Estudos Empíricos dos Métodos de Balanceamento para a Classificação

Daiany Francisca Lara
Universidade do Estado de Mato Grosso - UNEMAT
Colíder, Brasil
dflara@unemat.br
Aurora Trinidad Ramirez Pozo
Universidade Federal do Paraná - UFPR
Curitiba, Brasil
aurora@inf.ufpr.br
Léo Manoel Lopes da Silva Garcia
Universidade do Estado de Mato Grosso - UNEMAT
Colíder, Brasil
leoneto@unemat.br
Cláudia Alves Perez
Faculdade do Pantanal - FAPAN
Cáceres, Brasil
claudiaperez$_{@}hotmail.com$
Franciano Antunes
Universidade do Estado de Mato Grosso - UNEMAT
Cáceres, Brasil
franciano@unemat.br

***Resumo*—A classificação tem o objetivo de rotular eventos ou objetos de acordo com classes pré-estabelecidas. No entanto, a maioria dos algoritmos perdem a capacidade de predição, quando o conjunto de dados possui uma distribuição desbalanceada entre suas classes. Para tentar resolver esse problema diversos métodos tęm sido propostos na literatura. O presente estudo tem como objetivo analisar e comparar os métodos mais conhecidos que se propõem a resolver o problema de classificação com bases desbalanceadas. Para isto, os métodos foram testados usando cinco classificadores tradicionais, e 13 bases provenientes do UCI *Machine Learning Repository*. Os resultados demonstram que é possível melhorar a taxa de classificação, mas é difícil dizer o método que se comporta melhor, pois tudo depende de como o algoritmo de classificação generaliza a base.**

I. Introdução

As informações geradas pelos processos automatizados tem aumentado de forma considerável e estão dispostas e agrupadas de maneira irregular e desbalanceada, tornando difícil o entendimento dessas informações. De forma a aprimorar a análise dessas informações, algoritmos de aprendizado supervisionados como os classificadores são frequentemente utilizados [1]. Contudo, a utilização de bases com a distribuição desbalanceada entre as classes representa um dos aspectos que tem comprometido significativamente o desempenho dos algoritmos de classificação, pois não estão preparados para induzir bons modelos [2]. Normalmente, os resultados obtidos com esse tipo de base apresentam uma boa acurácia para a classe majoritária, e uma acurácia baixa para a classe minoritária. O que pode ser um problema quando a classe de interesse é justamente a classe minoritária.

Trabalhos encontrados na literatura apresentam métodos para tentar resolver o problema da classificação com classes desbalanceadas, tentando evitar os problemas causados pelos métodos aleatórios *Oversampling e Undersampling* [3][4][2]. Na maioria desses trabalhos, os testes dos métodos de balanceamento são realizados no máximo com três algoritmos de classificação. Os classificadores mais utilizados são: a árvore de decisão *C4.5*, e o classificador *NaiveBayes* e RIPPER. Geralmente, a métrica de avaliação utilizada é apenas a acurácia, o que pode ser um problema, pois nem sempre os métodos conseguem melhorar a classe minoritária, as vezes a melhora acontece apenas na classe majoritária, e que consequentemente melhora a acurácia.

Este trabalho faz uma análise comparativa dos principais métodos encontrados na literatura, com o intuito apresentar os métodos que realmente melhoram a taxa de acurácia e principalmente os valores dos verdadeiros positivos representada pela classe minoritária, e também melhorar, ou no mínimo manter os resultados da classe majoritária. Para tanto, foram testados esses métodos com classificadores que neste trabalho são chamados de tradicionais, como: *Naive Bayes*, *Bayes Net*, *SMO*, *MultilayerPerceptron*, *J48* e *JRip*. As métricas de avaliação utilizadas são: RecallP, RecallN e Acurácia. Os resultados serão demonstrados de acordo com os valores dessas métricas, pois dessa forma pode-se observar o que realmente acontece em cada classe. Os testes foram realizados com 13 bases

provenientes do repositório UCI.

Os métodos de balanceamento utilizados para a análise na fase de pré-processamento apresentados neste trabalho são: *Adasyn, Borderline, Smote, Oversampling e Undersampling*. A análise é feita também com os métodos conhecidos como Híbridos, pois faz uma junção dos métodos de balanceamento na fase de pré-processamento com a classificação, sendo eles: *BalanceCascade, EasyEnsemble, SMOTEBoost e SMOTEBoosting*.

O objetivo deste trabalho, de acordo com os resultados obtidos nos experimentos, é responder algumas questões que são de grande importância, como: (i) A Classificação pode ser melhorada aplicando algum método de balanceamento? (ii) Existe um método que sempre se comporta bem? (iii) Os classificadores podem influenciar nos resultados? (iv) O que significa melhorar a taxa de classificação da classe minoritária e no mínimo manter a taxa de classificação da classe majoritária? (v) Quais são as métricas de avaliação que realmente apontam a melhora da classe minoritária?

O Artigo está organizado da seguinte forma: Na Seção II são apresentadas as etapas do processo extração de conhecimento. Na Seção III estão relacionados os problemas causados pelo conjunto de dados desbalanceados e os métodos para resolver esses problemas. As ferramentas e classificadores utilizados neste trabalho assim como os experimentos com bases do UCI são apresentados nas Seções IV e V, respectivamente. E finalmente as conclusões são apresentadas na Seção VI.

II. Descoberta de Conhecimento em Base de dados - KDD

O processo de Descoberta de Conhecimento em Base de Dados (Knowledge Discovery in Databases - KDD) tem como intuito extrair conhecimento de grandes bases de dados, e abrange cinco etapas que exigem do usuário capacidade de análise e de tomada de decisão,as etapas envolvidas são: Seleção, Pré-processamento, Transformação, Mineração de Dados e Interpretação [Fayyad et al. 1996].

O pré-processamento tem o objetivo de aprimorar a qualidade dos dados coletados, visto que frequentemente os dados apresentam diversos problemas, como grande quantidade de valores desconhecidos, ruídos, desproporção entre o número de classe, dentre outros [2].

A mineração de dados é umas das etapas mais importantes do processo KDD, pois caracteriza-se pela existęncia de um algoritmo que diante da tarefa proposta será eficiente em extrair conhecimento implícito e útil de um banco de dados. As tarefas de mineração de dados podem extrair diferentes tipos de conhecimento, sendo necessário decidir, já no início do processo de mineração de dados, qual o tipo de conhecimento que o algoritmo deve extrair [5]. As principais tarefas de mineração de dados tanto supervisionadas como não supervisionadas são: classificação, regressão, regras de associação e agrupamento.

III. Classificação em conjuntos de dados Desbalanceados

As informações do mundo real são distribuídas de maneira desigual, principalmente em situações nas quais determinadas classes são mais difíceis de se obter, causando assim o problema de classes desbalanceadas.

Muitos aspectos podem influenciar o desempenho do modelo criado por um sistema supervisionado. Um desses aspectos está relacionado com a distribuição entre o número de exemplos pertencentes a cada uma das classes, ou seja, uma base de dados é dita desbalanceada, quando o número de exemplos de uma classe é maior que o número de exemplos da outra classe. Tal problema pode prejudicar o desempenho de classificação, pois tende a classificar exemplos da classe minoritária como sendo da classe majoritária [2].

