“Who is their mother?”: A classification work to get answers over registration people databases

Gustavo C. G. van Erven*, Rommel N. Carvalho*, Maristela Holanda†, Marcelo Ladeira†, Henrique Rocha*,
and Gilson Mendes*
* Department of Research and Strategic Information (DIE)
Brazilian Office of the Comptroller General (CGU)
Brasília, Brazil
Email: {gustavo.erven, rommel.carvalho, henrique.rocha, liborio}@cgu.gov.br
† Department of Computer Science (CIC)
University of Brasília (UnB)
Brasília, Brazil
Email: {mladeira, mholanda}@cic.unb.br

***Abstract*—Discovering how people are related contributes to accurately assessing several scenarios within a criminal investigation. If two people have family ties, such as, brothers or cousins, for example, and one of them has been involved in criminal activities, there is a high probability that the other has also been involved in these same or other similar activities. However, in mapping relationships, certain variables arise that make it a difficult task: in some cases there are gaps in the information about people involved; tracking relevant facts on large databases to entertain several possibilities, and subsequently crosschecking these facts for accuracy, is cumbersome, often precarious and time consuming. To facilitate this effort, we have explored the problem of identifying relationships from registration forms in a database, starting with the mother because this role is the relational base of several others. We summarize the data in a proposal in which information is collected about two people and processed in order to attribute a similarity score for both their name and address. Then, we apply machine learning to classify their relationship based on these scores using two well-known algorithms: Support Vector Machines, based on hyperplanes; and Naïve Bayes, a model which uses conditional probability to classify the input. At the end of the process, we present a minimum set of derived attributes which contributes to get answers to this problem with the selected model. We anticipate that this model will help the specialists of the Brazilian Office of the Comptroller General (CGU) to identify collusion in the government.**

Keywords*—Parents Relationship, Corruption Control, Collusion Identification, Brazil, CGU

I. Introduction

Information about people is an important part of the data for private or public offices. From these datasets we can search profiles or clusters that can guide us to make better decisions about a product release, public policies, audits, or investigations [1].

From the perspective of intelligence, the knowledge about how people are related is as important as information about the target itself. People influence and are influenced by the individuals in social groups of which they are members and with whom they share interests and goals. So, identifying how people are related to each other within a social group is critical in the fight against corruption because this knowledge makes it easier to outline possible scenarios and infer what really happened. For example, when an audit starts, we can identify evidence of fraud in a process by the connections between agents who make decisions in a public office and other people who benefit from them, and are associated with them through strict ties.

Among the main possible groups that can be organized and threaten public resources through several mechanisms, family relationships stand out. People who want to hide goods and valuables from the law, can try to use some relative to help them, taking advantage of the natural organization and strength that this bond provides.

Therefore, in this study we have selected, among the several types of bonds, one which is fundamental and helps to derive other relationships - the mother-child bond. With this relation defined, we find offspring and possible fathers for the chosen targets. Although there are databases that have data about the relatives, these may not guarantee the credibility of the information through some kind of confirmation process. Sometimes it is incomplete, inconsistent, or even different instances with the same data values.

In this paper we investigate a technique to increase the reliability in classifying a pair of records in a database and use this information later to let specialists decide whether they are related. Section II presents how some papers handled the theme of family and crimes, such as corruption. In Section III, we summarize the methodology and techniques used in the paper. Section IV presents a description of the scenario that is carried out and tested. Section V describes the data collected and which features are used in Section VI to train and select the appropriate model. Section VII constitutes an analysis of the results. Finally, Section VIII proposes new perspectives where we conclude the paper and outline future works.

II. Related Works

Several studies address the relationship between family and corruption or other kinds of crime. Some of the studies show how family structure can negatively influence its members [2], [3]. Other studies do not specifically focus on the relationship, but on the problem of crime or corruption and at least one example addresses issues regarding activity of a relative or friend as collusion or nepotism [4], [5], [6]. Other kind of study proposes social network analysis concepts applied to the structure of organized crime [7], [8] and the objective of identifying the members and their roles in the network.

Considering these studies, the importance of identifying ties between individuals - improving our understanding of who it is that certain agents can influence and how they can influence them - is very clear. Although our bibliographic search did not return any studies specifically dealing with this theme, some of the papers focusing on criminology use the idea of similarity and kinship in their work [9], [10].

