Classificação de fraudes em licitações públicas através do agrupamento de empresas em conluios

David P. Galvão Júnior, Gilberto F. de Sousa Filho, Lucídio dos Anjos F. Cabral

[1]Centro de Informática - Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

{gilberto,lucidio}@ci.ufpb.br, galvao2010@icloud.com

Abstract. *Bid rigging causes significant damage to society and is the subject of intense investigation by the authorities. Many works try to analyze the financial values of the proposals during the competition of a bidding, trying to classify them a fraud. In this work, we propose to group companies that participate in the same fraudulent bids, quantifying the probability of each group being collusive, and apply this metric in machine learning algorithms to classify new bids. Results demonstrate an improvement in the cross-validation correlation of up to* 9% *compared to the classification obtained by the literature metrics.*

Resumo. *Fraudes em licitações causam enormes prejuízos à sociedade, e são alvo de intensa investigação das autoridades. Muitos trabalhos procuram analisar os valores financeiros das propostas durante a concorrência de uma licitação, procurando classificá-la como fraude. Neste trabalho, propomos agrupar as empresas que participam das mesmas licitações fraudulentas, quantificando a probabilidade de cada grupo ser conluio, e aplicar esta métrica em algoritmos de aprendizado de máquina para classificar novas licitações. Resultados demonstram uma melhora na correlação de validação cruzada de até* 9% *comparada à classificação obtida pelas métricas da literatura.*

1. Introdução

Na administração pública, a licitação é um processo adotado para a contratação de bens e serviços. Em um processo licitatório, entidades como empresas e pessoas competem para serem escolhidas como a vencedora do processo, essa competição visa possibilitar ao poder público acesso a melhores preços nos seus insumos. O Banco Mundial estima que as licitações em 2018 totalizaram 11 trilhões de dólares, ou cerca de 12% do produto interno bruto (PIB) global daquele ano [Bosio and Djankov 2020]. No contexto do Brasil, foi estimado que as licitações foram responsáveis por cerca de 20% do PIB de 2018.

Por sua vez, fraudes em processos licitatórios consistem em adulterar o caráter competitivo do processo, com fins de obter vantagens ilícitas com o resultado deste. A descoberta dessas fraudes é um processo não trivial, que requer por parte das autoridades competentes intensa investigação. Trabalhos passados relacionados buscaram criar ferramentas para auxiliar na detecção de fraudes por meio de uma análise estrutural [Imhof et al. 2017, Imhof et al. 2018] e/ou comportamental [Signor et al. 2020] das licitações e dos licitantes, essa análise contempla a investigação de fatores, como por exemplo, condições de mercado, preços das propostas e a distribuição dessas. Estes fatores foram utilizados para a classificação de licitações através do uso de aprendizagem de máquina [Huber and Imhof 2019, Rodríguez et al. 2022].

O conluio em uma licitação é caracterizado pela atuação coordenada entre várias empresas participantes, também chamado de cartel, com fins de aumentar seus lucros, sendo considerado uma prática ilegal. É a formação destes conluios que a análise da literatura procura encontrar de forma indireta através da análise econômica das licitações.

Neste trabalho, propomos agrupar as empresas guiado por sua participação concomitante em licitações fraudulentas, para isto, propomos um modelo de agrupamento cuja função objetivo é juntar empresas que concorreram ao maior número de licitações em comum e em seguida quantificar a probabilidade desta participação conjunta está superior ao acaso indicando uma possível formação de conluio. Esta probabilidade foi introduzida como uma nova característica de entrada aos algoritmos de aprendizado de máquina.

O trabalho está organizado da seguinte forma: Seção 2 apresenta trabalhos relacionados à classificação de licitações; Seção 3 descreve o modelo de agrupamento de empresas por interseção de licitações; a Seção 4 descreve como o agrupamento de empresas ajuda na classificação de novas licitações; os experimentos computacionais são discutidos na Seção 5; enquanto que as considerações finais são trazidas na Seção 6.

2. Detecção de fraude em licitações

Nesta seção são apresentados alguns métodos existentes na literatura, com fins de realizar a detecção automática de fraude em licitações. Em suma, trabalhos anteriores empregaram técnicas envolvendo a análise estrutural e/ou análise comportamental das licitações.

2.1. Análise do comportamento de participantes de uma licitação

Em [Signor et al. 2020], é proposto um método probabilístico, baseado na análise do comportamento conjunto de participantes que agem juntos com fins de fraudar uma licitação. Para cada licitação, é elaborada por parte do organizador uma estimativa de valor pré-licitação (PTE), tal métrica estabelece um norte para o preço das propostas submetidas pelos participantes de uma licitação. Em uma proposta honesta, se assume que o licitante faz sua própria estimativa de preço e aplica uma margem que reflete fatores, como por exemplo, estratégia da empresa e condições de mercado.