Existem métodos que consistem em efetuar ajustes no conjunto de dados uniformizando uma distribuição de exemplos entre as classes utilizando amostragem (em ingleş, *sampling*). Dentre os métodos existentes na literatura, seja para adicionar ou remover instâncias das classes, os mais conhecidos são: Oversampling (*Sobreamostragem*) aleatória, que replica exemplos da classe minoritária com o objetivo de obter uma distribuição mais balanceada [6]. E Undersampling (*Subamostragem*) aleatória, que tem como objetivo balancear o conjunto de dados pela eliminação de exemplos da classe majoritária [6]. São métodos equivalentes no que diz respeito ao balanceamento artificial das classes, mas com um conjunto próprio de problemas consequentes que podem atrapalhar o aprendizado. O problema causado pelo *undersampling* acontece na eliminação de exemplos da classe majoritária fazendo com que o classificador perca informações importantes pertencentes a essa classe. No caso do *oversampling* aumenta a probabilidade de ocorrer *overfitting* [7].

A. Oversampling Orientado

Uma das técnicas mais conhecidas que tem se mostrado como um poderoso método em várias aplicações é o SMOTE (*Synthetic Minority Oversampling Technique*). Este método cria dados artificiais com base nas semelhanças existentes no espaço de característica entre os exemplos da classe minoritária. O algoritmo funciona criando exemplos sintéticos baseados em exemplos minoritários e seus k vizinhos mais próximos.

$$x_{new} = x_i + (\hat{x}_i - x_i)\delta \quad (1)$$

Esse método pode aumentar a taxa de acurácia, mas ocorre o efeito indesejado de modelos muito específicos para estes casos replicados, prejudicando a generalização para a classe de interesse, fato conhecido como overfitting. Para resolver esse problema vários métodos adaptativos como Borderline-SMOTE e ADASYN (Adaptative Synthetic Sampling) tęm sido propostos [1], para que esses novos dados sejam gerados na vizinhança de cada caso da classe minoritária, de forma a fazer crescer a região de decisão e, assim, aumentar o poder de generalização dos classificadores [8].

Borderline-SMOTE - O algoritmo Borderline-SMOTE tem o objetivo de identificar, as amostras próximas a superfície de decisão, essas amostras são identificadas como conjunto DANGER. Este conjunto é formado por instâncias que possuem mais vizinhos da classe majoritária do que da classe minoritária. O Borderline-SMOTE então irá aplicar o SMOTE somente nas instâncias do conjunto DANGER [2][1].

A Figura 1 mostra a ocorrência de um ruído e nenhum exemplo sintético gerado por ele, como mostra a instância C [1].

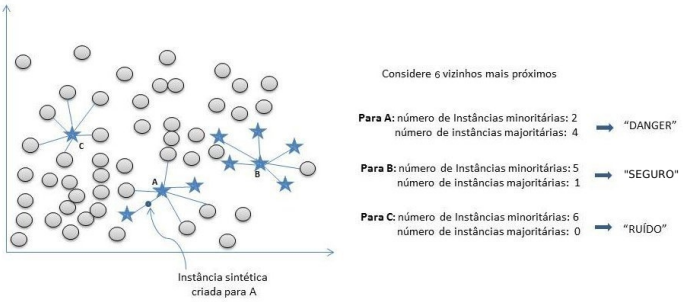

Figura 1. Geração de dados baseados no método Borderline [1]

ADASYN - O algoritmo ADASYN é baseado no algoritmo SMOTE, o objetivo é reduzir o viés introduzidos pela distribuição desbalanceada, e principalmente deslocar adaptativamente a fronteira de decisão para focar nos exemplos mais difíceis de aprender. A ideia, é usar a distribuição de densidade como um critério para decidir automaticamente o número de dados sintéticos que precisam ser gerados para cada exemplo da classe minoritária.

Primeiramente é obtido o grau de desbalanceamento da classe pela Equação $d = \frac{m_p}{m_n}$. Onde m_p é o número de exemplos da classe minoritária e m_n é o número de exemplos da classe majoritária, e $d \in (0, 1)$. Se $d < d_{th}$ então o cálculo do número de exemplos sintéticos que precisam ser gerados, é feito pela Equação $G = (m_n - m_p).\beta$. Onde $\beta \in (0, 1)$ e é um parâmetro usado para especificar o nível de balanceamento desejado depois da geração dos dados sintéticos. Para cada exemplo x_i no conjunto de instâncias positivas P, sendo i = 1,...,P, são encontrados os K vizinhos mais próximos baseados na distância euclidiana no espaço dimensional n, e então calcular a proporção de $D_i = \frac{\Delta_i}{k}$. Onde Δ_i é o número de exemplos dos k vizinhos mais próximos de x_i que pertence a classe majoritária, portanto $D_i \in (0, 1)$. Uma normalização é feita de D_i para $\hat{D}_i = D_i / \sum_{i=1}^{mp} D_i$, de modo que $\hat{D}_i$ é a distribuição de densidade ($\sum_i \hat{D}_i = 1$). Então o número de exemplos sintéticos que precisam ser gerados para cada exemplo minoritário x_i é dado pela Equação $g_i = \hat{D}_i \times G$. Onde G é o total de números de exemplos minoritários que precisam ser gerados para a classe minoritária. A ideia principal do ADASYN é usar a distribuição de densidade $\hat{D}_i$ como um critério para decidir automaticamente o número de exemplos sintéticos que precisam ser gerados para cada exemplo da classe minoritária.

Cluster baseado em Oversampling - Algoritmos de Cluster baseados em oversampling tem como objetivo minimizar o problema de pequenos disjuntos que são conceitos aprendidos que cobrem poucos casos do conjunto de treinamento, podendo ser aplicável a conjuntos de dados multiclasses. Utiliza técnica de agrupamento k-means, ou seja, leva um conjunto aleatório K exemplos de cada cluster e calcula a média do vetor de características desses exemplos, chamados como centros do cluster. Em seguida, o restante dos exemplos de treinamento é apresentado um de cada vez, e para cada exemplo, é feito o cálculo da distância euclidiana entre ele e cada centro do cluster. Então cada exemplo de treinamento é atribuído ao cluster que apresenta a menor magnitude do vetor de distância. Todos os meios de cluster são atualizados e o processo é repetido até que todos os exemplos sejam esgotados [1], [8].

B. Undersampling Orientado

Xu-Ying Liu [9] apresenta dois exemplos de métodos que tem o objetivo de reduzir a perda de informações causadas pelo método tradicional de undersampling aleatória. Esses métodos são conhecidos como híbridos, pois pertencem a um grupo de métodos que juntam os algoritmos de pré-processamento com classificação. Os algoritmos apresentados são: EasyEnsemble e BalanceCascade.

