https://dx.doi.org/10.22161/ijaers.73.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

Applying Data Mining to Support in the Extrajudicial Inspection

Harly Carreiro Varão[1], Marcelo Lisboa Rocha[2], David Nadler Prata[2]

[1]Developer at Tocantins Court of Justice (TJTO) and master's student at PPGMCS, UFT, Brazil
[2]Post-graduate program in Computational Modelling of Systems (PPGMCS), UFT, Brazil

Abstract**—***The Tocantins Court of Justice (TJTO) - Brazil has achieved high levels of computerization of its processes, whether in the judicial or extrajudicial areas. This scenario brings with it the need for transparency in carrying out such procedures. With regard to extrajudicial service, analyzing the data resulting from inspections of extrajudicial services are aspects that need attention. In this sense, a data mining technique based on association rules was proposed to analyze the data resulting from on-site extra-judicial inspections. As in general, the number of association rules is very large, a second step was taken in order to optimize/reduce the number of rules that represent the data set. Computational tests were carried out with other classic techniques of the literature, such as decision trees, Support Vector Machine and Naive Bayes, where the proposed technique performed better.*

Keywords**— Association Rules, Data Mining, Extrajudicial Inspection, Optimization.**

I. INTRODUCTION

Law and Information Technology (IT), two areas studied by humanity and which, despite great resistance, converged at some point in the 20th century towards mutual evolution. Historically, the Law appears in a certain society when its sources are evidenced as its generating fact, and among these sources is the custom. Costa et. Al. [1] argues that customs are characterized as the most important and oldest source of law, being defined by the repeated behavior of a certain social group. This means that a law or legal decision arises only after the Law generated and evidenced through direct, daily and habitual expression of the social group in which they are inserted.

The years have passed and the Law and Information Technology have increasingly narrowed their relationship. Justice, which is often referred to because of its slowness and little change in its legal provisions, has lived days of computerization of its judicial processes, unification of systems, communication between courts of different spheres and a decrease in the physical role. This makes the state courts, which deal in the first instance with access to justice, turn their IT teams towards a reality of constant technological innovation.

It is in this context that the Tocantins Court of Justice (TJTO) is inserted in the effectiveness of judicial and extrajudicial jurisdictional provision. One of the bodies that contributes to this is the General Internal Affairs Department (CGJUS) of the State of Tocantins, which acts in the control, guidance and inspection of the judicial and extrajudicial services provided within the state of Tocantins. With regard to extrajudicial, CGJUS periodically carries out some inspection procedures in which the notary is subjected to the confrontation of what is determined by the rules established by the competent bodies. These procedures, called extrajudicial inspection, must be carried out in person in all 273 active services spread over 139 municipalities in the state of Tocantins. This demonstrates a huge amount of information for a small personal contingent available, which leads to more intermittent in-person extrajudicial correction cycles and directly impacts your results.

This work proposes to facilitate the process of inspection and inspection carried out by the CGJUS with regard to the extrajudicial inspections. This is a massive, manual process and often has very small results in view of the amount of data available to be analyzed and, in case of divergence, apply the necessary corrections. However, this scenario can be changed, bringing speed to the process, with the help of data mining and mathematical tools that obtain relevant information from a given data set. It can be used as an ally in the optimization of procedures for guidance, control and inspection of services through the competent agencies.

https://dx.doi.org/10.22161/ijaers.73.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

The data mining technique used was Apriori method, to find association rules to characterize the data set. As the number of association rules generated is very huge, a second step, using mathematical programming was used to optimize this number of rules. Making feasible for a human expert to analyze this smaller number of rules and verify if they are relevant to the problem in hand, before applying them in practice.

Here, were executed computational experiments and statistical tests over the results of some well-established data mining techniques as Decision Tree, Support Vector Machine (SVM) and Naïve Bayes against the proposed method, which proved to perform better.

II. THEORETICAL FOUNDATIONS

This section will present the theoretical foundations regarding association rules and the problem of covering sets that are fundamental in this work.