Em suma, os autores consideram que dada a natureza aleatória de cada proposta, as diferenças entre as propostas e a estimativa pré-licitação seguem uma distribuição normal com média $\mu = 0$ e um desvio padrão, que teoricamente é difícil de mensurar, no entanto, é assumido arbitrariamente que 90% das propostas honestas ficam a uma distância de até 20% da estimativa pré-licitação, o que leva a um desvio padrão $\sigma = 0.12$, sendo assim, é possível identificar casos suspeitos, pois estes tendem a fugir significativamente da média. O modelo então consiste nos seguintes passos:

1. Computar as probabilidades individuais, denominadas de $P(i)$, por meio da função de distribuição acumulada, de uma proposta aleatória possuir valores inferiores à cada proposta observada i;
2. Calcular a probabilidade conjunta, denominada de $P(x)$, de um conjunto aleatório $B_1, B_2, ..., B_n$ de n propostas serem observados ao acaso, multiplicando suas probabilidades segundo a equação $P(x) = \prod_{i=1}^{n} P(i)$;
3. Comparar se $P(x)$ é superior aos valores limites, para um conjunto de n propostas e um dado intervalo de confiança, os valores limites são computados utilizando a função gama incompleta $\pi_n(x) = \frac{\Gamma(n,-ln(x))}{(n-1)!}, 0 \leq x \leq 1$;
4. Se $P(x) > \pi_n(x)$, houve fraude na licitação.

2.2. Estatística descritiva para detecção de fraude

Por meio de estatística descritiva, é possível fazer a análise de variáveis estratégicas, como por exemplo, preços e participação de mercado para definir se uma ou mais empresas fogem de um comportamento competitivo. Na literatura, diversos trabalhos se dedicam a propor fórmulas para definir características adicionais a uma licitação, e assim fornecer padrões que auxiliem o modelo de detecção de conluio. Seja t uma licitação, [Imhof et al. 2017] definem as seguintes características :

- *Coeficiente de Variação (CV)*: $CV(t) = \frac{\sigma_t}{\mu_t}$, onde σ_t e μ_t são respectivamente, o desvio padrão e a média dos valores das ofertas em uma licitação;
- *Distância Relativa*: $RD(t) = \frac{b_{2,t} - b_{min,t}}{\sigma_{l,t}}$, onde $b_{min,t}$ e $b_{2,t}$ são respectivamente as propostas de menor e segundo menor valor, e $\sigma_{l,t}$ é o desvio padrão das propostas perdedoras de uma licitação;
- *Amplitude*: $SPD(t) = \frac{b_{max,t} - b_{min,t}}{b_{min,t}}$, é a razão entre a diferença da oferta de maior e menor valor pela oferta de menor valor;
- *Diferença entre os mínimos*: $DIFFP(t) = \frac{b_{2,t} - b_{min,t}}{b_{min,t}}$, fornece a razão entre a diferença das duas ofertas de menor valor pela oferta de menor valor;
- *Excesso de Curtose:* sejam n_t o número de propostas submetidas à licitação t, e b_{it} o valor da i-ésima proposta submetida, seu valor é dado pela equação $KURT(t) = \frac{n_t^2 + n_t}{(n_t-1)(n_t-2)(n_t-3)} \sum_{i=1}^{n_t} (\frac{b_{it} - \mu_t}{\sigma_t})^4 - \frac{3(n_t-1)^3}{(n_t-2)(n_t-3)}$;
- *Assimetria:* busca identificar possíveis assimetrias na distribuição dos valores das ofertas através da equação $SKEW(t) = \frac{n_t}{(n_t-1)(n_t-2)} \sum_{i=1}^{n_t} (\frac{b_{it} - \mu_t}{\sigma_t})^3$;
- *Teste de Kolmogorov-Smirnov (KSTEST):* mede a similaridade da distribuição dos valores das ofertas em comparação com outra distribuição, no caso, é feita a comparação com uma distribuição uniforme.

Utilizando dados referentes à licitações na Suíça, em [Imhof et al. 2018] são empregadas as características CV e RD para separar, dos dados, licitações e empresas que apresentam um padrão suspeito, para tanto, são selecionadas licitações que apresentam valores baixos para o CV e altos para o RD.

Adicionalmente, em [Huber and Imhof 2019] e [Rodríguez et al. 2022], os autores buscam adicionar as características apresentadas acima aos dados originais como entrada para algoritmos de aprendizagem de máquina, é constatado que esta adição é capaz de prover ganhos na precisão do modelos para detecção de fraudes.

Diferente das estratégias da literatura que trabalham com a identificação de comportamentos escusos das empresas pela análise dos valores monetários das propostas às licitações, propomos a identificação de conluios de empresas guiada pelas participações concomitante em licitações fraudulentas contidas na base de dados. Esta técnica de agrupamento e seu uso na classificação de licitações estão descritas nas seções seguintes.

3. Agrupamento por Interseção de Conjuntos

Nesta seção definimos formalmente o Problema de Agrupamento por Interseção de Conjuntos (PAIC), além de um algoritmo para sua resolução.