EasyEnsemble - Aprendizagem considerada não supervisionada, pois explora o conjunto N (instâncias da classe majoritária) utilizando uma amostragem aleatória com substituição [9]. A implementação do EasyEnsemble é considerada simples, pois dado um conjunto de treinamento minoritário P, e o conjunto de treinamento majoritário N, criam-se t subconjuntos $N_i{}', i = 1, ..., i = t$, compostos de exemplos retirados de N, onde $N_i{}' < |N|$, e utiliza-se comumente $|N_i{}'| = |P|$. Os subconjuntos $N_i{}', ..., N_t{}'$ são criados e treinados separadamente, onde um classificador H_i é treinado usando $N_i{}'$ e todos os exemplos de P (O classificador H_i utilizado foi o *Adaboost*). Todos os classificadores gerados são combinados para uma decisão final [2] [9].

BalanceCascade - Considerada uma abordagem supervisionada por explorar exemplos da classe majoritária *N* de um conjunto de treinamento. A ideia é depois que o H_1 é treinado, se um exemplo $x_1 \in N$ for corretamente classificado por H_1, é aceitável pressupor que x_1 é um pouco redundante em N, dado que H_1 já aprendeu a classificar esse exemplo. Desta forma pode-se remover alguns exemplos classificados corretamente do conjunto N [9]. No BalanceCascade, a dependência sequencial entre classificadores é principalmente explorada para reduzir a informação redundante na classe principal. Assim a estratégia de amostragem leva para um espaço de amostra restrita para o seguinte processo undersampling, na esperança de explorar o máximo de informação útil possível [9].

1) Integração de Amostragem e Boosting: Chawla [10] propôs um método que faz uma combinação da técnica SMOTE com o procedimento de Boosting. O SMOTE é utilizado para melhorar a predição das classes minoritárias, e o Boosting para não sacrificar a precisão do conjunto de dados, ou seja, é um método iterativo de combinação de classificadores, onde a cada iteração um classificador é induzido a partir de uma amostra de instâncias do conjunto de treinamento [10].

IV. FERRAMENTAS

Um estudo empírico foi realizado com diferentes bases de dados e métodos de balanceamento com o objetivo de entender mais profundamente os efeitos de cada um dos métodos. Para a realização dos experimentos foram utilizadas algumas ferramentas como: (i) O software Weka (*Waikato*

Environment for Knowledge Analysis) que foi desenvolvido pela Universidade de Waikato na Nova Zelândia; (ii) e o Software Keel (*Knowledge Extraction based on Evolutionary Learning*), desenvolvida pelos projetos nacionais Espanhóis.

No processo de classificação, foram utilizados 6 (seis) algoritmos implementados no Software Weka e que neste trabalho são chamados de Classificadores tradicionais. Os classificadores são: *Naive Bayes, Bayes Net, SMO, J48, JRip e MultilayerPerceptron*.

Para os experimentos deste trabalho, foram utilizadas bases do repositório UCI *Machine Learning Repository*. O repositório UCI, abrange bases relacionadas a diversos problemas, e estão disponíveis para todos que precisam trabalhar com aprendizagem de máquina. [11]. As Bases utilizadas nos experimentos e suas principais características estão relacionadas na Tabela I.

Tabela I
CARACTERISTICAS DA BASE UCI

Base	Instâncias	Atributos	Classe	
			Positivo	Negativo
Breast	683	10	239	444
Bupa	345	7	145	200
Ecoli	336	8	35	301
Flag	194	29	17	177
Glass	214	10	17	197
Haberman	306	4	81	225
Heart	270	14	120	150
Ionosphere	351	34	126	225
Lettera	19999	17	789	19210
New-thyroid	215	6	35	180
Pima	768	9	500	268
Satimage	6435	37	626	5809
Vehicle	846	19	199	647

A. *Métricas de avaliação*

Na classificação, a utilização de alguma métrica para avaliar os resultados obtidos é necessária, e um exemplo para essa avaliação é a matriz de confusão que mostra o número de classificações corretas em relação as esperadas por cada classe. A matriz de confusão pode representar o problema de duas classes como também multiclasses, e o número de acertos da classe se encontra na diagonal principal da matriz [2].

As métricas que são relevantes para os problemas de classes desbalanceadas são: Precisão, Recall e F-measures, métricas essas que também utilizam as variáveis da matriz de confusão, e são apropriadas quando se está preocupado com o desempenho das classes [2].

V. RESULTADOS

Primeiramente, foram aplicados os classificadores tradicionais nas bases originais com o objetivo de poder observar o comportamento destes, com as bases que possuem um alto nível de desbalanceamento. Vale lembrar que todas as tabelas estão no **Apêndice**.

A Tabela III apresenta os resultados das bases retiradas do repositório UCI, classificadas com os algoritmos tradicionais. É possível perceber que os classificadores nem sempre conseguem predizer corretamente algumas bases, principalmente para a classe positiva (RecallP), pois provavelmente estas bases possuem um número menor de exemplos que determinam essa classe, ou também, podem possuir os problemas como ruídos, dificultando assim o poder de generalização do classificador.

Com o intuito de melhorar os resultados obtidos anteriormente, foram aplicados os métodos de balanceamento (Adasyn, Borderline, Smote, Oversampling e Undersampling). Os resultados para esse experimento são apresentados na Tabela V.

A seguir, é feito uma análise por base, considerando os diferentes métodos de balanceamento com os diferentes classificadores (Tabela V). Uma comparação é feita com os resultados obtidos com a base original (Tabela III), o objetivo é poder dizer os métodos que se destacam.

Breast - é possível perceber que o método Oversampling utilizando os classificadores NaiveBayes, J48 e BayesNet, conseguem melhorar os valores da classe positiva mantendo a classe negativa. Outro método, que se destaca nesta base melhorando ambas as classes é o Borderline classificado pelo J48 e JRip. As 4 bases **Bupa, Ecoli, Flag e Haberman**, conseguem uma melhora na classe positiva mas diminui a classe negativa, e principalmente a taxa de classificação. Vale lembrar que as bases Flag e Ecoli possuem um nível de desbalanceamento razoável em relação a Bupa e Haberman, o que não justifica a não melhora, este fato pode estar relacionado a problemas de generalização ou ruídos. **Heart -** O método Undersampling é o mais frequente nesta base, ou seja, este melhora os valores da classe positiva e mantem da classe negativa com o classificador NaiveBayes e melhora os valores de ambas as classes com o J48. O classificador J48 também apresenta melhoras com os métodos Borderline e Smote. **Ionosphere -** O método Undersampling melhora as classes positivas e negativas, consequentemente melhorando a taxa de classificação, isto é possível com a utilização dos classificadores J48 e BayesNet. Outro método que consegue uma significativa melhora nesta base é o Borderline com os classificadores NaiveBayes e J48. **Lettera -** O método Smote consegue melhorar a classe positiva e manter os valores da classe negativa com os classificadores J48 e JRip. O classificador BayesNet melhora ambas as classes com os métodos Borderline e Smote. **New-Thyroid -** Os classificadores NaiveBayes, MultilayerP e J48 conseguem melhorar a classe positiva e manter a classe negativa com o método Oversampling. O método Adasyn consegue também melhorar e manter com o classificador J48 e JRip. Acontece o mesmo com o método Smote e classificado pelo JRip. **Pima -** Nesta base, o único método que consegue melhorar significativamente o valores das classes e taxa de classificação é o Adasyn, classificado pelo algoritmo NaiveBayes. **Satimage -** É possível perceber que a melhora na classe positiva é significativa na maioria dos métodos e classificadores, mas o método que consegue melhorar ambas as classes nesta base é o Oversampling com o classificador BayesNet. **Vehicle -** Nesta base a maioria dos métodos conseguem uma melhora na maioria dos classificadores. O método Adasyn classificado pelo J48 e JRip melhora os valores das classes positivas e negativas e consequentemente a Acurácia. O método Smote melhora todos os valores com os classificadores NaiveBayes, J48 e JRip. O método Oversampling com os classificadores NaiveBayes e BayesNet e o método Undersampling com os

classifacadores NaiveBayes, JRip e BayesNet.

