SMOTE Generalizado: Uma técnica de sobreamostragem de geração universal para todos os tipos de dados em aprendizado desequilibrado

Documento da Conferência· Setembro de 2021

6 autores, Incluindo:

Carlos Cernuda
Mondragon Unibertsitatea
23PUBLICAÇÕES**304**CITAÇÕES
VER PERFIL

Daniel Reguera-Bakhache
Mondragon Unibertsitatea
12PUBLICAÇÕES**12**CITAÇÕES
VER PERFIL

Aitor Aguirre
Mondragon Unibertsitatea
8PUBLICAÇÕES**13**CITAÇÕES
VER PERFIL

Mikel Iturbe
Mondragon Unibertsitatea
12PUBLICAÇÕES**115**CITAÇÕES
VER PERFIL

SMOTE Generalizado: Uma técnica de sobreamostragem de geração universal para todos os tipos de dados em aprendizagem desequilibrada

Carlos Cernuda
Departamento de Eletrônica e Computação
Universidade de Mondragon
Arrasate-Mondragon, Espanha
ccernuda@mondragon.edu

Daniel Reguera-Bakhache
Departamento de Eletrônica e Computação
Universidade de Mondragon
Arrasate-Mondragon, Espanha
dreguera@mondragon.edu

Aitor Aguirre
Departamento de Eletrônica e Computação
Universidade de Mondragon
Arrasate-Mondragon, Espanha
aaguirre@mondragon.edu

Mikel Iturbe
Departamento de Eletrônica e Computação
Universidade de Mondragon
Arrasate-Mondragon, Espanha
miturbe@mondragon.edu

Iñaki Garitano
Departamento de Eletrônica e Computação
Universidade de Mondragon
Arrasate-Mondragon, Espanha
igaritano@mondragon.edu

Urko Zurutuza
Departamento de Eletrônica e Computação
Universidade de Mondragon
Arrasate-Mondragon, Espanha
uzurutuza@mondragon.edu

Abstrato— Um problema comum que surge ao enfrentar tarefas de classificação é o problema de desequilíbrio de classes, que acontece quando uma ou mais classes estão fortemente sub-representadas em relação às demais, sendo geralmente aquelas classes minoritárias as de interesse. Uma solução natural consiste em corrigir o desequilíbrio por métodos de amostragem, sendo a Técnica de Sobreamostragem de Minorias Sintéticas (SMOTE) o método mais utilizado. Da mesma forma que todas as outras técnicas de sobreamostragem, ela se baseia no uso de distâncias/semelhanças para focar nas vizinhanças de amostras minoritárias no procedimento de geração de amostras sintéticas, portanto, destina-se a dados numéricos puros. No entanto, é muito comum coletar dados categóricos ou discretizar atributos numéricos como uma etapa de pré-processamento, limitando-se a abordagens de amostragem aleatória para corrigir o desequilíbrio.

Propomos o GSMOTE, uma generalização do método SMOTE, adequado para qualquer tipo de dado. Para a determinação das vizinhanças, a distância entre as amostras é obtida por meio da transformação do Coeficiente Geral de Similaridade de Gower em um novo Coeficiente Geral de Distância, no qual a parte correspondente à forma de medir semelhanças entre categorias em variáveis categóricas foi substituída por um recentemente apresentou medida de similaridade denominada medida de Entropia Variável, inspirada na Entropia de Shannon.
O GSMOTE foi testado em seis conjuntos de dados públicos desbalanceados, com diferentes características e níveis de desequilíbrio.

Termos de índice—Aprendizagem Desequilibrada, Dados Categóricos, Dados de Tipo Misto, Técnicas de Sobreamostragem, SMOTE

eu. euNTRODUÇÃO

Ao enfrentar uma tarefa de classificação é bastante comum lidar com conjuntos de dados em que uma ou várias classes estão claramente sub-representadas quando comparadas às demais. Tal cenário é conhecido como*aprendizagem desequilibrada*. Por exemplo, em aplicações de aprendizado de máquina industriais ou médicas do mundo real, como detecção de falhas ou diagnóstico de doenças, as classes de interesse, respectivamente estados defeituosos e de doença, são bastante infrequentes. As razões subjacentes ao desequilíbrio são diversas. Os processos de fabricação industrial [1] são altamente otimizados porque nenhuma empresa poderia arcar com níveis defeituosos mesmo moderados, portanto, geralmente menos de 2% das observações correspondem a estados defeituosos. Em aplicações biomédicas, os tipos de doenças que podem se beneficiar de modelos de aprendizado de máquina, por exemplo, são graves e difíceis de detectar por métodos tradicionais com precisão aceitável, o que não acontece com frequência [2] [3] [4] [5]] [6]. Em outros casos, como na detecção de fraudes em transações monetárias [7], o desequilíbrio se deve à própria natureza do processo, pois a maioria das transações são legais.