3.1. Problema de Agrupamento por Interseção de Conjuntos

Sejam $X = \{1, 2, .., n\}$ um conjunto de n objetos, L o conjunto de recursos compartilhados entre os objetos, $L_i \subseteq L$ o conjunto de recursos utilizados pelos objetos $i \in X$, $P = \{C_1, C_2, .., C_m\}$ um particionamento de X,

$$L_{C_k} = \begin{cases} \bigcap_{i \in C_k} L_i & \text{, se } |C_k| > 1 \\ \emptyset & \text{, caso contrário} \end{cases}$$

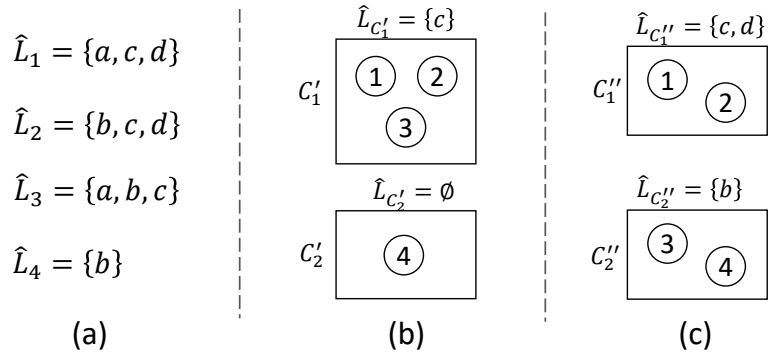
o conjunto interseção dos recursos utilizados pelos objetos contidos em uma parte C_k. No PAIC deseja-se encontrar um particionamento P que

$$\begin{aligned} \text{Maximize} \quad & \sum_{k=1}^{m} |L_{C_k}| \\ \text{s.a.} \quad & \bigcup_{k=1}^{m} C_k = X, && (1) \\ & C_k \neq \emptyset, \qquad \forall k \in \{1, 2, ..., m\}, && (2) \\ & C_k \cap C_l = \emptyset, \qquad \forall k, l \in \{1, 2, ..., m\}, k \neq l. && (3) \end{aligned}$$

Neste modelo o agrupamento é um particionamento de X obtida pelas restrições (1), (2) e (3), ou seja, tem-se como saída do problema um agrupamento P onde todo objeto $i \in X$ está em uma e somente uma parte de P, e a soma do tamanho das interseções de cada parte é máxima.

A Figura 1(a) ilustra uma instância exemplo $\hat{X} = \{1, 2, 3, 4\}$ para o PAIC através dos conjuntos de recursos $\hat{L}_i$. A Figura 1(b) ilustra um particionamento $P' = \{C'_1, C'_2\}$ de $\hat{X}$, onde $C'_1 = \{1, 2, 3\}$ possui apenas o recurso c em sua interseção, e a parte $C'_2 = \{4\}$, por ser unitária, não possui recursos em sua interseção, logo, P' contém apenas um recurso. Já a Figura 1(c) apresenta o particionamento ótimo de $\hat{X}$, pois com a ida do objeto 3 de C'_1 para C''_2 obtemos $\hat{L}_{C''_1} = \{c, d\}$ e $\hat{L}_{C''_2} = \{b\}$, aumentando para três os recursos nas interseções de P''.

Figura 1. Soluções para o PAIC. (a) Recursos dos objetos da instância $\hat{X} = \{1, 2, 3, 4\}$, (b) solução P' de $\hat{X}$ e (c) solução ótima P'' de $\hat{X}$.



Proposição 1. *Seja X uma instância para o PAIC, existe pelo menos um particionamento ótimo P^* de X com no máximo dois objetos em cada parte.*

Demonstração. É trivial que para toda parte C_k, com $|C_k| \geq 2$, a adição de um novo objeto não aumenta sua interseção, podendo permanecer a mesma ou perder recursos. E para toda parte C_k, com $|C_k| \geq 3$, a remoção de um objeto não diminui sua interseção, podendo permanecer a mesma ou ganhar recursos.

Sem perda de generalidade, considere $P^* = \{C_1^*, C_2^*, .., C_k^*\}$ uma solução ótima qualquer de X cujas partes estão ordenadas de forma não crescente pelos seus tamanhos, ou seja, $|C_1^*| \geq |C_2^*| \geq .. \geq |C_k^*|$. Seja u o menor índice onde $|C_u^*| = 1$, então para todo $i \in \{u, .., k\}$ remova um objeto x de qualquer parte $C_j^* \in \{C_1^*, .., C_{u-1}^*\}$, caso exista $|C_j^*| \geq 3$, e adicione em C_i^*. A saída de x de C_j^* não aumenta sua interseção, pois bastaria criar uma parte só com x para termos uma solução melhor que P^*, o que é impossível pois P^* é ótimo, do mesmo modo, o objeto x não possui interseção com o único objeto de C_i^*, caso contrário, a parte $C_i^* \cup x$ já estaria em P^*, logo, a adição de x nas partes unitárias C_i^* não altera suas interseções.