Considerando os diferentes métodos de balanceamento com os diferentes classificadores, é possível perceber uma frequęncia de alguns métodos em cada base, ou seja, alguns métodos conseguem melhorar os valores de RecallP e RecallN com diversos classificadores. De acordo com essa frequęncia uma análise foi feita entre os classificadores, com o intuito de comparar esse resultado com o resultado da base original utilizando o mesmo classificador. A Tabela II apresenta a porcentagem de melhoria do método que se destaca com a base original.

Tabela II
% DE MELHORIA DOS MÉTODOS COM A BASE ORIGINAL

Base	Métodos	Classificador	RecallP	RecallN	Acurácia
breast	Original	Jrip	0,958	0,953	95,460
	Borderline	Jrip	0,975	0,957	96,339
	% de Melhoria		**1,77%**	**0,42%**	**0,92%**
Glass	Original	J48	0,283	0,964	91,118
	Oversampling	J48	0,417	0,964	92,049
	% de Melhoria		**47,35%**	**0,00%**	**1,02%**
Heart	Original	J48	0,717	0,780	75,185
	Undersampling	J48	0,767	0,787	77,778
	% de Melhoria		**6,97%**	**0,90%**	**3,45%**
Ionosphere	Original	J48	0,818	0,911	87,738
	Borderline	J48	0,849	0,956	91,726
	% de Melhoria		**3,79%**	**4,94%**	**4,55%**
Lettera	Original	BayesNet	0,848	0,992	98,605
	Smote	BayesNet	0,937	0,994	99,19
	% de Melhoria		**10,50%**	**0,20%**	**0,59%**
NewThyroid	Original	Jrip	0,857	0,978	95,814
	Adasyn	Jrip	0,914	0,983	97,209
	% de Melhoria		**6,67%**	**0,57%**	**1,46%**
Pima	Original	NaiveBayes	0,848	0,608	76,433
	Adasyn	NaiveBayes	0,921	0,909	91,755
	% de Melhoria		**8,61%**	**49,48%**	**20,05%**
Vehicle	Original	BayesNet	0,945	0,683	74,467
	Undersampling	BayesNet	0,992	0,968	97,656
	% de Melhoria		**4,94%**	**41,78%**	**31,14%**

As bases Bupa, Ecoli, Flag, Haberman e Satimage não aparecem na Tabela II por não ter um método que melhore ambas as classes em nenhum dos classificadores. As bases Pima e Vehicle por exemplo, se destacam entre as bases, pois conseguem um significativo aumento na taxa de classificação, mas evidentemente com métodos e classificadores diferentes. O destaque da base Pima é o método Adasyn classificado pelo NaiveBayes, e da base Vehicle, o método Undersampling com o classificador BayesNet. Métodos esses que trabalham de forma totalmente diferente, o Adasyn por exemplo, tem a função de aumentar a classe minoritária acrescentando exemplos sintéticos próximos a superfície de decisão, e Undersampling, tem a função de diminuir os exemplos da classe majoritária.

Pode-se observar na Tabela II que não é possível dizer que um método é melhor que o outro, pois os métodos possuem comportamentos diferentes em cada base, é importante ressaltar também, que o classificador pode influenciar nos resultados. Comprovando este fato, pode ser visto nos resultados obtidos no trabalho de [12], os resultados para a base Pima com o Algoritmo Adasyn aplicado e comparado com o algoritmo Smote e ambos classificados por um algoritmo de arvore de decisão. Os resultados do trabalho de [12] demonstram que o Adasyn obtém o melhor resultado em relação a taxa de acurácia 68% e 60% de Recall, e pode ser observado nos resultados obtidos nos experimentos deste trabalho que o classificador que se destaca com resultados melhores é o NaiveBayes, tendo 92% de RecallP e 91,77% de acurácia.

A Tabela IV apresenta os resultados das bases originais utilizando os métodos híbridos. Esses Métodos são chamados híbridos por terem uma combinação entre métodos de pré-processamento e classificadores.

Observando os resultados com os métodos híbridos (Tabela IV), é possível notar resultados razoáveis para a classe minoritária, sem diminuir drasticamente os valores da classe majoritária, se comparado com os resultados obtidos na Tabela III. Os métodos que se destacam são: *BalanceCascade, EasyEnsemble e SmoteBoost*, é importante observar também, que esses métodos conseguem valores razoáveis da classe positiva, principalmente nas bases que o nível de desbalanceamento é grande, como por exemplo, Lettera, Glass, Flag, Satimage, Ecoli, New-thyroid e Vehicle.

A Tabela VI apresenta os resultados com os métodos de balanceamento classificados com os métodos híbridos. É notório que os resultados para a classe positiva (RecallP) são melhores em todas as bases, no entanto, os resultados são piores para a classe negativa. A análise da Tabela VI é comparada com os métodos híbridos aplicados ā base original (Tabela IV).

Breast - O método Smote melhora a classe positiva mantendo a classe negativa se classificado pelo método hibrido EasyEnsemble. O mesmo acontece com os métodos Adasyn, Borderline e Oversampling classificados pelo SmoteBoost. **Bupa -** Pode-se observar a melhora significativa da classe negativa com o método BalanceCascade com todos os métodos de pré-processamento, mas é notório também os péssimos valores da classe de interesse (positiva). Nesta base um método que pode ser destacado é SmoteBoost com o método Smote, pois quase mantem os valores das classes. **Ecoli -** O método híbrido SmoteBagging consegue melhorar a classe positiva e manter a classe negativa com o método Oversampling. O destaque vai para o método Borderline, que consegue melhorar ambas as classes com este mesmo classificador. **Flag -** Nesta base os resultados são satisfatórios, pois o método Smote consegue aumentar os valores de ambas as classes, com os classificadores SmoteBagging e SmoteBoost, Oversampling com o SmoteBoost e o método Undersampling com o EasyEnsemble. **Glass, Haberman e Satimage -** Os métodos, tanto para pré-processamento quanto os híbridos, não conseguem melhorar a classe positiva sem diminuir drasticamente os valores da classe negativa. **Heart -** O método híbrido SmoteBoost, utilizando o método de pré-processamento Adasyn e Oversampling, consegue melhorar ambas as classes nesta base. . **Ionosphere -** SmoteBoost com o método Smote e Undersampling melhora a classe negativa mantendo o valor da classe positiva, consequentemente melhorando a acurácia. **Lettera -** O único método de pré-processamento que consegue predizer a classe negativa é o Undersampling. O restante dos métodos não conseguem predizer nenhum exemplo da classe positiva, e acerta 100% da classe positiva. **New-Thyroid -** Os métodos Borderline e Smote conseguem melhoras nas classes positiva e negativa, classificados pelos métodos híbridos EasyEnsemble. O método Oversampling mantem uma das classes com o EasyEnsemble, o mesmo acontece com Borderline com SmoteBagging e Adasyn com SmoteBoost. **Pima -** O Método híbrido SmoteBoost aumenta razoavelmente os valores das classes com o método Undersampling, enquanto que o Borderline com o EasyEnsemble aumenta apenas a

classe positiva. **Vehicle -** Nesta base, o método híbrido SmoteBoost consegue aumentar os valores de uma das classes mantendo os valores da outra, melhorando assim a acurácia, os métodos em destaque são Adasyn, Borderline e Oversampling.