They have used name and hometown attributes to create links between people that share these features in different criminal networks. This idea can be extended using also the mother-son relationship. Besides, attributes available in other databases, besides criminal records, could be used to identify new links, such as owners of the same enterprise. Therefore, we believe, this work can help not only CGU, but also other offices in the government, such as the Federal Police.

III. Methodology, model and metrics

In this section, the methodology and tools used in this paper are briefly outlined. The process model, CRISP-DM served to guide the data mining project. To infer the class of the relationship - whether it is ”mother” or not two supervised techniques of the ten most popular data mining models algorithms[11] were used: The Naíve-Bayes[12], a statistical model and the Support Vector Machines (SVM)[13][1]. Both implemented in R software used in this paper. To compare results, we used the confusion matrix as well as some statistics described below.

A. CRISP-DM

The fundamental objective of CRISP-DM [14] is to provide a methodology to conduct data mining projects. The CRISP-DM is formed by the six phases as follows:

1) Business Understanding
2) Data Understanding
3) Data Preparation
4) Modeling
5) Evaluation
6) Deployment

Section IV focuses on Business Understanding, where the goals, requirements, and accepted terms are defined. In the Data Understanding phase we have verified which attributes are more relevant and can be used in the models. Section VI covers Data Preparation and Modeling phases of CRISP-DM. Information, such as names, have been transformed to scores with values between 0 and 1 to be more tractable in the models. Finally, the results are evaluated and the model that has the best fit over the business requirements is chosen as a final model. The deployment process has not been treated in this paper.

B. Measurements

The confusion matrix represents the quantitative set of instances correctly classify in its diagonal. These numbers are also called TP or True Positive, for mother, and TN or True Negative, for non-mother. The non-diagonal values are defined as FP or False Positive, for instances incorrectly classified as mother, and FN or False Negative, for values incorrectly classified as non-mother. In this paper we use F-Measure [15] for binary classification based on precision (see Equation 1) and recall (see Equation 2) as shown in Equation 3.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$FMeasure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3}$$

C. Datasets and Cross-Validation

Three data sets are used in this study. A 60% (55,091) group to training and the others 40% split in two groups of 20% (18,364): one to validate and to compare the models; and the other to test if the selected model can be generalized to unseen cases.

Another technique used was ten-fold cross-validation[15], which splits the data in folds, where one is used for testing while the rest is used for training. This process is repeated until all folds are used for testing.

IV. Scenario

The Department of Research and Strategic Information (DIE) of the Brazilian Office of the Comptroller General (CGU)[2] provide information to several departments to assist their works. Therefore, a large set of data has been used to extract useful information and generate reports.

Among daily and recurrent problems, identifying the people associated with a particular individual or enterprise is one that inevitably appears when we need to check information about public employees or bidding processes.

Results of corrupt activities are usually hidden, but the relationship between people can provide us with clues, as indirect links that guide us to the collusion where friends or relatives play a central role.

Identifying this relationship can be helpful even if it is only an estimate. The specialist can direct his search to a world smaller than he the one he started with, and save several hours of research. This is why the knowledge of relationships, as used in Social Network Analysis[17], is important. Since many relationships derive from the mother of relationship, we decided to focus our work on it. Later, we can expand this work to try to infer other relations, such as siblings.

This scenario was utilized as the base to start the study, which contributes to the work done in CGU. The scope has

[1] Although other algorithms can be applied, we chose these two in order to validate this proof of concept. In the near future, we will use other algorithms to learn new models in order to evaluate if we can improve performance.

[2] http://www.cgu.gov.br/english/default.asp

been defined to use the minimum information possible from two individuals. To be accepted, the model must have an F-measure equal or greater than 70%, because the methodology must be better than a manual triage. A system could present a limited collection of related names to the target, but we expect that only in 30% of the time the search would be useless to avoid a greater waste of time.

V. Data Understanding

When we were looking for what data to use on the model, the specialists helped us identify which information were important and available. The main database used has more than 200 million records. However, initially, we only use a fraction of this, in order to facilitate the verification process. Moreover, in order to guarantee efficiency and minimize the running time for learning and using the model, we have focused on a minimal number of attributes. We have “CPF”[3], “name”, “gender code”, “birth date”, “mother’s name”, “address”, “zip code”, “city”, and “federation unit” in the main table for person.