Seja $z = k+1$, para todo $i \in \{z, .., \frac{n}{2}\}$ remova um objeto x de qualquer parte $C_j^* \in \{C_1^*, .., C_{u-1}^*\}$, caso exista $|C_j^*| \geq 3$, e crie uma parte unitária C_i^* com $L_{C_i^*} = \emptyset$ e adicione em P^*. Agora repita o procedimento anterior com as partes unitárias $\{C_z^*, .., C_{\frac{n}{2}}^*\}$ e obtenha um novo particionamento P^*, cujas partes possuem no máximo dois objetos cada, e com a mesma interseção do particionamento ótimo original. □

3.2. *Branch and Bound* para o PAIC

Da proposição 1 definimos um algoritmo *branch and bound* para resolução do PAIC chamado $PartitionTree$. Sejam $I(P) = \sum_{c \in P} L_c$ o valor solução do particionamento P, $|P|$ a quantidade de partes contidas em P e $C_{cand} = \{\{i, j\} \mid \forall i, j \in X$ tal que $L_i \cap L_j \neq \emptyset\}$ a lista de partes candidatas que contém dois objetos com interseção não vazia. C_{cand} é ordenado de forma não crescente pelo tamanho de suas interseções e fornecida para a execução $P^* = PartitionTree(C_{cand}, 0, \{\})$, caso exista objetos em X que não estão em P^*, então $P^* = P^* \cup \{x \mid x \in X$ e $x \notin P^*\}$, tornado assim P^* uma partição ótima de X.

Algoritmo 1: PartitionTree(C_{cand}, i, $\hat{P}$)

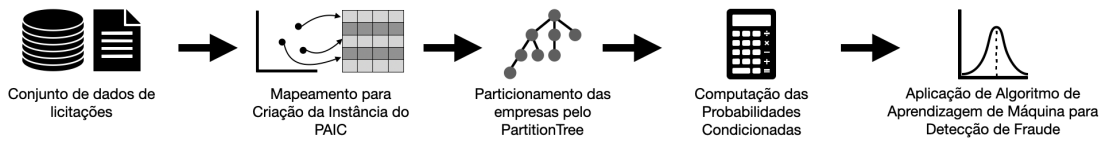
```
1  se I(P̂) > I(P*) então
2  |  P* ← P̂
3  fim
4  enquanto i ≤ |C_cand| faça
5  |  c ← C_cand[i]
6  |  se I(P̂) + |L_c| · (n/2 − |P̂|) ≤ I(P*) então
7  |  |  retorna
8  |  fim
9  |  se os objetos de c ∉ P̂ então
10 |  |  PartitionTree(c, i + 1, P̂ ∪ c)
11 |  fim
12 |  i ← i + 1
13 fim
14 retorna P*
```

O algoritmo $PartitionTree$ é recursivo e, nas iterações das linhas 4-13, visita todos os possíveis agrupamentos formados pelas partes de C_{cand} numa abordagem de busca em profundidade. O teste da linha 1 garante que P^* retorne o melhor agrupamento para o valor de $I(.)$. O teste da linha 6 estima um valor máximo a ser alcançado para o $\hat{P}$ corrente, se a estimativa for pior do que o valor atual de P^*, então esta solução $\hat{P}$ é descartada. Por fim, o teste da linha 9 garante a propriedade de particionamento na construção de $\hat{P}$ corrente.

4. Classificação de licitações

Nesta seção apresentamos a metodologia adotada no trabalho para a classificação de licitações fraudulentas. A Figura 2 ilustra as etapas envolvidas no modelo de detecção de conluios proposto, sendo melhor detalhadas nas subseções seguintes.

Figura 2. Metodologia do modelo de detecção de fraudes



A primeira etapa (Seção 4.1) consiste no mapeamento das licitações em estruturas específicas do PAIC, construindo, através do algoritmo $PartitionTree$, o particionamento da lista de empresas E, guiado pelo histórico de licitações contida no *dataset* fornecido. A segunda etapa (Seção 4.2) consiste na computação da probabilidade condicionada de uma licitação l ser fraudulenta caso a lista de empresas E participe de l. E na terceira etapa (Seção 4.3) utilizamos as métricas associadas às licitações juntamente com a probabilidade condicionada proposta para construir um classificador de licitações com o auxílio de algoritmos de aprendizado de máquina.

4.1. Mapeamento

As bases de dados das licitações públicas utilizadas neste trabalho possuem formato tabular, cada linha corresponde a uma proposta de uma empresa e em uma licitação l, contendo entre diversas informações, os identificadores da empresa e da licitação associada. Cada licitação possui uma categoria binária $c \in \{fraude, legit\}$ indicando a presença ou não de fraude nesta. Extraímos a lista de empresas que deram lances e mapeamos no conjunto de objetos X, já a lista de licitações é mapeada no conjunto de recursos L, entretanto, como as licitações possuem duas categorias distintas, definimos L^c como a parte do conjunto de licitações rotulados na categoria c e L^c_e a parte do conjunto de licitações rotuladas em c que a empresa e participou. Ou seja, L^{fraude} é o conjunto de licitações rotuladas como fraudada, já L^{legit} o conjunto de licitações legítimas.

Em cada linha encontramos uma relação entre uma empresa e e uma licitação l de rótulo c, que é mapeada na lista L^c_e, fazendo $L^c_e = L^c_e \cup l$ para cada relação. Com a entrada (X, L^c, L^c_e) formada, utilizamos o algoritmo $PartitionTree$ para nos fornecer o particionamento P^c das empresas de X guiado pela participação concomitante nas licitações da categoria c. Na próxima seção apresentaremos o uso de P^c para a análise da formação destes agrupamentos de empresas na classificação de licitações ainda não classificadas.