Os resultados afirmam que os métodos de balanceamento melhoram a capacidade de predição dependendo é claro de como o classificador generaliza a base, e principalmente o nível de desbalanceamento. A piora em algumas bases com alguns métodos de balanceamento pode ter ocorrido por terem movido a fronteira de decisão, dificultando ainda mais o poder de generalização dos métodos híbridos.

VI. Conclusão

Neste trabalho foram apresentados métodos propostos na literatura para tratar o problema de classificação em bases desbalanceadas. Estes métodos foram aplicados em treze (13) bases de dados provenientes do UCI *Machine Learning Repository*.

Primeiramente, foram efetuados os experimentos aplicando 5 classificadores chamados de classificadores Tradicionais. Com o intuito de observar o comportamento dos métodos de balanceamento foram aplicados na base original os métodos de balanceamento: Adasyn, Borderline, Oversampling, Smote e Undersampling e novamente classificados pelos algoritmos tradicionais. Com o mesmo objetivo os métodos híbridos *BalanceCascade, EasyEnsemble, SMOTEBagging e SMOTEBoost* foram aplicados na base original e depois aplicados com a base já balanceada com o métodos de pré-processamento.

Para as 13 bases, cada método apresenta resultados diferentes para cada classificador. De acordo com os resultados obtidos nos experimentos deste trabalho é possível responder as questões realizadas ao assunto, que são: (i) A Classificação pode ser melhorada aplicando algum método de balanceamento? A melhora da classificação é possível, mas para tal, é preciso da combinação correta do método aplicado com o classificador. Observar utilizando uma métrica de avaliação que realmente indique a melhoria nos valores da classe de interesse (classe minoritária). Para isso, vários testes deveriam ser realizados para descobrir o método que resulta melhores valores para uma base específica. (ii) Existe um método que sempre se comporta bem? Com os resultados obtidos é possível dizer que ainda não existe um método que resolva todos os problemas da classificação desbalanceada, pois como pode ser visto, os métodos possuem comportamentos diferentes, dependendo da base e do classificador. Geralmente, as bases utilizados para testar a eficácia dos métodos são bases provenientes do UCI *machine learning repository*, bases que são pré-processadas e que não possuem os problemas como ruídos, que normalmente são mais frequentes em bases do mundo real. (iii) Os classificadores podem influenciar nos resultados? Como dito anteriormente, cada classificador é capaz de construir um modelo a partir do conjunto de dados de treinamento, por este motivo, o classificador pode sim influenciar nos resultados, como mostra os resultados das bases provenientes do UCI *machine learning repository*. Ou seja, a mesma base com o mesmo método de balanceamento aplicado, utilizando classificadores diferentes possuem resultados diferentes. Pode-se concluir, que é de suma importância a utilização de vários classificadores para observar melhor o comportamento dos métodos aplicados. (iv) O que significa melhorar a taxa de classificação da classe minoritária e no mínimo manter a taxa de classificação da classe majoritária? Quando acontece a melhora nos valores da classe minoritária e os valores da classe majoritária são no mínimo mantidos, significa que o método realmente consegue evitar os problemas indesejados como por exemplo os ruídos, facilitando assim o poder de generalização do classificador, e consequentemente melhorando a acurácia. (v) Quais são as métricas de avaliação que realmente apontam a melhora da classe minoritária? Um dos fatores importantes observado neste trabalho é o tipo de métrica de avaliação utilizada como critério de avaliação, pois a maioria dos trabalhos relacionados citados neste trabalho utilizam a acurácia e o F-Measure como o critério de avaliação, o que não significa que com os resultados dessas métricas tenha melhorado a classe minoritária, pois na maioria das vezes a melhora acontece apenas na classe majoritária, mascarando o verdadeiro objetivo dos métodos de balanceamento. Desta forma, é muito importante observar os resultados obtidos entre as classes separadamente. Uma maneira interessante é utilizar a métrica de avaliação Recall, fazendo os cálculos dos verdadeiros positivos e dos verdadeiros negativos.

APÊNDICE

Tabela III
CLASSIFICADORES TRADICIONAIS - BASE ORIGINAL

Base	Class. Tradicionais	RecallP	RecallN	Acurácia
breast	NaiveBayes	0,975	0,957	96,340
	SMO	0,962	0,973	96,924
	MultilayerP	0,941	0,969	95,900
	J48	0,937	0,953	94,729
	JRip	0,958	0,953	95,460
	BayesNet	0,979	0,971	97,364
bupa	NaiveBayes	0,772	0,395	55,362
	SMO	0,007	0,995	57,971
	MultilayerP	0,634	0,770	71,304
	J48	0,634	0,690	66,667
	JRip	0,490	0,785	66,087
	BayesNet	0,172	0,870	57,681
Ecoli	NaiveBayes	0,800	0,894	88,402
	SMO	0,029	1,000	89,877
	MultilayerP	0,714	0,963	93,766
	J48	0,486	0,973	92,270
	JRip	0,571	0,960	91,971
	BayesNet	0,886	0,860	86,308
Flag	NaiveBayes	0,500	0,808	77,908
	SMO	0,067	0,960	88,138
	MultilayerP	0,117	0,949	87,625
	J48	0,000	1,000	91,242
	JRip	0,100	0,960	88,650
	BayesNet	0,300	0,904	85,088
Glass	NaiveBayes	0,783	0,457	48,084
	SMO	0,000	1,000	92,060
	MultilayerP	0,183	0,970	90,642
	J48	0,283	0,964	91,118
	JRip	0,067	0,980	90,664
	BayesNet	0,000	1,000	92,060
haberman	NaiveBayes	0,199	0,942	74,521
	SMO	0,000	1,000	73,533
	MultilayerP	0,286	0,902	73,882
	J48	0,470	0,809	71,893
	JRip	0,345	0,867	72,871
	BayesNet	0,257	0,893	72,544
heart	NaiveBayes	0,792	0,853	82,593
	SMO	0,775	0,907	84,815
	MultilayerP	0,733	0,820	78,148
	J48	0,717	0,780	75,185
	JRip	0,708	0,833	77,778
	BayesNet	0,808	0,860	83,704
ionosphere	NaiveBayes	0,865	0,809	82,893
	SMO	0,714	0,960	87,167
	MultilayerP	0,810	0,964	90,881
	J48	0,818	0,911	87,738
	JRip	0,826	0,929	89,175
	BayesNet	0,810	0,929	88,600
Lettera	NaiveBayes	0,858	0,992	98,650
	SMO	0,845	0,998	99,160
	MultilayerP	0,888	1,000	99,540
	J48	0,958	0,999	99,715
	JRip	0,949	0,998	99,635
	BayesNet	0,848	0,992	98,605
New-thyroid	NaiveBayes	1,000	0,972	97,674
	SMO	0,543	1,000	92,558
	MultilayerP	0,943	0,989	98,140
	J48	0,829	0,989	96,279
	JRip	0,857	0,978	95,814
	BayesNet	0,886	0,994	97,674
pima	NaiveBayes	0,848	0,608	76,433
	SMO	0,896	0,534	76,956
	MultilayerP	0,838	0,563	74,219
	J48	0,816	0,567	72,916
	JRip	0,836	0,605	75,525
	BayesNet	0,828	0,630	75,910
Satimage	NaiveBayes	0,866	0,815	82,020
	SMO	0,000	1,000	90,272
	MultilayerP	0,660	0,970	94,017
	J48	0,545	0,957	91,671
	JRip	0,514	0,968	92,432
	BayesNet	0,835	0,863	86,030
Vehicle	NaiveBayes	0,889	0,592	66,192
	SMO	0,909	0,975	95,981
	MultilayerP	0,939	0,983	97,279
	J48	0,864	0,957	93,503
	JRip	0,859	0,944	92,436
	BayesNet	0,945	0,683	74,467