Independentemente da causa raiz do desequilíbrio, as ações para evitar que um classificador preveja sempre a classe majoritária (classificador ingênuo) são bastante limitadas. Apesar da existência de certas taxonomias com mais de duas possibilidades, poderíamos incluir todas em duas:

(eu)*Aprendizado sensível ao custo*[8]. Uma matriz de custos é definida para os diferentes erros que um classificador pode cometer. Em seguida, o classificador é treinado de tal forma que seu objetivo seja minimizar o custo global de todos os seus erros. Dessa forma, atribuir um alto custo à falha na classificação da classe (ou classes) minoritárias levaria a modelos mais precisos em predizê-la (resp.-los). A contrapartida é que pequenas melhorias nos erros indesejados provocariam um aumento nos demais muito maior [9]. O trade-off entre os erros fornecido pelo nosso controle sobre a matriz de custo é difícil de ajustar, devido a problemas de overfitting devido ao fato de forçar os algoritmos a se concentrarem em amostras de classes minoritárias durante o treinamento [10].

(ii)*Estratégias de amostragem*. As diferenças na quantidade de amostras das classes majoritária e minoritária são compensadas

aumentando a quantidade de amostras de classe minoritária (oversampling) e/ou diminuindo a quantidade de amostras de classe majoritária (undersampling). Como ambos têm prós e contras [11] [12] [13], também é comum optar por uma abordagem híbrida (*amostragem mista*) [14] [15] [16] combinando-os.

No caso de *sobreamostragem*, além de repetir aleatoriamente algumas amostras de classes minoritárias, a maioria das abordagens *gerar* novas amostras sintéticas de classe minoritária nas vizinhanças das originais existentes. O algoritmo mais utilizado é *técnica de sobreamostragem de minoria sintética (SMOTE)* [17]. No Smote, cada amostra sintética, x_{novo}, é selecionado aleatoriamente no segmento que conecta uma das amostras originais da classe minoritária, x_{eu}, com outra amostra de classe minoritária de seu bairro, x_j. A menos que optemos por abordagens aleatórias, que geralmente apresentam desempenho ruim [15], focaremos em vizinhanças, portanto, abordagens baseadas em distância, o que significa que estamos limitados a recursos numéricos. Além disso, não poderíamos discretizar características numéricas e depois corrigir o desequilíbrio, mas o contrário, levando a bins que são influenciados pelos dados sintéticos que introduzimos artificialmente.

No artigo original do SMOTE, os autores propõem uma adaptação para dados de tipo misto, chamada *SMOTE Nominal Contínuo* (SMOTE-NC), que atribui a cada categoria não correspondente de um atributo categórico uma distância fixa igual à mediana dos desvios padrão das características numéricas para as amostras da classe minoritária. Portanto, ele usa apenas as informações fornecidas pela parte numérica das amostras, levando a suposições de distâncias perigosas caso apenas alguns dos atributos sejam numéricos. Além disso, ignora qualquer informação dos atributos categóricos, como a quantidade de categorias e sua distribuição.

Outra adaptação é proposta para características nominais puras, denominadas *SMOTE Nominal* (SMOTE-N), que usa uma métrica chamada Value Difference Metric [18] para caracterizar distâncias/ dissimilitudes. A forma como as amostras sintéticas são geradas promove a manutenção de peças comuns a qualquer um dos vizinhos. Isso faz com que amostras sintéticas provenientes de atributos com poucos valores sejam principalmente cópias, reduzindo a *sobreamostragem aleatória*.

Neste artigo, propomos uma estratégia de sobreamostragem adequada para todos os tipos de dados, ou seja, dados categóricos puros ou numéricos e do tipo misto, destinados a problemas binários. Chamamos isso de Generalized Synthetic Minority Oversampling TEchnique (GSMOTE) porque obtemos SMOTE original quando aplicado a dados numéricos puros, portanto, pode ser considerado como uma generalização de SMOTE. A razão para considerar o SMOTE original como ponto de partida é ver mais claramente o potencial da abordagem. No entanto, a generalização proposta pode ser feita diretamente a partir de quase qualquer variante.

No GSMOTE, as distâncias são calculadas por meio de uma transformação do Coeficiente Geral de Similaridade de Gower [19] em um coeficiente de distância. Neste novo Coeficiente de Distância Geral, a parte relacionada às feições categóricas foi substituída por uma distância derivada da medida de Entropia Variável (VE) [20], uma medida de similaridade originalmente desenvolvida para agrupamento hierárquico de dados categóricos.

A criação de amostras sintéticas de classes minoritárias a partir de pares de amostras existentes é desacoplada para partes categóricas e numéricas das amostras. Em ambas as partes mantém-se a mesma ideia empregada pelo SMOTE, ou seja, selecionar aleatoriamente um ponto no segmento que liga os dois pontos originais por meio de uma combinação convexa, sendo adaptado para as partes categóricas mantendo o conceito de proximidade aos extremos. As amostras sintéticas finais são obtidas pela união de ambas as partes.

O restante do trabalho está organizado da seguinte forma. Na Seção II, apresentamos todos os trabalhos anteriores relacionados ao nosso método. Na Seção III, apresentamos nossas contribuições, consistindo em um novo Coeficiente Geral de Distância (GDC), e na generalização do algoritmo SMOTE, que utiliza o GDC. A seção IV descreve os conjuntos de dados, esquema experimental e resultados alcançados. Por fim, na Seção V apresentamos as conclusões e possíveis trabalhos futuros.