Although the mother’s name is in the table, there is no key to join with the exact person referenced by it. It is possible to work around if the name is unique in this base, but in other cases we cannot absolutely confirm the identity of the mother. Misspelling can occur as well, which can cause the return of an empty set when performing the join. Therefore, we do not use this field to explore the worse case scenario, *i.e.*, the absence of this information.

Beside the CPF field, which is only used to identify the instance, we analyzed all other attributes to define which ones should be used and how. Address and zip code are relevant attributes, according to the specialist. The fact that two people are living together is a good indication that they are friends or relatives. Although addresses and zip codes are similar in kind, addresses necessarily contain more specific information, because zip codes cover a broader range and are more general. However, one of the challenges of using addresses is misspellings, which is frequent with this type of data. Some people usually write their addresses with abbreviations, so the same information can be harder to join. One hypothetical sample is shown in Table I. These three cases are the same, but a query over them will return an empty set even when using upper case in all instances.

The attributes that supply the most important information for identifying relatives are the family names. Both mother and father usually give their last names to their children, thereby guaranteeing that the information will be passed on, as presented in Table I.

Cities and federation units are well defined and do not present problems in the main database. This information is then considered reliable, but it is more useful as a Boolean field that indicates if people are living in the same city, to avoid cases where we have two cities with the same address, but in different federation units.

Finally, the birth date can be used to measure the difference between the ages of two people, but this field has several errors or null values, so we chose not to use it. Indication of gender was removed as well because only women can be mothers and the distribution of gender for sons and daughters does not modify the fact of motherhood.

Table I: Sample Data.

TITLE	NAME
FATHER	FULANO SILVA SALGUEIRO
MOTHER	FULANA LIRA SALGUEIRO
SON	BELTRANO LIRA SALGUEIRO
ADDRESS	
R. Osvaldo Aranha 1500 CS10	
RUA OSVALDO ARANHA Numero 1500 CASA 10	
Rua Osvaldo Aranha N 1500 Cs 10	

With the first prune done, we obtained the attributes “CPF”, “name”, “address”, “city”, and “federation unit” that were used to derive other attributes.

VI. Defining the model

The analysis of raw data helps us understand what kind of preprocessing we need to do over the input. Initially, all character attributes were changed to upper case to avoid errors in comparative processes. The original database does not have accents, so we do not need to change this either.

The states (federation units) and cities have been grouped to avoid ambiguity and to generate a final attribute to represent whether these people are living in the same city.

The names and addresses were the attributes that went through the biggest transformation. We use the functionlevenshteinSim[4] to calculate a similarity for the pair of variables in each instance.

Table II: Similarity example for addresses and names.

FIELD 01	FIELD 02	levenshteinSim
R. OSVALDO ARANHA 1500 CS10	RUA OSVALDO ARANHA N 1500 CASA 10	0.788
R. OSVALDO ARANHA 1500 CS10	RUA RUI BARBOSA 1500 CS10	0.518
FULANA LIRA SALGUEIRO	BELTRANO LIRA SALGUEIRO	0.783
FULANA LIRA SALGUEIRO	CICLANA SILVA MONTEIRO	0.500

The Levenshtein distance [18] calculates the minimum number of changes needed to make two strings equal. It can be 0 when the inputs are already the same up to the size of the biggest string. The levenshteinSim function applies this algorithm with some transformation to generate a number between 0 and 1, where a value closer to 1 represents strings that present greater similarities as shown in Table II.

$$levenshteinSim = 1 - \frac{levenshteinDist(str1, str2)}{maxLenBetween(|str1|, |str2|)} \quad (4)$$

Equation 4 receives two strings as inputs, str1 and str2. After the distance has been calculated, the result is divided by the largest string length (maxLenBetween function) and subtracted by one. If the strings are the same, the distance will be 0 and the levenshteinSim will be 1. Otherwise, the function will return some value between 0 and 1.

[3] Brazilian social security number.

[4] http://cran.r-project.org/web/packages/RecordLinkage/index.html

Thus, the addresses and names were replaced by ranges between 0 and 1 in an attempt to avoid the slight differences that can occur in the records, and the family names were reduced to a simple score number. The first and last name could have been split, but this would have added more complexity, so this analysis was left for future work.

The final field to define the input instances is the attribute class, which was used to train and verify the model. The records with the relationship "mother" were tagged with "MOTHER_OF" and the others with "OTHERS". Therefore, we were able to validate the performance over the data and present the direction of the relationship (`Person 01` is "MOTHER_OF" `Person 02`). We have defined to the final instance's attributes as "proximity score for names", "proximity score for address", "lives in same city", and "class".