Tabela IV
MÉTODOS HÍBRIDOS - BASE ORIGINAL

Base	Híbridos	RecallP	RecallN	Acurácia
breast	BalanceCascade	0,980	0,940	96,00
	EasyEnsemble	0,960	0,960	96,00
	SMOTEBagging	0,940	0,960	95,00
	SMOTEBoost	0,940	0,960	95,00
bupa	BalanceCascade	0,910	0,240	52,00
	EasyEnsemble	0,710	0,640	67,00
	SMOTEBagging	0,510	0,760	66,00
	SMOTEBoost	0,600	0,720	67,00
Ecoli	BalanceCascade	0,940	0,830	84,00
	EasyEnsemble	0,880	0,790	80,00
	SMOTEBagging	0,820	0,880	88,00
	SMOTEBoost	0,800	0,940	92,00
Flag	BalanceCascade	0,820	0,610	63,00
	EasyEnsemble	0,820	0,550	57,00
	SMOTEBagging	0,470	0,840	81,00
	SMOTEBoost	0,290	0,900	85,00
Glass	BalanceCascade	0,880	0,540	57,00
	EasyEnsemble	0,760	0,650	66,00
	SMOTEBagging	0,580	0,850	83,00
	SMOTEBoost	0,580	0,900	87,00
haberman	BalanceCascade	0,660	0,610	63,00
	EasyEnsemble	0,560	0,760	70,00
	SMOTEBagging	0,550	0,750	70,00
	SMOTEBoost	0,590	0,720	68,00
heart	BalanceCascade	0,920	0,360	61,00
	EasyEnsemble	0,830	0,740	78,00
	SMOTEBagging	0,760	0,860	81,00
	SMOTEBoost	0,750	0,765	76,00
ionosphere	BalanceCascade	0,890	0,869	86,00
	EasyEnsemble	0,860	0,890	89,00
	SMOTEBagging	0,840	0,894	90,00
	SMOTEBoost	0,840	0,903	91,00
Lettera	BalanceCascade	0,980	0,970	97,00
	EasyEnsemble	0,980	0,970	97,00
	SMOTEBagging	1,00	0,000	3,00
	SMOTEBoost	1,00	0,000	3,00
New-thyroid	BalanceCascade	0,970	0,930	93,00
	EasyEnsemble	0,850	0,960	94,00
	SMOTEBagging	0,910	0,950	94,00
	SMOTEBoost	0,970	0,960	96,00
pima	BalanceCascade	0,540	0,860	65,00
	EasyEnsemble	0,670	0,760	70,00
	SMOTEBagging	0,770	0,690	74,00
	SMOTEBoost	0,740	0,710	73,00
Satimage	BalanceCascade	0,870	0,830	83,00
	EasyEnsemble	0,860	0,820	83,00
	SMOTEBagging	0,720	0,940	92,00
	SMOTEBoost	0,710	0,950	93,00
Vehicle	BalanceCascade	0,980	0,890	91,00
	EasyEnsemble	0,960	0,910	92,00
	SMOTEBagging	0,950	0,930	94,00
	SMOTEBoost	0,950	0,950	95,00