II. Premissão

A. Coeficiente Geral de Similaridade

Para medir a semelhança entre dois pontos de dados de tipo misto, x_{eu} e x_j, Gower [19] definiu o Coeficiente Geral de Similaridade, dado por

$$sG(x_{eu}, x_j) = \sum_{k=1}^{d} \frac{1}{W(x_{ik}, x_{jk})} \sum_{k=1}^{d} W(x_{ik}, x_{jk}) s(x_{ik}, x_{jk}) \quad (1)$$

Onde $s(x_{ik}, x_{jk})$ é um componente de similaridade para o k-th atributo e $W(x_{ik}, x_{jk})$ é uma constante binária com valor 1 se a comparação entre os valores para o k-th atributo é válido, e 0 por outro lado.

Os componentes de similaridade são definidos de forma diferente dependendo dos tipos de atributos como segue

- Por *quantitativo* atributos, a semelhança $s(x_{ik}, x_{jk})$ é definido como

$$s(x_{ik}, x_{jk}) = 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$

  com R_k o intervalo do atributo.

- Por *binário* atributos, a similaridade é definida como

$$sb(x_{ik}, x_{jk}) = \begin{cases} 1, & \text{E se } x_{ik} = x_{jk} = verdadeiro \\ 0, & \text{por outro lado} \end{cases}$$

- Por *nominal* ou *categórico* atributos, a semelhança $s(x_{ik}, x_{jk})$ é definido como

$$s(x_{ik}, x_{jk}) = \begin{cases} 1, & \text{E se } x_{ik} = x_{jk} \\ 0, & \text{E se } x_{ik} \neq x_{jk} \end{cases}$$

B. Medida de similaridade de entropia variável

Dada uma variável aleatória X, tomando valores $x_1, \ldots, x_n$, *Entropia de Shannon* do X é definido como

$$H(X) = -\sum_{k=1}^{n} P(x_k) \log_b P(x_k)$$

Onde b é a base do logaritmo (geralmente tomando valores 2, e ou 10). Pode ser interpretada como a incerteza ou a informatividade inerente aos possíveis resultados de X, e isso é

profundamente relacionado com a variabilidade da variável aleatória. De fato, a entropia máxima acontece quando todos os valores da variável aleatória são igualmente prováveis, ou seja, incerteza máxima. Com essa interpretação em mente, pensamos em duas variáveis categóricas diferentes, uma com grande variabilidade e todos os valores quase igualmente prováveis, e outra com pequena variabilidade, tendo uma categoria predominante e as demais bastante esparsas. Nesse caso, é desejável que uma correspondência em qualquer categoria da primeira variável seja interpretada como maior similaridade do que uma correspondência na categoria predominante da segunda, mas menor similaridade que uma correspondência em qualquer uma das esparsas. Assim, uma medida de similaridade seguindo a filosofia acima deve dar maior peso às correspondências no caso de alta variabilidade, portanto, entropia.

Considerando as frequências relativas amostrais das categorias como suas probabilidades, a similaridade entre as categorias é definido como [20]

$$S_k(x_{ik}, x_{jk}) = \begin{cases} -\frac{1}{\ln N_k} \sum_{você=1}^{N_k} p_{você} \ln p_{você}, & \text{E se } x_{ik} = x_{jk} \\ 0, & \text{E se } x_{ik} \neq x_{jk} \end{cases} \quad (2)$$

Observe que, no caso de uma correspondência, teoricamente poderia assumir valores no intervalo de unidade [0,1],Incluindo0.Esse valor nulo só pode ser obtido se uma das variáveis for constante. No entanto, nesse caso essa categoria não é relevante para a similaridade, portanto o valor nulo parece razoável. Ao contrário, um valor1só é possível se todas as categorias em uma das variáveis forem igualmente prováveis.

Para determinar a semelhança entre duas amostras x_{eu}ex_j, nós definimos o*Entropia Variável*(VE) medida de similaridade como

$$S_{ve}(x_{eu}, x_j) = \frac{\sum_{k=1}^{d} S_k(x_{ik}, x_{jk})}{d} \quad (3)$$

Novamente, a medida de similaridade assume valores em [0,1],sendo os valores extremos acessíveis se e somente se todos$d$os valores médios assumem esse valor extremo exato.

C. Técnica de sobreamostragem de minoria sintética

A técnica Synthetic Minority Oversampling [17], brevemente descrita acima, é uma técnica de oversampling que, em sua versão original, gera amostras sintéticas de classes minoritárias dentro de vizinhanças de amostras de classes minoritárias selecionadas aleatoriamente, que são obtidas por$k$-Algoritmo de vizinhos mais próximos com um pré-definido$k$. As novas amostras sintéticas da classe minoritária são determinadas por pontos de amostragem aleatórios no segmento que conecta a amostra inicial com um de seus$k$vizinhos. Se denotarmos porx_{eu}a amostra de classe minoritária original central, e por$x_{m_1}, \ldots, x_{n_k}$EstáKvizinhos mais próximos da classe minoritária, então um dos$K$vizinhos é selecionado aleatoriamente (vamos denotar porx_j). Matematicamente, o segmento que une dois pontos é dado por sua combinação linear convexa. Neste caso, corresponde a

$$x_{novo} = \lambda x_j + (1 - \lambda) x_{eu}, \text{com} \lambda \in [0,1] \quad (4)$$

A seleção de um único ponto é obtida por amostragem de um λvalor. Observe que os dois limitesλvalores replicariam os pontos extremos do segmento.