The class information was constructed using the data retrieved from SIAPE[5]. There is a table for people who receive a government pension due to the death of a relative who was a civil servant. We can identify these people using the CPF key. This data is submitted to complex legal processes that verify this information. So the positive class "MOTHER_OF" is loaded from the pairs that have this relationship in the SIAPE database.

Subsequently, we crossed the mothers with other people on the sons/daughters side (who are not relatives) to generate the "OTHERS". "MOTHER_OF" is our positive class, any combination of names which differs from this set is treated as a negative class. This is called the closed-world assumption and is "the idea of specifying only positive examples and adopting a standing assumption that the rest are negative" [15]. Other names from the main databases were also added with sons/daughters in the class "OTHERS", to increase the number of similar names on the mother's side.

The final sample achieved a total of 91,820 registers where 13,479 were in the class "MOTHER_OF" and was split in training, validation, and test sets keeping the proportion of the original distribution. To avoid imbalance, the training set was complemented with sample data from the same set and class "MOTHER_OF" going to 94,008 (each class having 50% in the end). After training the model, we checked it against the validation set.

Table III: Confusion Matrix for selected models.

<table>
<tr><th colspan="6">Confusion Matrix</th></tr>
<tr><th colspan="3">Naíve-Bayes</th><th colspan="3">SVM</th></tr>
<tr><td>Prediction</td><td>MOTHER_OF</td><td>OTHER</td><td>Prediction</td><td>MOTHER_OF</td><td>OTHER</td></tr>
<tr><td>MOTHER_OF</td><td>2109</td><td>1072</td><td>MOTHER_OF</td><td>2104</td><td>873</td></tr>
<tr><td>OTHER</td><td>587</td><td>14596</td><td>OTHER</td><td>592</td><td>14795</td></tr>
<tr><th colspan="6">Statistic Board</th></tr>
<tr><td colspan="2">Statistics</td><td colspan="2">Naíve-Bayes</td><td colspan="2">SVM</td></tr>
<tr><td colspan="2">F-Measure</td><td colspan="2">0.7177131</td><td colspan="2">0.7417592</td></tr>
</table>

Table IV: Hypothesis tests for selected models.

<table>
<tr><th colspan="2">Chi-Square with 3 Degrees of Freedom and $\alpha = 5\%$</th></tr>
<tr><td>X^2 = 39.7088</td><td>$X_\alpha^2 = 7.81$</td></tr>
</table>

Table III presents the confusion matrix for the applied algorithms and their performances. Both of them beat the fixed target and both with similar results as well. In order to verify if one model is significantly better than the other, we applied the Chi-Square hypothesis test where the null hypothesis is that there is no significant difference between the two models. Since the computed X^2 is greater than $X^2_.05$, as shown in Table IV, we have sufficient statistical evidence to refute the null hypothesis. Therefore, we decided to use the SVM model because it presents better performance than the NB model for this scenario. With the model defined, we can go on to test the data.

VII. Data analysis

After the model was selected it was applied over the test set to check its performance. Table V presents its confusion matrix and statistics. Since the results are good and similar to those obtained from the validation set, we can infer that this model can generalize to unseen data.

Table V: SVM - Data Test.

<table>
<tr><th colspan="3">Confusion Matrix</th></tr>
<tr><td>Prediction</td><td>MOTHER_OF</td><td>OTHER</td></tr>
<tr><td>MOTHER_OF</td><td>2118</td><td>943</td></tr>
<tr><td>OTHER</td><td>578</td><td>14726</td></tr>
<tr><th colspan="3">Statistics for SVM</th></tr>
<tr><td>F-Measure</td><td colspan="2">0.7357999</td></tr>
</table>

One las change we decided to perform was to remove the "lives in same city" attribute, since it does not provide a significant contribution to final performance. Thus, we were able to simplify and compact the input instances. The result of removing this attribute is shown in Table VI.

Table VI: SVM: data test without Lives in same City.

<table>
<tr><th colspan="3">Confusion Matrix</th></tr>
<tr><td>Prediction</td><td>MOTHER_OF</td><td>OTHER</td></tr>
<tr><td>MOTHER_OF</td><td>2116</td><td>955</td></tr>
<tr><td>OTHER</td><td>580</td><td>14714</td></tr>
<tr><th colspan="3">Statistics for SVM</th></tr>
<tr><td>F-Measure</td><td colspan="2">0.7338304</td></tr>
</table>

These derived attributes are better to achieve a good performance over such a large database. The results obtained here also serve as a baseline for other works on this theme. Even though we obtained good results with SVM, we pretend to perform new tests with other algorithms in order to verify if we can improve the results.

