they inherit the best features, they can be more adapted to the medium. Also, through the work it can be seen that mutation happens in live beings, making the selection better or worse.

Differential evolution is the technique of the evolutionary algorithm proposed by Storn and Prince (Storn and Price, 1997), from one initial population randomly chosen through of uniform distribution, except in some cases. The differential evolution tries to evolve for the better solution.

First, one initial population is created, being the operators make the population evolves from generation to generation. When optimal parameters are almost reached, new vectors are generated through the addition of the weighted difference between two population vectors to a third vector, this operation is called mutation. One vector, the trial vector, receives characteristics inherited from the vector derived from the mutation, and the target vector through the pre-determined probability, this operation is called crossover. If the target vector returns the cost lower than the one of the trial vector, the target vector is chosen for the next generation, on the contrary, the trial vector is chosen. Each vector of populations has to serve once as a target vector in each generation.

Each target vector of $x_{i,G} = 1, 2, 3, \dots, NP$, where x is the population vector; three index

are randomly chosen, α, β, γ and represent vectors $x_\alpha, x_\beta, x_\gamma$ of population in generation G, all of them different for each order, and also different from the target vector. Therefore, for this model, the population have a minimum of four individuals. The mutant vector can be write as:

$$v_{i,G+1} = x_{\alpha,G} + F \cdot (x_{\beta,G} - x_{\gamma,G}) \quad (1)$$

Where, $(x_{\beta,G} - x_{\gamma,G})$ is the difference, and F is the weight, also called weight difference. F can be between $[0,2]$, the control of magnitude of the difference.

Crossover: Because the algorithm has the diversification power, the crossover is introduced, where the vector derived from the mutation is mixed with the target vector, creating the trial vector. The crossover is generated according to:

$$u_{i,G+1} = \begin{cases} v_{i,G+1}, & \text{se}(rand \leq CR) \text{ ou } j = rnbr(i); \\ x_{i,G}, & \text{se}(rand > CR) \text{ e } j \neq rnbr(i). \end{cases} \quad (2)$$

Where $j=1,2,\dots,D$, where D is the quantity of features of each individual of the population. CR is the probability of the trial vector inherit the features of the mutation vector or target vector, being between $[0,1]$, and rnbr ensures that the new individual receives at least one feature of the mutation vector.

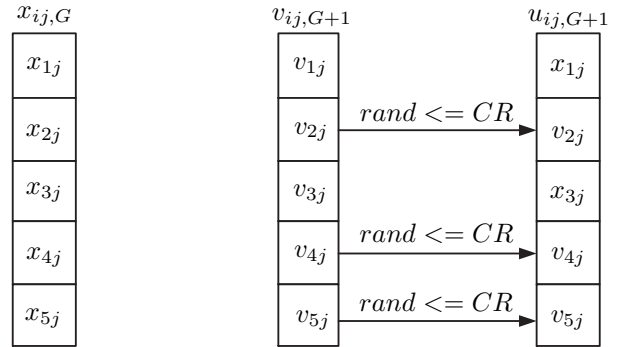


Figure 1: Crossover

Selection: This operation compares the trial vector with the target vector, throughout a cost function. The individual more adapted to the medium will be chosen for the next generation.

3 Codification of the Problem

The objective of this work is to find the optimal step and ramp values. Both parameters directly indicate the fraudulent electrical energy consumers, because they are indicators of behavior change. A line is necessary, since, if there is a huge demand for hits, all fraudulent consumers will not be indicated. In other words, the higher the accuracy, the lower the coverage.

Therefore, the following logic is take: if the concessionary manager wants the maximum number of fraudulent consumers, the error rate is very high, if a higher number of hits is wanted, not all fraudulent consumers will be covered.

So, using data available from the electrical energy concessionary, this work has to find a point that agrees with the manager needs and that has to be adapted to the daily fraud inspections, made in loco, which generates costs to the concessionary. To maximize the return, the system can't indicate all the fraudulent consumers, in this case the accuracy would be low, generating a lot of inspections, and would not be a financial return for the concessionary.

For solution, the problem was codified in such a way that using the desired accuracy it brings an optimal point of maximization of return.

The manager indicates a acceptable accuracy band, where infinite possible points of step and ramp exist. Therefore, it is necessary a technique that find the desired points without traveling to all the possibilities, so the differential evolution technique was used.

The collected data are real consumption data provided by the concessionaire. The inspections were made by previous studies, from which were set the values of step and ramp. This study will optimize these parameters. To find the optimal values we need a codification of the problem, such decoding should ensure the development of the differential evolution, so the cost function is very import, especially because it is the one embedded with intelligence.

The cost function was proposed so that it syn-tonize accuracy, coverage and manager of the system, always seeking to maximize the return. So, the system will go among the track that was chosen by the manager and will find accuracy with the greater coverage.

The system generates a random population, where each individual has a population of values of step and ramp, generating accuracy and coverage.