Tabela V
MÉTODOS DE PRÉ-PROCESSAMENTO

Base	Métodos	NaiveBayes			SMO			MultilayerP			J48			JRip			BayesNet		
		RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia
breast	ADASYN	0,987	0,944	95,901	0,996	0,955	96,926	0,979	0,946	95,754	0,950	0,939	94,292	0,971	0,948	95,606	0,992	0,968	97,656
	Boderline	0,987	0,948	96,193	0,992	0,962	97,219	0,975	0,959	96,487	0,941	0,959	95,317	0,975	0,957	96,339	0,983	0,966	97,217
	Smote	0,983	0,955	96,487	0,983	0,971	97,510	0,967	0,957	96,048	0,950	0,948	94,873	0,966	0,953	95,754	0,979	0,968	97,217
	Oversampling	0,979	0,957	96,487	0,979	0,966	97,072	0,958	0,962	96,048	0,950	0,953	95,165	0,954	0,966	96,193	0,992	0,971	97,804
	Undersampling	0,979	0,946	95,756	0,975	0,971	97,217	0,992	0,960	97,073	0,966	0,948	95,460	0,967	0,944	95,172	0,983	0,971	97,511
Bupa	ADASYN	0,834	0,275	51,014	0,745	0,525	61,739	0,786	0,570	66,087	0,621	0,610	61,449	0,710	0,610	65,217	0,883	0,125	44,348
	Boderline	0,807	0,300	51,304	0,717	0,490	58,551	0,724	0,665	68,986	0,517	0,685	61,449	0,648	0,650	64,928	0,669	0,430	53,043
	Smote	0,828	0,295	51,884	0,745	0,480	59,130	0,703	0,640	66,667	0,614	0,645	63,188	0,634	0,610	62,029	0,717	0,330	49,275
	Oversampling	0,821	0,260	49,565	0,724	0,495	59,130	0,703	0,660	67,826	0,607	0,670	64,348	0,641	0,650	64,638	0,614	0,515	55,652
	Undersampling	0,834	0,230	48,406	0,786	0,430	57,971	0,703	0,565	62,319	0,545	0,680	62,319	0,669	0,605	63,188	0,883	0,125	44,348
Ecoli	ADASYN	0,943	0,840	85,114	0,857	0,807	81,247	0,800	0,897	88,696	0,800	0,874	86,611	0,829	0,870	86,611	0,914	0,844	85,119
	Boderline	0,886	0,867	86,901	0,886	0,834	83,924	0,743	0,937	91,677	0,714	0,927	90,478	0,600	0,934	89,890	0,914	0,857	86,308
	Smote	0,943	0,844	85,404	0,886	0,824	83,029	0,886	0,900	89,881	0,743	0,914	89,587	0,743	0,910	89,289	0,914	0,851	85,720
	Oversampling	0,914	0,844	85,105	0,857	0,830	83,332	0,800	0,937	92,270	0,657	0,937	90,777	0,629	0,907	87,801	0,800	0,864	85,711
	Undersampling	0,971	0,840	85,404	0,971	0,787	80,632	0,800	0,837	83,332	0,800	0,784	78,595	0,914	0,837	84,522	0,914	0,807	81,831
Flag	ADASYN	0,617	0,514	52,105	0,400	0,740	71,134	0,283	0,791	74,750	0,217	0,769	72,186	0,250	0,893	83,522	0,367	0,791	75,263
	Boderline	0,433	0,825	78,934	0,267	0,893	83,522	0,183	0,927	86,100	0,200	0,905	84,049	0,067	0,943	86,599	0,067	0,994	91,242
	Smote	0,550	0,429	43,900	0,417	0,752	72,173	0,533	0,769	74,764	0,617	0,875	85,020	0,317	0,870	81,957	0,367	0,849	80,459
	Oversampling	0,550	0,746	72,753	0,050	0,893	81,943	0,133	0,870	80,391	0,283	0,898	84,521	0,167	0,909	84,521	0,350	0,893	84,575
	Undersampling	0,550	0,493	49,649	0,700	0,696	69,663	0,700	0,662	66,545	0,683	0,566	57,800	0,550	0,482	48,596	0,650	0,700	69,568
Glass	ADASYN	0,850	0,431	46,213	1,000	0,314	36,866	0,633	0,822	80,864	0,583	0,888	86,390	0,650	0,797	78,505	0,533	0,777	75,703
	Boderline	0,800	0,523	54,142	1,000	0,319	37,298	0,433	0,924	88,317	0,433	0,954	91,096	0,300	0,934	88,283	0,433	0,888	85,039
	Smote	0,783	0,431	45,759	0,867	0,380	42,027	0,533	0,878	85,028	0,650	0,832	81,761	0,550	0,878	84,994	0,417	0,832	79,867
	Oversampling	0,733	0,461	48,084	0,883	0,390	42,924	0,483	0,939	90,188	0,417	0,964	92,049	0,233	0,939	88,328	0,233	0,934	87,829
	Undersampling	0,833	0,406	43,865	1,000	0,263	32,193	0,883	0,461	49,446	0,800	0,542	56,489	0,800	0,415	44,762	0,933	0,411	45,293
Haberman	ADASYN	0,371	0,898	75,833	0,368	0,884	74,870	0,297	0,884	72,898	0,629	0,693	67,652	0,667	0,667	66,653	0,704	0,622	64,363
	Boderline	0,334	0,898	74,849	0,346	0,902	75,505	0,471	0,804	71,581	0,580	0,778	72,559	0,582	0,738	69,614	0,556	0,800	73,538
	Smote	0,358	0,893	75,177	0,346	0,902	75,510	0,349	0,858	72,242	0,555	0,733	68,646	0,642	0,693	67,980	0,506	0,764	69,603
	Oversampling	0,346	0,898	75,177	0,382	0,889	75,516	0,455	0,836	73,548	0,531	0,680	64,056	0,568	0,684	65,341	0,457	0,809	71,565
	Undersampling	0,396	0,858	73,543	0,454	0,747	67,023	0,692	0,627	64,379	0,485	0,782	70,270	0,592	0,747	70,608	0,663	0,587	60,894
Heart	ADASYN	0,842	0,847	84,444	0,833	0,827	82,963	0,750	0,787	77,037	0,758	0,760	75,926	0,700	0,773	74,074	0,825	0,813	81,852
	Boderline	0,808	0,867	84,074	0,842	0,860	85,185	0,733	0,760	74,815	0,767	0,787	77,778	0,717	0,813	77,037	0,808	0,840	82,593
	Smote	0,808	0,840	82,593	0,783	0,853	82,222	0,800	0,793	79,630	0,783	0,780	78,148	0,642	0,813	73,704	0,825	0,827	82,593
	Oversampling	0,792	0,853	82,593	0,817	0,847	83,333	0,650	0,787	72,593	0,667	0,753	71,481	0,742	0,780	76,296	0,817	0,827	82,222
	Undersampling	0,817	0,853	83,704	0,842	0,853	84,815	0,767	0,793	78,148	0,767	0,787	77,778	0,742	0,787	76,667	0,825	0,840	83,333
Ionosphere	ADASYN	0,873	0,844	85,453	0,738	0,920	85,473	0,905	0,907	90,588	0,888	0,880	88,306	0,841	0,822	82,893	0,905	0,871	88,314
	Boderline	0,865	0,849	85,445	0,746	0,920	85,755	0,794	0,947	89,163	0,849	0,956	91,726	0,833	0,920	88,889	0,858	0,893	88,028
	Smote	0,865	0,858	86,016	0,738	0,929	86,040	0,818	0,933	89,163	0,818	0,929	88,889	0,850	0,849	84,926	0,873	0,907	89,449
	Oversampling	0,865	0,787	81,469	0,714	0,933	85,461	0,834	0,933	89,734	0,873	0,920	90,314	0,779	0,911	86,342	0,802	0,942	89,171
	Undersampling	0,865	0,787	81,473	0,738	0,929	86,036	0,786	0,929	87,738	0,841	0,951	91,163	0,873	0,880	87,751	0,834	0,933	89,738
Lettera	ADASYN	0,350	0,794	77,614	0,989	0,860	86,534	0,944	0,991	98,880	0,953	0,999	99,680	0,913	0,997	99,340	0,757	0,998	98,875
	Boderline	0,247	0,964	93,565	0,437	0,824	80,899	0,973	0,904	90,634	0,971	0,996	99,490	0,956	0,997	99,535	0,900	0,998	99,370
	Smote	0,173	0,995	96,270	0,896	0,976	97,280	0,951	0,971	97,000	0,978	0,999	99,805	0,957	0,998	99,595	0,937	0,994	99,190
	Oversampling	0,782	0,993	98,490	0,896	0,974	97,050	0,949	0,972	97,105	0,987	0,975	97,505	0,972	0,970	96,990	0,947	0,968	96,745
	Undersampling	0,868	0,983	97,855	0,890	0,973	97,010	0,949	0,969	96,820	1,000	0,928	93,953	1,000	0,922	93,488	0,971	0,911	92,093
NewThyroid	ADASYN	0,902	0,923	92,969	1,000	0,822	85,116	1,000	0,950	95,814	0,943	0,989	98,140	0,914	0,983	97,209	0,943	0,978	97,209
	Boderline	1,000	0,967	97,209	1,000	0,933	94,419	1,000	0,978	98,140	1,000	0,967	97,209	0,914	0,950	94,419	0,914	0,961	95,349
	Smote	0,971	0,950	95,349	1,000	0,928	93,953	1,000	0,967	97,209	0,943	0,983	97,674	0,886	0,978	96,279	0,886	0,983	96,744
	Oversampling	1,000	0,983	98,605	1,000	0,939	94,884	1,000	0,989	99,070	0,971	0,989	98,605	0,886	0,961	94,884	0,886	0,944	93,488
	Undersampling	0,943	0,989	98,140	1,000	0,939	94,884	0,914	0,994	98,140	0,688	0,742	70,705	0,660	0,758	69,399	0,696	0,769	72,140
Pima	ADASYN	0,921	0,909	91,755	0,750	0,720	73,959	0,738	0,765	74,747	0,724	0,732	72,659	0,740	0,743	74,086	0,768	0,706	74,613
	Boderline	0,706	0,750	72,142	0,774	0,679	74,090	0,752	0,731	74,483	0,690	0,750	71,107	0,736	0,728	73,318	0,750	0,672	72,272
	Smote	0,730	0,732	73,054	0,772	0,683	74,093	0,744	0,720	73,574	0,752	0,735	74,607	0,748	0,649	71,358	0,730	0,739	73,299
	Oversampling	0,722	0,724	72,275	0,774	0,675	73,960	0,752	0,705	73,574	0,702	0,739	71,481	0,714	0,739	72,266	0,690	0,732	70,445
	Undersampling	0,724	0,694	71,361	0,774	0,694	74,613	0,772	0,713	75,134	0,797	0,884	87,584	0,673	0,919	89,495	0,574	0,931	89,650
satimage	ADASYN	0,795	0,712	74,173	0,925	0,678	70,194	0,940	0,573	60,839	0,701	0,947	92,354	0,620	0,938	90,707	0,552	0,952	91,329
	Boderline	0,186	0,957	88,174	0,879	0,744	75,695	0,885	0,614	63,994	0,818	0,900	89,215	0,695	0,918	89,666	0,689	0,921	89,868
	Smote	0,395	0,951	89,728	0,872	0,756	76,721	0,928	0,588	62,067	0,706	0,945	92,199	0,594	0,945	91,096	0,707	0,915	89,510
	Oversampling	0,557	0,938	90,070	0,891	0,743	75,742	0,911	0,594	62,471	0,895	0,842	84,755	0,827	0,822	82,284	0,871	0,826	83,015
	Undersampling	0,850	0,838	83,916	0,891	0,741	75,524	0,947	0,570	60,699	0,949	0,943	94,448	0,909	0,920	91,728	0,939	0,870	88,669
Vehicle	ADASYN	0,848	0,790	80,107	0,834	0,660	70,065	0,995	0,913	93,258	0,924	0,974	96,222	0,889	0,944	93,144	0,945	0,909	91,738
	Boderline	0,915	0,677	73,293	0,889	0,653	70,892	0,990	0,940	95,151	0,919	0,937	93,265	0,895	0,930	92,198	0,900	0,892	89,378
	Smote	0,924	0,708	75,896	0,940	0,643	71,248	0,995	0,921	93,850	0,935	0,972	96,339	0,884	0,961	94,327	0,919	0,900	90,437
	Oversampling	0,944	0,705	76,131	0,955	0,644	71,719	0,990	0,943	95,386	0,965	0,943	94,803	0,935	0,923	92,555	0,945	0,843	86,663
	Undersampling	0,940	0,699	75,542	0,945	0,644	71,483	0,995	0,910	93,021	0,950	0,939	94,292	0,971	0,948	95,606	0,992	0,968	97,656