4.2. Teorema de Bayes na classificação de licitações

Diversos estudos demonstraram que o uso Teorema de Bayes é eficaz para problemas de classificação [Sakkis et al. 2003, Langley et al. 1992, Domingos and Pazzani 1996]. Neste trabalho, propomos seu uso para computar a probabilidade condicionada de uma licitação l ser fraudulenta caso a lista de empresas E participe desta licitação.

Do Teorema de Bayes e da lei da probabilidade total, a probabilidade de uma licitação l, cujas empresas $E = \{e_1, ..., e_m\}$ lançaram proposta, pertencer a categoria c é:

$$Pr(l = c \mid E) = \frac{Pr(E \mid l = c) \cdot Pr(l = c)}{\sum_{k \in \{fraude, legit\}} Pr(E \mid l = k) \cdot Pr(l = k)}. \quad (4)$$

Por causa da dificuldade de computar $Pr(E \mid l = c)$ o classificador *Naive Bayesian* assume de forma simplista que o evento de participação individual de cada empresa $E = e_i$ é independente das outras, logo, temos que

$$Pr(E \mid l = c) = \prod_{i=1}^{m} Pr(E = e_i \mid l = c).$$

Já para computar as probabilidades a priori básicas contidas nos dados, temos que

$$Pr(l = c) = \frac{|L^c|}{|L|} \text{ e}$$
$$Pr(E = e_i \mid l = c) = \frac{k + |L^c_{e_i}|}{2k + |L^c|},$$

sendo $k = 0,5$ uma constante necessária para evitar erros de classificação quando a ocorrência do evento $E = e_i$ for rara na base de dados.

Como citamos acima, a suposição de independência dos eventos $E = e_i$ é simplista, e supomos que exista a formação de conluios entre empresas para fraudar licitações. Para quantificar estas relações contidas nos dados, propomos a construção do particionamento P^c das empresas $E = \{e_1, ..., e_m\}$ através do algoritmo $PartitionTree$, sendo cada grupo de empresas $S \in P^c$ uma relação de dependência medida pela suas ocorrências concomitantes em outras licitações. Logo, redefinimos a probabilidade condicionada do evento de participação das empresas E em uma licitação de categoria c como

$$Pr(E \mid l = c) = \prod_{S \in P^c} Pr(E = S \mid l = c).$$

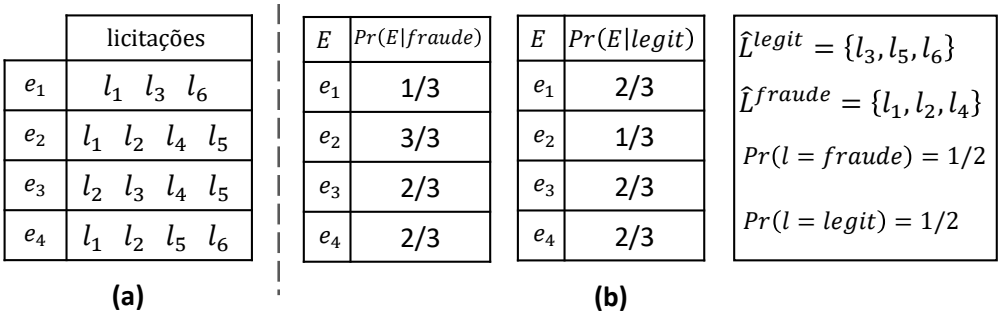
E por fim, a probabilidade a priori do evento $E = S$ pode ser computada por

$$Pr(E = S \mid l = c) = \frac{k + |L_S|}{2k + |L^c|},$$

aplicando a mesma constante k com o objetivo de tratamento de eventos raros.

Para o entendimento do efeito da identificação de dependências entre eventos pelo PAIC para a classificação de licitações, propomos computar a probabilidade condicionada

Figura 3. Licitação exemplo $\hat{l}$. (a) Histórico das licitações que as empresas de $\hat{E}$ concorreram e (b) probabilidades a priori contidas no histórico de $\hat{E}$.



$Pr(\hat{l} = fraude \mid \hat{E})$ da licitação exemplo $\hat{l}$ cuja lista de empresas concorrentes $\hat{E} = \{e_1, e_2, e_3, e_4\}$ tem seu histórico de participação em licitações ilustradas na Figura 3(a).

A Figura 3(b) apresenta as probabilidades a priori extraídas do histórico de licitações das empresas de $\hat{E}$. Por exemplo, a probabilidade da empresa e_3 participar de uma licitação caso ela seja fraudulenta é $Pr(E = e_2 \mid l = fraude) = 2/3$, pois das 3 empresas rotuladas como fraudulentas, e_3 participou de 2 (l_2 e l_4). A chance de uma licitação l qualquer ser fraudulenta é $Pr(l = fraude) = 1/2$. A categorização destas licitações do histórico está ilustrada nas listas $\hat{L}^{fraude}$ e $\hat{L}^{legit}$. Neste exemplo ilustrativo consideramos a constante de eventos raros $k = 0$ para facilitar o seu entendimento.