Ultimately, the present analysis shows that our objective in this study has been achieved. The F-measure beat the minimum required of 70% and we were able to reduce dimensionality by removing some and aggregating some attributes without compromising the results. In the near future, we pretend to incorporate this model in our daily analysis with some user-friendly interface, but this is beyond the scope of this paper.

VIII. Conclusion and Future Works

The information of relationships is very important in several different domain areas. In the criminal investigation scenario it plays a fundamental role in understanding how targets

[5] Federal system for public employees management

can influence or be influenced by the network. Therefore, this paper has focused on the basic bond of mothers to begin the studies in this filed.

Initially, we had several attributes and two models, Naíve-Bayes and SVM. Later, we checked which model presented better performance, which was the SVM model, since it obtained a higher F-measure in this scenario with a significant difference.

In future works, we can test other algorithms, besides expanding the model to several types of ties, such as fathers (which can be derived from other ties combined with mothers), until we achieve the identification of the entire family core and friends.

References

[1] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973. [Online]. Available: http://www.jstor.org/stable/2776392

[2] S. A. Cernkovich and P. C. Giordano, "Family relationships and delinquency," *Criminology*, vol. 25, no. 2, pp. 295–319, 1987. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1745-9125.1987.tb00799.x/abstract

[3] R. Agnew and T. Brezina, *Juvenile Delinquency: Causes and Control*, 4th ed. New York: Oxford University Press, Jul. 2011.

[4] R. Klitgaard, *Controlling Corruption*. University of California Press, 1988. [Online]. Available: http://www.jstor.org/stable/10.1525/j.ctt1pnj3b

[5] F. A. J. Ianni, *A Family Business: Kinship and Social Control in Organized Crime*. New York: Russell Sage Foundation, Jul. 1972.

[6] T. Gong, "Dangerous collusion: corruption as a collective venture in contemporary china," *Communist and Post-Communist Studies*, vol. 35, no. 1, pp. 85–103, Mar. 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967067X01000265

[7] J. S. McIllwain, "Organized crime: A social network approach," *Crime, Law and Social Change*, vol. 32, no. 4, pp. 301–323, Dec. 1999. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1008354713842

[8] J. Xu and H. Chen, "Criminal network analysis and visualization," *Commun. ACM*, vol. 48, no. 6, p. 100107, Jun. 2005. [Online]. Available: http://doi.acm.org/10.1145/1064830.1064834

[9] F. Ozgul and Z. Erdem, "Detecting criminal networks using social similarity," in *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2012, pp. 581–585.

[10] ——, "Which crime features are important for criminal network members?" in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2013, pp. 1058–1060.

[11] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Jan. 2008. [Online]. Available: http://link.springer.com/10.1007/s10115-007-0114-2

[12] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, no. 2-3, p. 131163, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1023/A:1007465528199

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995. [Online]. Available: http://link.springer.com/article/10.1023/A%3A1022627411411

[14] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "CRISP-DM 1.0 step-by-step data mining guide," The CRISP-DM consortium, Tech. Rep., Aug. 2000. [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf

[15] I. H. Witten, E. Frank, and M. A. Hall, *Data mining practical machine learning tools and techniques, third edition.* Burlington, Mass.: Morgan Kaufmann Publishers, 2011.

[16] L. Sales, "Risk prevention of public procurement in the brazilian government using credit scoring," OBEGEF - Observatrio de Economia e Gesto de Fraude & OBEGEF Working Papers on Fraud and Corruption, OBEGEF Working Paper 019, 2013. [Online]. Available: http://ideas.repec.org/p/por/obegef/019.html

[17] R. C. v. d. Hulst, "Introduction to social network analysis (SNA) as an investigative tool," *Trends in Organized Crime*, vol. 12, no. 2, pp. 101–121, Jun. 2009. [Online]. Available: http://link.springer.com/article/10.1007/s12117-008-9057-6

[18] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics-Doklady*, vol. 10, no. 8, pp. 707–710, 1966. [Online]. Available: http://www.bibsonomy.org/bibtex/2dedd0a2babcdb70996bb49e19d7de6fa/wvdaalst