2.1 Association Rules

The generation of association rules is a data mining technique used to find useful and valuable information in large databases [2]. Mining of association rules according to Agrawal et al. [3] is generally defined as follows. Given $I = \{i_1, ..., i_n\}$ as a set of items (attributes) and D as a database, each data line consists of a subset of items of I. An association rule is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$. The rule $X \rightarrow Y$ occurs in D with confidence c if c% of D data that occurs in X also occurs in Y. The rule is supported s in D if s% of data in D contains $X \cup Y$ (X and Y occur simultaneously in s% of D data). The problem of mining association rules is the generation of all association rules that have greater support and trust than those specified by the user.

Mathematically, support and trust can be defined as:

$$\text{(support) } s(X \rightarrow Y) = \frac{\tau(X \cup Y)}{N} \quad (1)$$

$$\text{(confidence) } c(X \rightarrow Y) = \frac{\tau(X \cup Y)}{\tau(X)} \quad (2)$$

where τ (•) is the number of occurrences of D data lines containing the specified set of items, and $N = |D|$ is the number of lines in the database.

Therefore, considering the weather.nominal database, which is an example toy (5 attributes and 14 data lines) available at the UCI Machine Learning Repository [4], shown in Figure 1, and considering, for example, the humidity = normal windy = FALSE → play = yes ($X \rightarrow Y$) rule, we have the following values for support (s) and confidence (c), as specified in (1) and (2):

$N = 14$, $\tau(X \cup Y) = 4$, $s(X \rightarrow Y) = 4/14 = 0{,}2857$ and $c(X \rightarrow Y) = 4/4 = 1{,}0$

@attribute outlook {sunny, cloudy, rainy}
@attribute temperature {hot, medium, cold}
@attribute humidity {high, normal}
@attribute windy {YES, NO}
@attribute play {yes, no}
@data
sunny, hot, high, NO, no
sunny, hot, high, yes, no
cloudy, hot, high, NO, yes
rainy, medium, high, NO, yes
rainy, cold, normal, NO, yes
rainy, cold, normal, yes, no
cloudy, cold, normal, YES, yes
sunny, medium, high, NO, no
sunny, cold, normal, NO, yes
rainy, medium, normal, NO, yes
sunny, medium, normal, yes, yes
cloudy, medium, high, yes, yes
cloudy, hot, normal, NO, yes
rainy, medium, high, yes, no

Fig. 1: Description of weather.nominal database.

Source: Dua & Graff [4]

A typical association rules mining algorithm works in two steps. The first step finds all the largest itemsets (set of items) that satisfy the minimum support constraint. The second step generates the rules from all major itemsets that satisfy the minimum confidence constraint.

2.2 Set Covering Problem

The Set Covering Problem (SCP) is a well-known combinatorial optimization problem, with a variety of applications in different research fields [5] and [6].

The description of the SCP adopted here is that defined in [7], as shown below: given a finite set $K = \{1, 2, \cdots, m\}$, of m elements and the family $J = \{S_1, S_2, \cdots, S_p\}$ of subsets of K, the SCP aims to find a subset of minimum size $T \subseteq J$, such that all members of K are covered by members of T with a minimum total cost, that is, for each $k \in K$, there is at least one $S_j \in J$, such that k is covered by S_j. Let $A = (a_{kj})$ be an m x p matrix, such that the j-th column is the characteristic vector of the subset S_j, that is, $a_{kj} = 1$, if k is

https://dx.doi.org/10.22161/ijaers.73.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

covered by S_j, otherwise, $a_{kj} = 0$. Thus, the formulation of the SCP's entire binary programming (0-1) is

$$\text{Minimize } \sum_{j=1}^{p} x_j \quad (3a)$$

Subject to

$$\sum_{j=1}^{p} a_{kj} x_j \geq 1, \text{ for } k = 1 \text{ to } m \quad (3b)$$

$$x_j \in \{0,1\}, \text{ for } j = 1 \text{ to } p \quad (3c)$$

The variable x_j receives the value 1 if the subset S_j is selected in the coverage of the set, otherwise it receives 0.