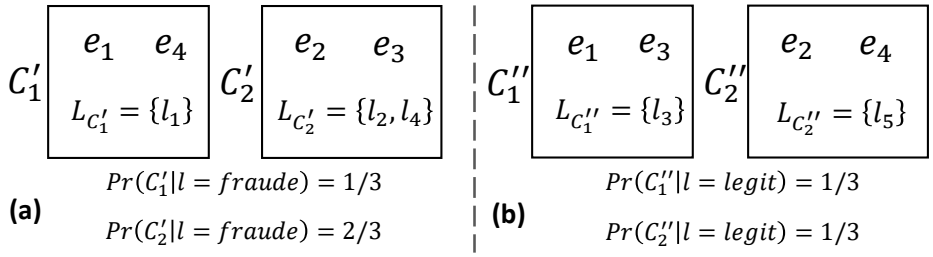
Em seguida computamos $Pr(\hat{E} \mid l = fraude)$ e $Pr(\hat{E} \mid l = legit)$. Inicialmente, para efeito de comparação, assumiremos que os eventos são independentes, logo,

$$Pr(\hat{E} \mid l = fraude) = \frac{1}{3} \cdot 1 \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{27},$$

de forma análoga computamos $Pr(\hat{E} \mid l = legit) = 8/81$, e aplicando a equação (4) temos que $Pr(\hat{l} = fraude \mid \hat{E}) = 3/5$.

Agora iremos considerar a dependência dos eventos entre as empresas de $\hat{E}$ e utilizando o PAIC construímos os particionamentos P^{fraude} e P^{legit} que estão ilustrados na Figura 4. Temos também as probabilidades a priori dos grupos de empresas das partições, como por exemplo, da partição $P^{fraude} = \{C'_1, C'_2\}$, ilustrada na Figura 4(a), podemos observar que a probabilidade $Pr(E = C'_2 \mid l = fraude) = 2/3$ pois sua interseção $L_{C'_2}$ possui duas (l_2 e l_4) das três licitações rotuladas como fraude.

Figura 4. Particionamentos de $\hat{E}$ para as listas de empresas $\hat{L}^{fraude}$ e $\hat{L}^{legit}$. (a) Particionamento P^{fraude} e (b) particionamento P^{legit}.



De posse das probabilidades a priori das partições, podemos recalcular

$$Pr(\hat{E} \mid l = fraude) = \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9},$$

$Pr(\hat{E} \mid l = legit) = 1/9$, e aplicando a equação (4) temos que $Pr(\hat{l} = fraude \mid \hat{E}) = 2/3$ que é superior à probabilidade de $3/5$ computada anteriormente assumindo independência dos eventos.

4.3. Algoritmo de Aprendizagem de Máquina

Identificar presença ou não de fraude em uma licitação é um problema de classificação, sendo assim, propomos o uso da técnica de aprendizado supervisionado, adicionando a probabilidade condicionada $Pr(l = fraude \mid E)$ como novo parâmetro das amostras rotuladas contidas na base. Para a classificação das licitações, são propostos os algoritmos: Máquina de Vetor de Suporte (SVM); K-Vizinhos mais Próximos (K-NN); Rede Neural (NN); Random Forest (RF) e Árvore de Decisão (DT).

Em alguns algoritmos, se pressupõe uma distribuição normal dos dados ou que as características estejam na mesma escala, caso contrário, algumas das características apresentadas ao algoritmo podem fazer com que este seja incapaz de aprender corretamente com os dados apresentados. Sendo assim, com fins de mitigação desse problema, cada variável x dos dados passa por uma padronização seguindo a equação $z = \frac{x-\mu}{\sigma}$, sendo μ a média aritmética e σ o desvio padrão.

Para a avaliação dos modelos de aprendizagem de máquina, a base de dados orginal é dividida em base de treino e base de teste, sendo feita uma divisão dos objetos do conjunto de dados entre estas bases. Considerando que uma licitação pode ter mais de um participante, a divisão entre bases é realizada por licitação.

Para evitar dar informações sobre o teste para a função hipótese treinada (bisbilhotagem de dados) propomos as seguinte ações: (a) a padronização dos dados é feita sobre a base de treinamento, e utilizando os valores μ e σ computados no treino efetuamos a padronização da base teste; (b) na fase de treinamento não são consideradas as licitações da base teste para a construção do particionamento P^c utilizado na computação das probabilidades condicionadas, enquanto que na computação das probabilidade condicionadas de licitações do teste é considerado todo o conjunto de licitações original.

5. Experimentos computacionais

O algoritmo proposto na Seção 3.2 e os algoritmos de aprendizagem de máquina listados na Seção 4.3 foram implementados na linguagem de programação Python versão 3.9.6 com auxílio da biblioteca *scikit-learn* [du Boisberranger et al. 2022]. Todos os experimentos computacionais deste trabalho foram executados em um computador com processador Intel Core i5, de 4 núcleos de 2,3 GHz, a máquina dispõe de 8 GB de memória RAM, e possui como sistema operacional o MacOS Ventura.