III. Our abordagem

A. Coeficiente Geral de Distância

Com base no coeficiente de similaridade geral de Gower, e levando em conta a relação entre dissimilaridade e distância, podemos derivar o seguinte*Coeficiente Geral de Distância*, dado por

$$d_{gdc}(x_{eu}, x_j) = \left(\frac{1}{\sum_{k=1}^{d} W(x_{ik}, x_{jk})} \sum_{k=1}^{d} W(x_{ik}, x_{jk}) d_2(x_{ik}, x_{jk}) \right)^{-1/2} \quad (5)$$

Onde$d_2(x_{ik}, x_{jk})$é um componente quadrado da distância para o k-th atributo e$W(x_{ik}, x_{jk})$é uma constante binária com valor 1se a comparação entre os valores para ok-th atributo é válido, e0por outro lado.

Os componentes de distância são definidos de forma diferente dependendo dos tipos de atributos da seguinte forma

- Por*contínuo*e*ordinal*atributos, a distância $d(x_{ik}, x_{jk})$é definido como

$$d(x_{ik}, x_{jk}) = \frac{|x_{ik} - x_{jk}|}{R_k},$$

comR_ko intervalo do atributo.

- Por*quantitativo*atributos,$d(x_{ik}, x_{jk})$é definido como

$$d(x_{ik}, x_{jk}) = |x_{ik} - x_{jk}|$$

se nós padronizássemos$x_{\bullet k}$Como

$$x^*_{\bullet k} = \frac{x_{\bullet k} - \mu_k}{\sigma_k}$$

Ondeμ_ke$sigma_k$são, respectivamente, a média e o desvio padrão da$k$-th atributo, também pode ser normalizado, tal que

$$d(x_{ik}, x_{jk}) = \frac{\left| x^2_{ik} - x^2_{jk} \right|}{\sigma_k}.$$

- Por*binário*atributos, a distância é definida, também filosofia VE, como

$$d(x_{ik}, x_{jk}) = \begin{cases} 1 - \frac{1}{em\ 2} \sum_{você=1}^{2} p_{você} \ln p_{você}, & \text{E se } x_{ik} = x_{jk} = verdadeiro \\ 1, & \text{por outro lado} \end{cases}$$

Observe que a distância nula é alcançável se e somente se

$$p_1 = P(verdadeiro) = 1/2 = P(falso) = p_2$$

- Por*nominal*ou*categórico*(não*binário*) atributos, a distância$d($$x_{ik}, x_{jk})$é definida como a dissimilaridade obtida da medida de similaridade VE

$$d(x_{ik}, x_{jk}) = \begin{cases} 1 - \frac{1}{\ln N_k} \sum_{você=1}^{N_k} p_{você} \ln p_{você}, & \text{E se } x_{ik} = x_{jk} \\ 1, & \text{E se } x_{ik} \neq x_{jk} \end{cases}$$

Onde$N_k \geq 3$.

O motivo da separação em *contínuo e ordinal* e a *resto do quantitativo* atributos na forma de normalização, evitando o intervalo para o último, é a falta de determinação da variabilidade usando apenas as informações fornecidas pelos valores máximos e mínimos nesse caso. Em vez disso, o desvio padrão é mais razoável [21].

B. Técnica de sobreamostragem de minoria sintética generalizada

O algoritmo Generalized Synthetic Minority Oversampling TEchnique (GSMOTE) compreende as seguintes etapas:

1) Entrada:

(i) Um conjunto de dados D, com dimensionalidade d (ou seja, número de recursos de entrada) e um vetor de classe de destino binário, dado por

$$\{(x_{eu} = (x_{eu1}, \ldots, x_{Eu\,iria}), e_{eu}), eu = 1, \ldots, m\}$$

Onde m_p as amostras correspondem à classe positiva (minoritária) e m_n para a classe negativa (majoritária), com $m = m_p + m_n$; e onde $d_c \geq 0$ características são categóricas/binárias e $d_n \geq 0$ são numéricos, de modo que $d = d_c + d_n$.

(ii) Um pré-definido *número de vizinhos* a ser considerado para geração de amostras sintéticas (padrão $K = 5$)

(iii) Um pré-definido *relação de desequilíbrio máxima permitida* (predefinição $\beta = 0.1$).

2) Procedimento:

(i) Cálculo do *grau de desequilíbrio*, que pertence a (0,1],

$$b = m_p / m_n \tag{6}$$

(ii) Cálculo do *quantidade de amostras minoritárias sintéticas*, m_s, necessário para cumprir β limitação, ou seja

$$m_s = dm_n(1 - \beta) - m_p e \tag{7}$$

Onde dze denota a função teto, ou seja, o menor inteiro maior ou igual que z. Observe que, se nenhum desequilíbrio for permitido, então $\beta = 0$ e, portanto, $m_s = m_n - m_p$, como esperado.

(iii) Para cada amostra x_{eu} nós calculamos o seu K vizinhos mais próximos, usando *coeficiente geral de distância* na Equação 5.