So, for each individual found within the band that was given by the manager, a weight is assigned. The following rule is proposed:

```

IF accuracy is within the range
  y = k + (accuracy-(mim(certainty)))
else
  y = -k + (accuracy-(mim(certainty)))

```

Y is calculated based on two terms. The first term is a constant value K that weights the solutions. The second term computes the difference between the accuracy found by the individual of the population and the lowest values of accuracy of desired range, as it can be seen in the figure 2:

Function y is not a *fitness* function, it does not have the degree of coverage embedded in it, so the degree of coverage is calculated as follow:

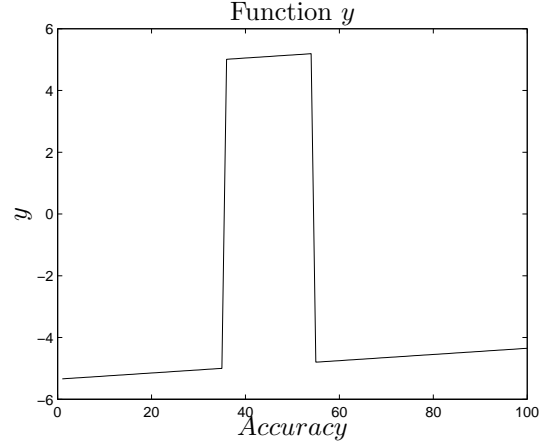


Figure 2: Function y for accuracy between 35% and 55%

$$cost = y + (0.8 \cdot (coverage)) \quad (3)$$

The cost function is dynamic for each value of step and ramp, so it will have a different coverage for each value of accuracy.

This function is tailored to each value that was desired by the manager. So, there is a convergence of the initial population through the following generations, until the point that has a controlled accuracy, maximizing the coverage, and then indicating a number of controlled suspected, ensuring a minimum number of mistakes and indicating the largest possible number of fraudulent consumers.

4 Results

The results are based on the real data of the local electrical energy concessionaire, and totalizes 14,273 customers, being 331 fraudsters. Through this study, it is possible to control the rate of success using step/ramp parameters. But, the higher the rate of success (accuracy), the lower the coverage, as seen in the results below.

The consumers were separated in classes: residential, commercial and industrial; and the tracks of connection in: single-phase, biphasic, three-phase; totalizing seven parameters for step and seven for degree, because there are not industrial consumers with tracks of connection single-phase or biphasic.

With degree of freedom between 25% and 35%, accuracy was 26.33% and coverage was 97.28%. So, the system indicates 322 fraudsters and 901 normal, better than the initial result. Showing lower coverage, proving that higher the accuracy the lowest the coverage.

With the degree of freedom between 30% and 40%, accuracy was 31.56% and coverage was 66.24%. Resulting 220 fraudsters and 477 normal consumers.

With degree of freedom between 40% and 50%, accuracy was 41.61% and coverage was 34.44%. With 114 fraudsters and 160 normal consumers.

As can be seen, the more it is required to hit the system, the number of suspects decreases, also the coverage decreases, so the concessionaire manager should manipulate and control the amount of information with the degree of certainty.

Analyzing the results, it appears that it is possible to control the margin of error, making the system fit the needs of the concessionaire.

If the aim is to identify the largest number of frauds in an attempt to eliminate them, the system indicates a high number of suspects, which results in high financial cost.

If the manager wants to minimize the costs of inspections, the rate of success should be high. However, it does not indicate all the fraudsters. So, a very small percentage of the population will be inspected, while the general population will see a minimum of inspections, and fraud can not be controlled, so there is a trend that it increases.

5 Conclusions

The solution to the problem of fraud identification through the history of consumption is complex because there are thousands of customers and also because the change of behavior in the electric energy consumption is natural. So, the improvement of each parameter becomes necessary, and the control of success through statistics from past inspections is a great tool, and this can be obtained from the knowledge that the concessionary manager, which is the estimate of the return of inspections that were carried out, fraudster or normal consumer.

The result of this research was satisfactory as it is possible to control the level of success in terms of coverage, thereby maximizing the return, and then developing a method of fraud controlling, and eliminating with time the commercial losses suffered by concessionary of electric energy.

The system should be used as a tool of control, because the random search of fraudsters generates a minimum success. With the system in operation, the company can control the amount of inspections, success and coverage, and still is a great system for detecting fraud.

References

- Arantes, M. B., da Silva Oliveira, G. T. and Sarago, S. F. P. (2006). Evolução diferencial aplicada a alguns problemas da engenharia de produção, *FAMAT em revista* **6**: 48–61.
- Cabral, J. E., Pinto, J. O. P., Gontijo, E. M. and Reis., J. (2004). Fraud detection in electrical energy consumers using rough sets., *2004 IEEE International Conference on Systems, Man, and Cybernetics.*, Vol. 4, pp. 3625–3629.
- Chan, P. K., Fan, W., Prodromidis, A. L. and Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection, *Intelligent Systems, IEEE [see also IEEE Expert]* **14**(6): 67–74.
- CODI (1998). Perdas comerciais, *Technical Report 08-05*, ABRADEE: Associação Brasileira de Distribuição de Energia Elétrica.
- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine* **17**: 37–54.
- Storn, R. and Price, K. (1997). Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces, *Journal of Global Optimization* **11**: 341–359.