Foram utilizados quatro conjuntos de dados envolvendo licitações no Brasil, Itália, Japão e Estados Unidos, disponibilizados em [Rodríguez et al. 2022]. A Tabela 1 descreve as principais características dos conjuntos de dados, a densidade da instância do PAIC é dada por Densisdade $= \frac{\text{\# propostas}}{\text{\# licitações} \times \text{\# empresas}}$. Todas as bases contém as características: quantidade de participantes e valor da proposta vencedora. As licitações do

Brasil, Japão e Itália incluem a estimativa de valor pré-licitação (PTE), e a razão entre o valor da proposta vencedora e o PTE.

Tabela 1. Descrição dos conjuntos de dados utilizados

País de origem dos dados	Brasil	Japão	Itália	Estados Unidos
Escopo	Licitações de infraestrutura da Petrobras	Construção de edificações e engenharia civil	Construção de estradas	Fornecimento de leite para escolas
# licitações	101	1080	278	3754
# empresas	272	1665	821	120
# propostas	683	13515	20286	7004
# licitações fraudulentas	33	123	143	487
Densidade (PAIC)	2,48%	0,75%	8,88%	1,55%

Resultados dos algoritmos de Aprendizagem de Máquina

Para a comparação das características propostas neste trabalho e as características definidas na literatura, criamos 4 configurações distintas a serem fornecidas aos algoritmos de aprendizagem. Todas as configurações incluem, quando disponível, a estimativa de valor pré-licitação (PTE), valor médio das propostas, razão entre o valor da proposta vencedora e o PTE, e quantidade de propostas. São definidas as seguintes configurações:

- *Configuração 1*: acrescenta ao conjunto de características originais as 7 características criadas por meio de estatística descritiva (CV, SPD, DIFFP, RD, KURT, SKEW, KTEST), utilizadas em [Rodríguez et al. 2022] e abordadas na seção 2.2;
- *Configuração 2*: adiciona a característica da probabilidade $P(x)$ de uma licitação, proposta em [Signor et al. 2020], e descrita na seção 2.1;
- *Configuração 3*: adiciona a probabilidade $Pr(l = fraude \mid E)$ definida na Seção 4.2;
- *Configuração 4*: contempla apenas as características originais;

Para comparar a eficácia dos modelos de classificação dividimos toda as bases em: base treino (80% das licitações) e base teste (20% restantes). Para toda combinação da base de dados fornecida, algoritmo de aprendizado e configuração das características, é realizada uma calibração dos hiperparâmetros dos algoritmos utilizados por meio da técnica de *k-fold cross validation* ($k = 5$), sendo considerados os seguintes valores:

- SVM: flexibilidade da função fronteira $C \in \{0.001, 0.1, 1, 2\}$ e coeficiente do kernel $\gamma \in \{\frac{1}{|F|}, \frac{1}{|F| \cdot var(X)}\}$, onde $|F|$ é o número de características de entrada, X é a entrada do algoritmo, e $var(X)$ é uma função que retorna a variância de X;
- K-NN: número de vizinhos $k \in \{1, 3, 5, 7\}$, ordem da função de distância de Minkowski $p \in \{1, 2, 3, 4\}$, e peso dado aos objetos $w \in \{\text{uniforme}, \text{proporcional}\}$;
- NN: taxa de aprendizagem inicial $\theta \in \{0.001, 0.01, 0.1\}$, taxa de aprendizagem $\in \{\text{constante}, \text{adaptativa}\}$, número de neurônios nas camadas ocultas $\in \{(100, 1), (24, 1), (8, 4, 2, 1)\}$, parâmetro de regularização $\alpha \in \{0.0001, 0.001, 0.1\}$;
- RF: número de classificadores $\in \{29, 99, 199\}$, profundidade máxima da árvore $\in \{\text{ilimitado}, 3, 4, 5, 7\}$, número mínimo de objetos em um nó folha $\in \{1, 2, 3\}$;
- DT: profundidade máxima da árvore, realizada com auxílio da heurística de poda por complexidade de custos [Breiman et al. 1984];

Todos os hiperparâmetros não envolvidos no processo de calibração foram definidos conforme o padrão da biblioteca *scikit-learn*.

A Tabela 2 apresenta os resultados dos modelos de detecção de fraudes para as quatro configurações. Valores em negrito indicam a melhor área abaixo da curva ROC (AUC) entre todas as configurações para um conjunto de dados, que mede a probabilidade de um exemplo positivo aleatório receber uma pontuação maior que um exemplo negativo aleatório [Fawcett 2006]. Por conta da ausência do PTE para as licitações dos EUA, a Configuração 2 não é aplicada nestes experimentos e está marcada com "-".