(iv) Assumindo, sem perda de generalidade, que o primeiro d_c as amostras são categóricas/binárias e a última d_n são numéricos e denotam por P o subconjunto de amostras de classe minoritária

$$P = \{x_{eu} = (x_{eu1}, \ldots, x_{Eu\,iria}), eu = 1, \ldots, m_p\} \tag{8}$$

Para cada amostra

$$x_{eu} = (\underbrace{x_{eu1}, \ldots, x_{Eu\,iria_c}}_{\text{Gato/Lixo}}, \underbrace{x_{eu(d_c+1)}, \ldots, x_{Eu\,iria}}_{\text{Numérico}}) \tag{9}$$

Por razões operacionais na etapa de cálculo de caminho a seguir, mantivemos juntos os atributos categóricos e binários na separação. No entanto, eles foram tratados de forma diferente no cálculo da distância.

(v) Repetimos G vezes o seguinte processo

1) Selecione aleatoriamente uma amostra x_{eu} a partir de P.

2) Considere sua vizinhança e selecione aleatoriamente um de seus vizinhos, x_j. Divida ambas as amostras em suas partes categóricas/binárias e numéricas. Nós os denotamos por $x_{c\,eu}$ e $x_{c\,j}$, para partes categóricas/binárias, e por $x_{n\,eu}$ e $x_{n\,j}$ para numérico.

3) Calcule o caminho da parte categórica/binária de x_{eu}

$$x_{ceu} = (x_{eu1}, \ldots, x_{Eu\,iria_c})$$

para aquele de x_j

$$x_{cj} = (x_{j1}, \ldots, x_{jd_c})$$

ordenar as amostras do mais próximo para o mais distante x_{eu}, de acordo com *distância de Manhattan*. Neste caso, com nossa notação, essa distância é definida como

$$d_M(x_{c\,eu}, x_{cj}) = \sum_{k=1}^{d_c} |x_{ik} - x_{jk}|$$

Como exemplo, considere que $x_{ceu} = (1,1,3,2)$ e $x_{cj} = (1,3,3,1)$. Os passos no caminho, incluindo todos distâncias de Manhattan para $x_{c\,eu}$, seria

$$\begin{aligned} x_{ceu} = x_{\zeta_0} &= (1,1,3,2) \Rightarrow d_M = 0 \\ x_{\zeta_1} &= (1,2,3,2) \Rightarrow d_M = 1 \\ x_{\zeta_2} &= (1,1,3,1) \Rightarrow d_M = 1 \\ x_{\zeta_3} &= (1,2,3,1) \Rightarrow d_M = 2 \\ x_{\zeta_4} &= (1,3,3,2) \Rightarrow d_M = 2 \\ x_{cj} = x_{\zeta_5} &= (1,3,3,1) \Rightarrow d_M = 3 \end{aligned}$$

4) Faça uma amostra aleatória λ dentro [0,1], e depois

4.1. Aplique a Eq. 4 para as partes numéricas, obtendo o novo parte numérica sintética $x_{n\,novo}$, como no SMOTE original algoritmo.

4.2. Se denotarmos por n_s a quantidade de etapas intermediárias no caminho, e definimos $\rho = \frac{1}{n_s+1}$, podemos definir o seguinte partição do intervalo de unidade

$$[0,1] = \underbrace{[0, \rho \cup \tfrac{1}{2}]}_{EU_0} \cup \bigcup_{v=1}^{n_s} \underbrace{(\tfrac{2v-1}{2}\rho, \tfrac{2v+1}{2}\rho]}_{4} \cup \underbrace{(\tfrac{2n_s+1}{2}\rho, 1]}_{\text{Entradas}+1} \tag{10}$$

Então, obtemos o novo categórico/binário sintético parte selecionando $x_{c\,novo} = x_{c\,s_w}$. Onde W é assim $\lambda \in EU_W$. No exemplo anterior, $n_s = 4, \rho = 1/5$, então a partição do intervalo unitário seria

$$\underbrace{[0, \tfrac{1}{10}]}_{EU_0} \cup \underbrace{(\tfrac{1}{10}, \tfrac{3}{10}]}_{EU_1} \cup \underbrace{(\tfrac{3}{10}, \tfrac{1}{2}]}_{EU_2} \cup \underbrace{(\tfrac{1}{2}, \tfrac{7}{10}]}_{EU_3} \cup \underbrace{(\tfrac{7}{10}, \tfrac{9}{10}]}_{EU_4} \cup \underbrace{(\tfrac{9}{10}, 1]}_{EU_5}$$

Por exemplo instance, se $\lambda = 0.47 \in (\tfrac{3}{10}, \tfrac{1}{2}] = EU_2$, então $x_{c\,novo} = x_{c\,s_2} = (1,1,3,1)$.

5) Obtenha a nova amostra sintética x_{novo} juntando-se aos dois novos partes $x_{c\,novo}$ e $x_{n\,novo}$. Sob nossas suposições

$$x_{novo} = (x_{c\,novo}, x_{n\,novo}) \tag{11}$$

(vi) Adicionamos o novo G amostras minoritárias sintéticas para o conjunto de dados.

TABELA I
SINFORMAÇÕES ÚNICAS DOUCICONJUNTOS DE DADOS USADOS.