2.3 Technique Developed to Minimize the Number of Mined Association Rules

This section presents the developed technique, based on mathematical programming. In this case, it is considered that the Apriori algorithm has already been executed on the data set *D* and the set of association rules *R* has been generated. Here, the problem in finding the smallest number of rules that cover the entire data set is the one specified, according to equations (3a) to (3c).

The relationship between a set of association rules and the set coverage problem is as follows. The SCP mathematical programming model presents *m* lines as a constraint, where each of the *m* lines represents a line of data from the data set and x_j represents each of the association rules generated and $nr=|R|$ is the number of rules. In this model, the variable $a_{kj} = 1$, if rule *j* covers data line k, and $a_{kj} = 0$, otherwise. Thus, we seek to find the least number of association rules x_j that cover all data lines, which in this case is the optimal solution. This solution will henceforth be called LPI, as defined in equations (4a) to (4c) as follows.

$$\text{Minimize } \sum_{j=1}^{nr} x_j \quad (4a)$$

Subject to

$$\sum_{j=1}^{p} a_{kj} x_j \geq 1, \text{ for } k = 1 \text{ to } m \quad (4b)$$

$$x_j \in \{0,1\}, \text{ for } j = 1 \text{ to } nr \quad (4c)$$

III. EXPERIMENTS AND COMPUTATIONAL RESULTS

Built in order to minimize the number of association rules necessary to cover the data set, the method presented in section 2.3 was implemented in the Java programming language 1.8.0_111 and in the solution of the entire linear programming model (exact method LPI), the GLPK API (GNU Linear Programming Kit) version 4.65 was used, which is a library of routines widely used to solve large-scale linear programming problems [8]. Among the available techniques, the exact Branch-and-Cut method was used to solve the problem.

In order to evaluate the performance and robustness of the technique proposed in this work (LPI), the results obtained were compared to those of mining techniques J4.8 (Decision Tree), SVM (Support Vector Machine) and Naive Bayes (NB). It is worth mentioning that all experiments were performed on a machine with an I7-4500U processor with 8GB of RAM.

3.1 Database of the Judiciary Used

This base was obtained through the terms of correction visits carried out between January 2015 and December 2018 in an entire region that covers an estimated population of 101,887 inhabitants by 2018. This represents dealing with data from 12 extrajudicial services divided into 5 municipalities in the state. The data is composed of acts performed in the services in any period, as well as the books found in the registry offices and the analysis of the fundamental requirements defined by CGJUS and CNJ. This resulted in a total of 1055 items inspected, each equivalent to a record in the database.

The classification attributes (Table 1) were established together with an analyst in the area of Registry and Notarial Law, these being: correction item, description of the error, if it has a stamp and the occurrence of an error. The analyzed fields were also chosen considering not to identify the services or to list possible monetary values.

Table 1: Data dictionary of classification attributes

Attribute	Description	Type
inspec_ item	It concerns the item that the analyst inspected during the inspection	Text
error_ desc	Reports the description of the	Text
has_ stamp	Presence or absence of stamp	Boolean
error	Presence or absence of error	Boolean

3.2 Computational Results of the Proposed Method

In order to obtain the minimum number of association rules that cover the entire database, in this specific case, the complete database (1055 lines), as specified in section 3.1, the Apriori algorithm was first run with the following parameters: confidence = 0 and support = 0.0009 (1/1055). These parameters were defined so that, the Apriori algorithm generated all possible rules for the data in question.

Initially, the Apriori algorithm generated 218 association rules in total. The exact LPI method obtained

https://dx.doi.org/10.22161/ijaers.73.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

100% accuracy, correctly classifying all instances and providing a reduced number of 3 rules that cover the entire database, which are: 122, 162 and 173. The description of each of these rules is specified next:

122. error_desc=no_error 650 ==> error=no 641 conf:(0.99)

162. has_stamp=yes 195 ==> error=yes 139 conf:(0.71)

173. has_stamp=no 677 ==> error=yes 275 conf:(0.41)

However, with expert analysis, rule 162 does not logically match the reality of the problem. Thus, it was excluded from the total set of 218 rules and the exact LPI method was executed again, obtaining another 3 rules that cover the entire database, which are: 122, 170 and 173. The description of one of these rules is specified at follow:

122. error_desc=no_error 650 ==> error=no 641 conf:(0.99)

170. inspec_item=stamp_act 471 ==> error=yes 231 conf:(0.49)

173. has_stamp=no 677 ==> error=yes 275 conf:(0.41)

Again, with the expert's analysis, one of the rules (170) does not logically match the reality of the problem. Therefore, this rule was excluded from the set of 217 rules and the exact LPI method was executed again, now obtaining 9 different rules that cover the entire database, which are: 4, 14, 17, 32, 38, 73, 92, 122 and 173. Rule 122 is implicit and its use is justified by the reason that the analysis covers the entire database in question. The description of each of these rules is specified below:

4. error_desc=stamp_not_found 96 ==> error=yes 96 conf:(1)

14. error_desc=incorrect_procedure 49 ==> error=yes 49 conf:(1)

17. error_desc=incomplete_data 36 ==> error=yes 36 conf:(1)

32. error_desc=unregistered act 17 ==> error=yes 17 conf:(1)

38. error_desc= divergent_value 12 ==> error=yes 12 conf:(1)

73. error_desc= non-corresponding_act 5 ==> error=yes 5 conf:(1)

92. error_desc=duplicate_stamp 2 ==> error=yes 2 conf:(1)

122. error_desc =no_error 650 ==> error=no 641 conf:(0.99)

173. has_stamp=no 677 ==> error=yes 275 conf:(0.41)

Now, according to the expert, these 9 rules, in addition to covering all data, providing 100% accuracy, are consistent with the reality of the problem and are useful for identifying acts practiced in extrajudicial services and that may contain some irregularity.

It is noteworthy that the exact LPI method, has an average execution time of the order of 0.08 seconds, for the amount of data presented, in addition to providing a reduction in the number of rules of the order of 91% (from 218 to 9), proving to be quite efficient.

3.3 Comparison of the Results of the Proposed Method with other Literature Methods

In order to prove the robustness of the exact LPI method proposed, in this section, performance comparisons will be made with the J4.8 (decision tree), SVM (Support Vector Machine) and Naive Bayes (NB) techniques, which are well established techniques in data mining literature, which are made available in the WEKA software package [9] and executed with the available standard parameters. It is important to note that in the comparisons with the other techniques there was no withdrawal of rules by the analyst.

The evaluation of the proposed technique and all the techniques in the literature were performed using the k-fold cross-validation procedure, with k = 10. Thus, 10 different classifiers were built.

Table 2 below shows the percentage of instances classified correctly and incorrectly by each of the techniques considered.

Table 2: Classification of instances by technique applied in the k-fold cross-validation procedure.

	LPI	**J4.8**	**SVM**	**NB**
Correctly classified instances	100%	97,79%	97,79%	96,87%
Incorrectly classified instances	0%	2,21%	2,21%	3,23%

From the data in Table 2, it is possible to observe that the proposed technique is superior to the others. However, additional statistical tests are carried out to confirm this hypothesis.

In order to provide a better comparison between the

https://dx.doi.org/10.22161/ijaers.73.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

proposed technique and the techniques in the literature, the AUC (Area Under Curve) for each one was calculated and then a hypothesis test to determine which of the techniques is superior to the others. AUC ∈ [0.5.1] provides an indication of the discriminating power of the model. The higher the AUC value, the better the model's ability to classify [10].

Thus, Table 3 presents the AUC values for each of the techniques used in the work, in order to assess their classification capacity for the presented database.

Table 3: AUC value calculated for each of the techniques used at work.

Technique	AUC
LPI	1,000
J4.8	0,989
SVM	0,989
NB	0,982

From Table 3, it is possible to observe that the technique proposed in this work (LPI based on the Apriori algorithm), has the best classification capacity. However, the other techniques also showed satisfactory results.

From the AUC values calculated for each technique, DeLong's hypothesis tests [11] were performed which verifies the difference between the AUC in order to allow the comparison of whether a technique has a better performance or not in relation to another for a given level of significance (α). For this test, if the P-Value is less than α, the null hypothesis is rejected (the difference between the AUC is equal to zero, that is, the two techniques in comparison do not show statistically significant differences).