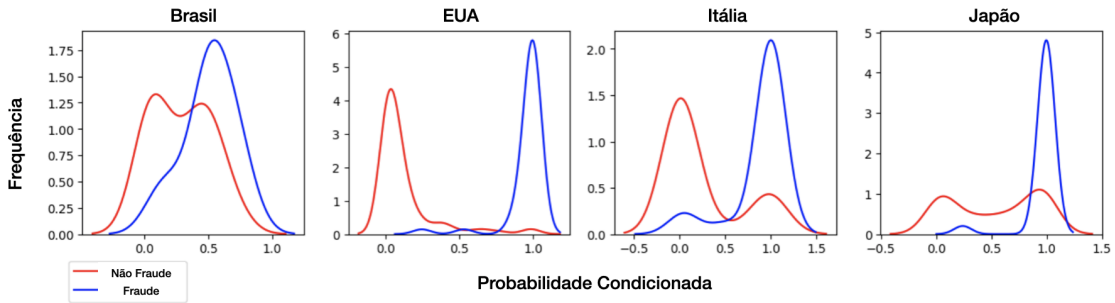
Tabela 2. Resultados dos modelos de classificação de fraude em licitações

			Conjunto de Dados			
			Brasil	EUA	Itália	Japão
Configuração	1	Melhor Algoritmo	SVM, NN, RF	RF	RF	NN
		Melhor AUC	**97,62%**	66,11%	94,01%	96,38%
	2	Melhor Algoritmo	SVM, NN	-	RF	RF
		Melhor AUC	95,24%	-	96,05%	95,21%
	3	Melhor Algoritmo	DT	NN	RF	NN
		Melhor AUC	97,02%	**99,53%**	**96,73%**	**96,57%**
	4	Melhor Algoritmo	RF	NN, RF	RF	NN
		Melhor AUC	94,05%	88,79%	96,19%	95,92%

Na Tabela 2 é possível observar que, os melhores modelos da Configuração 3 são capazes de obter uma melhor acurácia em relação aos melhores modelos das outras configurações em 3 dos 4 conjuntos de dados testados, o que demonstra uma robustez da característica proposta em possibilitar uma melhor detecção de fraudes.

Já a Figura 5 ilustra a distribuição de valores das probabilidades condicionadas de uma licitação em relação a presença ou ausência de fraude na partição teste dos quatro conjuntos de dados. É possível observar que para todos os conjuntos há uma nítida diferença entre os valores das probabilidades em licitações fraudulentas e não fraudulentas, com licitações fraudulentas assumindo valores superiores. No entanto, para as licitações do Brasil, há muita sobreposição dos valores, sendo um gerador de ruídos para o classificador. Vale salientar que esta base possui no total apenas 101 licitações, fato que pode inviabilizar o modelo de particionamento por falta de relações entre as empresas.

Figura 5. Distribuição do valor das probabilidades condicionadas nos *datasets*.



6. Considerações finais e trabalhos futuros

Ao representar uma parcela considerável do PIB, licitações são alvo de malfeitores que buscam fraudá-las para ganhos ilícitos, causando assim prejuízo aos cofres públicos.

Neste trabalho foi proposto um método para classificação de licitações fraudulentas, que faz uso de conceitos de agrupamentos guiados pela interseção de licitações em comum e a mensuração da probabilidade destes grupos não se formarem ao acaso. Para utilizar esta abordagem proposta em outras bases de licitações se faz necessária a identificação única das empresas em cada licitação, permitindo assim seu agrupamento.

A distribuição dos valores das probabilidades mostra uma clara diferença para licitações com e sem fraude. O modelo proposto, comparado aos modelos que utilizam métricas da literatura, apresentou um melhor AUC em 3 dos 4 conjuntos de dados. Trabalhos futuros incluem a criação de algoritmos mais eficientes de agrupamento e métricas com base em cobertura, que possibilita listar todos os grupos em que as empresas podem estar, abrindo um maior leque de possibilidades na identificação de conluios.

Referências

Bosio, E. and Djankov, S. (2020). How large is public procurement?

Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1984). *Classification and Regression Trees*. Taylor & Francis.

Domingos, P. M. and Pazzani, M. J. (1996). Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *International Conference on Machine Learning*.

du Boisberranger, J., Van den Bossche, J., Estève, L., and J. Fan, T. (2022). scikit-learn: machine learning in python.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874. ROC Analysis in Pattern Recognition.

Huber, M. and Imhof, D. (2019). Machine learning with screens for detecting bid-rigging cartels. *International Journal of Industrial Organization*, 65:277–301.

Imhof, D., Blatter, M., Brisset, K., BUhler, S., Egli, A., Karagök, Y., Madì, T., Schmutzler, A., and Wyssling, M. (2017). Simple statistical screens to detect bid rigging acknowledgement.

Imhof, D., Karagök, Y., and Rutz, S. (2018). SCREENING FOR BID RIGGING—DOES IT WORK? *Journal of Competition Law & Economics*, 14(2):235–261.

Langley, P., Iba, and, W., and Thompson, K. (1992). An analysis of bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, page 223–228. AAAI Press.

Rodríguez, M. J. G., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E., and Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, 133.

Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C. D., and Stamatopoulos, P. (2003). A memory-based approach to anti-spam filtering for mailing lists. *Inf. Retr.*, 6(1):49–73.

Signor, R., Love, P. E. D., Belarmino, A. T. N., and Olatunji, O. A. (2020). Detection of collusive tenders in infrastructure projects: Learning from operation car wash. *Journal of Construction Engineering and Management*, 146.