Nome	Alvo	Eu sou b. Razão	# Amostras	# Façanha.
Ecoli	imU	8,6: 1	336	7 num
Pen Dig.	Número 5	9,4: 1	10992	16 números
Vinho Qual.	Qualidade≤4	26: 1	4898	número 11
Doente Eute.	eutireóide doente	9,8: 1	3163	6 números, 36 gato
Avaliação do carro.	bom, bom	12: 1	1728	6 gato
Solar Fl.	Moderado = 0	19: 1	1389	10 gato

TABELA II
RRESULTADOS,SIGNIFICA±DST COM DEZ REPETIÇÕES,PARA DIFERENTES PORCENTAGENS DE CATEGORIA/DADOS BINÁRIOS EMECOLI, PPTDIGITS ECINEQCONJUNTOS DE DADOS DE UALIDADE.

Perc.	Método	Ecoli	Pen Dig.	Vinho Qual.
25%	Não	. 9123±.0140†	. 99984±10−5	. 8427±.0027†
	SM-NC	.9926±.0019	. 9981±10−5	. 8843±.0008†
	GSMOTE	.9970±0,0033	. 9993±10−6	.9942±.0008
50%	Não	.9176±.0107†	. 9984±10−5	. 8559±.0029†
	SM-NC	.9872±.0116	. 9988±10−5	.9865±.0017
	GSMOTE	.9931±.0014	. 9991±10−5	.9946±.0007
75%	Não	.9354±.0098†	. 9984±10−5	. 8614±.0032†
	SM-NC	.9822±.0087†	. 9990±10−5	.9933±.0007
	GSMOTE	.9934±.0021	. 9990±10−5	.9960±.0004

4. EEXPERIMENTOS

A. Esquema experimental

Consideramos seis conjuntos de dados do repositório UCI [22], da forma como foram propostos em [23]. A Tabela I resume as informações sobre eles.

Para fins de validação, compararemos o desempenho do GSMOTE com os conjuntos originais desbalanceados, ou seja, sem correção de desequilíbrio, no caso de dados categóricos/binários, e também com o SMOTE-NC em dados do tipo misto.

Neste último, a fim de verificar o comportamento dos algoritmos para diferentes razões de atributos categóricos versus numéricos, consideramos os três conjuntos de dados numéricos puros (Ecoli, Pen Digits e Wine Quality [24]). Para cada um deles, selecionamos aleatoriamente 25%, 50% e 75% de seus atributos, e os discretizamos por meio do algoritmo de discretização Chi2 [25]. Para mitigar possíveis desvios por causa da seleção de atributos por acaso, realizamos cada dez vezes. No primeiro, consideramos todos os seis conjuntos de dados descritos na Tabela I, em que todos os atributos numéricos também foram discretizados pelo algoritmo Chi2.

O algoritmo empregado para os testes é o Random Forest (RF) [26], por ser um algoritmo poderoso e bastante eficiente para dados do tipo misto, por ser baseado em árvores de decisão. Para a seleção dos hiperparâmetros do modelo, realizamos uma busca exaustiva em grade, variando o número de árvores (em $\{50,100,200\}$), a profundidade máxima de cada árvore (em $\{5,10,20,\infty\}$) e o $\sqrt{}$ quantidade de atributos visíveis em cada nó de divisão (em $\{d, d\}$), juntamente com uma estratégia de validação cruzada de 10 vezes, usando a área sob a curva do operador do receptor (AUC) para pontuação porque é robusta contra o efeito do desequilíbrio na ausência de uma abordagem sensível ao custo [27]. Além disso, como a floresta aleatória é um método estocástico, repetimos cada procedimento 10 vezes. Portanto, para cada método de aprendizado de desequilíbrio, fizemos 720 tentativas por conjunto de dados, incluindo os originais.

B. Resultados e discussão

Os resultados discutidos são apresentados a seguir. Primeiro, no caso de dados do tipo misto, relatamos os resultados da pontuação média das 10 repetições de RF, para os melhores hiperparâmetros da grade em cada caso na Tabela II. Por último, no caso de dados categóricos/binários puros, os resultados são apresentados na Tabela III. Em ambos os casos, foram realizados testes de significância estatística por meio do teste de postos sinalizados de Wilcoxon [28], sendo indicado nas tabelas caso a hipótese nula tenha sido rejeitada (significando diferença significativa) quando comparada com a melhor (em negrito).

No caso da Ecoli, podemos apreciar que o desempenho de ambas as abordagens é cerca de 10% melhor em média do que o conjunto de dados desequilibrado. Apesar de não se poder dizer que o GSMOTE se comporta muito melhor, é um pouco melhor na maioria das vezes, independentemente da porcentagem de atributos categóricos. Deve-se notar que a dimensionalidade deste conjunto de dados é baixa, portanto não há muita margem para apreciar a influência do tipo de estimativa feita pelo SMOTE-NC com base nos desvios padrão, exceto talvez 75%. De fato, este é o caso em que a diferença das AUCs médias é maior (2,5 vezes maior do que para 25%). Para dados categóricos puros, podemos apreciar que a pontuação AUC também é alta para atributos categóricos de 100%, mas inferior às médias do restante das porcentagens mais baixas.