Table 4: P-Value for hypothesis tests to compare performance between the techniques considered in this work.

	LPI	J4.8	SVM	NB
LPI	***	0,00245	0,00245	0,0004461
J4.8		***	1	0,08252
SVM			***	0,08252
NB				***

According to the P-Value values presented in Table 4, the following conclusions can be drawn: the proposed LPI technique is better than competing techniques in the literature with α = 5% and also with α = 1%; techniques J.48 and SVM present exactly the same performance; and that despite the techniques J4.8 and SVM present AUC better than NB, this difference is not statistically significant at α = 5%.

IV. CONCLUSION

The Tocantins Court of Justice has most of its activities converted to digital reality, which has led to extrajudicial jurisdictional provision at the same level in the services offered. Access to extrajudicial information enables greater transparency and engagement on the part of the citizen by contributing to the process of inspection and inspection of extrajudicial services.

Regarding the use of data resulting from in-person extrajudicial inspections to obtain relevant data through association rules, it is possible to state that the exact LPI method used in this work resulted in rules that are in accordance with the problem, being of valuable importance to identify acts practiced in extrajudicial registries that eventually contain any divergence with the current rule. In the end, when subjected to tests of comparison with some of the other existing techniques in the literature, the method based on the Apriori algorithm (exact LPI method) was shown to be superior to the others when obtaining greater capacity for classifying the rules. This method is even useful when subjected to virtual extrajudicial inspections, where the analyst can be suggested by the result of applying the technique when inspecting a particular service.

Regarding the association rules mining process, the minimum number of rules that characterize the database was obtained and assist in the task of identifying extrajudicial acts that potentially have a problem. However, this process can still be improved with the following future work:

- Make use of other measures of interest to the rules such as Lift, Leverage and Conviction in order to check if it is possible to find more meaningful rules without having to perform the LPI method repeatedly.
- Possibility of making use of fuzzy rules and values (fuzzy) in order to generate more general rules.
- Creation of a graphical interface in order to make the process of identifying and presenting the generated rules more friendly and easy to interpret for laypersons.

https://dx.doi.org/10.22161/ijaers.73.21 *ISSN: 2349-6495(P) | 2456-1908(O)*

REFERENCES

[1] Pietro Costa, Danilo Zolo and E. Santoro (2007) The Rule of Law History, Theory and Criticism (Law and Philosophy Library). Springer.

[2] Rakesh Agrawal and Ramakrishnan Srikant (1994). Fast algorithms for mining association rules in large databases. In Proc. of the VLDB Conference, Santiago, Chile.

[3] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami (1993). Mining association rules between sets of items in large databases. In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C.

[4] D. Dua, and C. Graff (2017), UCI Machine Learning Repository. Available in: <http://archive.ics.uci.edu/ml>.

[5] Z. Ren, Z. Feng, Z., L. Ke and Z. Zhang (2010). New ideas for applying Ant Colony Optimization to the set covering problem. Computers and Industrial Engineering, 58, 774–784.

[6] S. Al-Shihabi, M. Arafeh, and M. Barghash (2015), An improved hybrid algorithm for the set covering problem. Computer and Industrial Engineering, 85, pp. 328–334.

[7] S. Balaji and N. Revathi (2016), A new approach for solving set covering problem using jumping particle swarm optimization method. Natural Computing, Volume 15, Issue 3, pp 503–517.

[8] GLPK. (2018). Available in: https://www.gnu.org/software/glpk/

[9] Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal. (2016) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann - 4th ed.

[10] David W. Hosmer and Stanley Lemeshow (2000). Assessing the Fit of the Model. In Walter A. Shewhart and Samuel S. Wilks, editors, Applied Logistic Regression. John Wiley & Sons, Inc., Hoboken, New Jersey, chapter 5, pages 160–164. https://doi.org/10.1002/0471722146.ch5

[11] E. R. Delong, D. M. DeLong and D. L. Clarke-Pearson (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845.