Quando se trata de dados mistos no conjunto de dados Pen Digits, a margem para melhoria é realmente baixa porque a AUC é muito alta mesmo sem corrigir o desequilíbrio. No entanto, isso pode ser visto como um desafio que ambas as abordagens superaram. Novamente, o GSMOTE supera o SMOTE-NC em vitórias individuais, mas desta vez a maior diferença acontece com a menor proporção de recursos categóricos. Sem um estudo exaustivo da natureza e características dos conjuntos de dados, fora do escopo deste artigo, é difícil encontrar uma razão para isso. A tendência permanece a mesma para dados categóricos completos, incluindo a margem apertada para melhoria (ver Tabela III).

Agora, em dados de tipo misto para o conjunto de dados Wine Quality, podemos dizer que o GSMOTE é claramente o melhor algoritmo, vencendo todas as vezes e aumentando a pontuação AUC em cerca de 15% em relação aos dados originais desbalanceados. Também é perceptível que, nos três casos, quanto maior o percentual de características categóricas, melhor o desempenho. Isso não é mantido no caso de dados categóricos puros, como pode ser visto na Tabela III. Finalmente, a Tabela III apresenta os resultados em dados categóricos/binários puros para todos os conjuntos de dados. Em todos os casos, é benéfico realizar a correção do desequilíbrio. Além dos três conjuntos de dados numéricos, comentados anteriormente, o Eurotireoids é um exemplo claro da possibilidade de obter um conjunto do tipo misto e discretizar todos os atributos numéricos (6 a 36 neste caso) antes de corrigir o desequilíbrio, levando a resultados bem-sucedidos.

TABELA III
RESULTADOS PARA PURA CATEGORICAL/DADOS BINÁRIOS.

Método	Ecoli	Caneta	Vinho	Doente	Carro	Solar
Não	. 9477 †	. 9998.	8602 †	. 9835 †	. 9871.	7919 †
GSMOTE	. 9926	. 9999	. 9517	. 9950	. 9966	. 9030

e SolarFlare, originalmente categórica pura, também se beneficiam da correção (14% no caso de SolarFlare).

V. CINCLUSÃO

Apresentamos o SMOTE Generalizado, um método de sobreamostragem de geração de protótipos que pode ser empregado em dados puramente numéricos e categóricos, bem como em dados do tipo misto. No caso particular de dados numéricos puros, reduz-se ao SMOTE original.

A principal vantagem sobre outras técnicas existentes é o melhor aproveitamento das informações inerentes, pois leva em consideração a quantidade de categorias e sua distribuição no subespaço da classe minoritária, a fim de atribuir as distâncias entre amostras na busca de vizinhança. Além disso, desacopla a geração de protótipos em partes categóricas e numéricas, sendo capaz de manter a noção de proximidade implícita na forma como o SMOTE gera amostras sintéticas nos segmentos que conectam as originais.

Os resultados obtidos em seis conjuntos de dados desbalanceados públicos diversos suportam o GSMOTE como um método poderoso.

UMAAGRADECIMENTO

Este trabalho foi parcialmente financiado pelo projeto Integração Semântica de Conhecimento para Filtragem de Spam Baseada em Conteúdo (TIN2017-84658-C2-2-R) do Ministério da Economia, Indústria e Competitividade da Espanha (SMEIC), Agência Estatal de Pesquisa (SRA) e Fundo Europeu de Desenvolvimento Regional (FEDER), e pelo projeto REMEDY - Real Time control and embedded security do Departamento de Desenvolvimento Económico e Infraestruturas do Governo Basco ao abrigo da convenção de subvenção KK-2021/00091. Foi desenvolvido pelo grupo de sistemas inteligentes para sistemas industriais apoiado pelo Departamento de Educação, Política de Línguas e Cultura do Governo Basco.

REFERÊNCIAS

[1] Cernuda C.: Sobre a relevância do pré-processamento na manutenção preditiva de sistemas dinâmicos. In: Lughofer E., Sayed-Mouchaweh M. (eds) Manutenção preditiva em sistemas dinâmicos. Springer,Capítulo 3, 53–93 (2019)

[2] Davis DN, Nguyen TTT: Gerando e Verificando Modelos de Previsão de Risco Usando Mineração de Dados (Um Estudo de Caso da Cardiovascular Medicine). 57º Congresso Anual da Sociedade Europeia de Cirurgia Cardiovascular (ESCVS), Barcelona, Espanha (2008)

[3] Mar J., Gorostiza A., Arrospide A. et al.: Estimativa da epidemiologia da demência e sintomas neuropsiquiátricos associados aplicando aprendizado de máquina a dados do mundo real. Jornal de Psiquiatria e Saúde Mental (2021). DOI:*https://doi.org/10.1016/j.rpsm.2021.03.001*

[4] Laza R., Pavon R., Reboiro-Jato M., Fdez-Riverola F.: Avaliando o efeito de dados não balanceados na classificação de documentos biomédicos. Jornal de Bioinformática Integrativa8(3), 1–13 (2011)

[5] Rahman MM, Davis DN: Abordando o problema de desequilíbrio de classe em conjuntos de dados médicos. Revista Internacional de Aprendizado de Máquina e Computação3(2), 224–228 (2013)

[6] Mar J., Gorostiza A., Ibarrondo O. et al.: Validação de modelos de aprendizado de máquina florestais aleatórios para prever sintomas neuropsiquiátricos relacionados à demência em dados do mundo real. Jornal da Doença de Alzheimer77(2), 855–864 (2020)

[7] Phua, C., Alahakoon, D.: Relatório de minorias na detecção de fraudes: Classificação de dados distorcidos. Boletim ACM SIGKDD Explorations6(1), 50–59 (2004)

[8] Thai-Nghe N., Gantner Z., Schmidt-Thieme L.: Métodos de aprendizado sensíveis ao custo para dados desbalanceados. A Conferência Conjunta Internacional de 2010 sobre Redes Neurais (IJCNN), Barcelona Espanha, 1–8 (2010)

[9] Attenberg J., Ertekin S.: Desequilíbrio de Classe e Aprendizagem Ativa. Em: Ele H., Ma Y. (eds) Aprendizagem Desequilibrada: Fundamentos, Algoritmos e Aplicações. John Wiley & Sons, Capítulo 6, 101–149 (2013)

[10] Fernández A., García S., Galar M., Prati RC, Krawczyk B., Herrera F.: Aprendizagem Sensível ao Custo. In: Aprendendo com conjuntos de dados desequilibrados. Springer, Cham, Capítulo 4, 63–78 (2018)

[11] Chawla, NV: C4.5 e conjuntos de dados desequilibrados: Investigando o efeito do método de amostragem, estimativa probabilística e estrutura de árvore de decisão. Proceedings of the ICML'03 Workshop on Learning from Desbalanced Data sets, Washington, DC, EUA (2003)

[12] Drummond, C., Holte, R.: C4.5, desequilíbrio de classe e sensibilidade ao custo: Por que subamostragem supera superamostragem. Anais do Workshop ICML'03 sobre Aprendizagem com Conjuntos de Dados Desequilibrados, Washington, DC, EUA (2003)

[13] Maloof, M.: Aprendendo quando os conjuntos de dados estão desequilibrados e quando os custos são desiguais e desconhecidos. Anais do Workshop ICML'03 sobre Aprendizagem com Conjuntos de Dados Desequilibrados, Washington, DC, EUA (2003)

[14] Kubat, M., Matwin, S.: Abordando a maldição dos conjuntos de treinamento desequilibrados: seleção unilateral. Anais da Décima Quarta Conferência Internacional sobre Aprendizado de Máquina, 179-186, Nashville, Tennessee. Morgan Kaufmann (1997)

[15] Japkowicz, N.: O Problema do Desequilíbrio de Classes: Significado e Estratégias. Anais da Conferência Internacional sobre Inteligência Artificial (IC-AI'2000): Curso Especial sobre Aprendizagem Indutiva, 111–117, Las Vegas, Nevada, EUA (2000)

[16] Batista, GEAPA, Prati, RC, Monard, MC: Um estudo do comportamento de vários métodos para balancear dados de treinamento de aprendizado de máquina. Boletim ACM SIGKDD Explorations6(1), 20–29 (2004)

[17] Chawla, NV, Bowyer, KW, Hall, LO, Kegelmeyer, WP: SMOTE: Técnica de sobreamostragem de minoria sintética. Jornal de Pesquisa em Inteligência Artificial16,321–357 (2002)

[18] Custo S., Salzberg S.: Um Algoritmo do Vizinho Mais Próximo Ponderado para Aprendizagem com Características Simbólicas. Aprendizado de máquina10(1), 57-78 (1993)

[19] Gower, JC: Um coeficiente geral de semelhança e algumas de suas propriedades. Biometria27(4), 857-871 (1971)

[20] Sulc Z., Rezankova H.: Comparação de medidas de similaridade para dados categóricos em agrupamento hierárquico. Diário de Classificação35(1), 58–72 (2019)

[21] Wishart D.: Agrupamento de k-médias com detecção de outliers, variáveis mistas e valores ausentes. In: Schwaiger M. e Opitz O. (edts) Análise de Dados Exploratórios em Pesquisa Empírica, Springer, 216–226 (2003)

[22] Dua D., Graff C.: UCI Machine Learning Repository. Irvine, CA: Universidade da Califórnia, Escola de Informação e Ciência da Computação (2019). *https://archive.ics.uci.edu/ml/*

[23] Zejin D.: Classificadores de conjuntos diversificados para aprendizagem de dados altamente desequilibrada e sua aplicação em bioinformática. Dissertação, Georgia State University (2011)

[24] Cortez P., Cerdeira A., Almeida F., Matos T., Reis J.: Modelagem de preferências de vinho por mineração de dados a partir de propriedades físico-químicas. Sistemas de Suporte à Decisão47(4), 547–553 (2009)

[25] Liu H., Setiono R.: Seleção de recursos via discretização. Transações IEEE em Engenharia de Conhecimento e Dados9(4), 642-645 (1997)

[26] Breiman, L.: Florestas aleatórias. Aprendizado de Máquina, v. 45(1), pp. 5 - 32 (2001)

[27] Ferri C., Flach P., Orallo J., Lachice N. (edts). Primeiro Workshop ECAI'2004 em Análise ROC em Inteligência Artificial (2004)

[28] Wilcoxon F.: Comparações individuais por métodos de classificação. Boletim Biométrico1(6), 80-83 (1945)