Tabela VI
PRÉ-PROCESSAMENTO COM OS MÉTODOS HÍBRIDOS

Base	Métodos	BalanceCascade			EasyEnsemble			SmoteBagging			SmoteBoost		
		RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia	RecallP	RecallN	Acurácia
breast	ADASYN	0,920	0,950	94,000	0,960	0,950	96,000	0,990	0,920	94,000	0,950	0,960	96,000
	Boderline	0,770	0,970	90,000	0,950	0,960	96,000	0,970	0,940	95,000	0,960	0,960	96,000
	Smote	0,880	0,970	94,000	0,970	0,960	96,000	0,990	0,940	96,000	0,940	0,960	95,000
	Oversampling	0,910	0,970	95,000	0,950	0,960	96,000	0,980	0,940	95,000	0,950	0,960	96,000
	Undersampling	0,910	0,970	95,000	0,950	0,940	95,000	0,980	0,940	95,000	0,950	0,950	95,000
Bupa	ADASYN	0,280	0,840	61,000	0,520	0,740	64,000	0,810	0,470	61,000	0,600	0,620	61,000
	Boderline	0,240	0,920	64,000	0,570	0,770	68,000	0,760	0,540	63,000	0,640	0,650	64,000
	Smote	0,070	0,950	58,000	0,500	0,810	68,000	0,800	0,560	66,000	0,600	0,710	66,000
	Oversampling	0,270	0,930	65,000	0,460	0,750	63,000	0,780	0,570	66,000	0,560	0,710	65,000
	Undersampling	0,170	0,950	62,000	0,560	0,730	66,000	0,750	0,600	66,000	0,680	0,670	67,000
Ecoli	ADASYN	0,570	0,930	89,000	0,680	0,900	87,000	0,910	0,850	85,000	0,710	0,900	88,000
	Boderline	0,510	0,960	91,000	0,680	0,940	91,000	0,850	0,890	88,000	0,680	0,940	91,000
	Smote	0,480	0,980	92,000	0,600	0,940	90,000	0,800	0,890	88,000	0,570	0,960	92,000
	Oversampling	0,510	0,970	92,000	0,710	0,940	91,000	0,880	0,880	88,000	0,710	0,940	91,000
	Undersampling	0,480	0,950	90,000	0,800	0,910	90,000	0,850	0,860	86,000	0,740	0,850	83,000
Flag	ADASYN	0,110	0,940	87,000	0,110	0,870	80,000	0,410	0,720	69,000	0,170	0,920	85,000
	Boderline	0,110	0,960	88,000	0,050	0,930	85,000	0,170	0,900	84,000	0,170	0,950	88,000
	Smote	0,290	0,930	88,000	0,290	0,870	82,000	0,580	0,880	85,000	0,410	0,960	91,000
	Oversampling	0,110	0,970	89,000	0,640	0,880	86,000	0,640	0,750	74,000	0,290	0,940	89,000
	Undersampling	0,230	0,720	68,000	0,880	0,640	67,000	0,700	0,670	67,000	0,880	0,650	67,000
Glass	ADASYN	0,230	0,970	92,000	0,470	0,930	89,000	0,760	0,790	78,000	0,520	0,920	89,000
	Boderline	0,290	0,960	91,000	0,350	0,950	91,000	0,410	0,910	87,000	0,410	0,960	92,000
	Smote	0,290	0,990	93,000	0,410	0,950	91,000	0,470	0,900	87,000	0,350	0,950	91,000
	Oversampling	0,350	0,950	90,000	0,410	0,910	87,000	0,820	0,780	78,000	0,410	0,910	87,000
	Undersampling	0,230	0,900	85,000	0,640	0,630	63,000	0,520	0,590	59,000	0,640	0,620	62,000
Haberman	ADASYN	0,140	0,910	71,000	0,580	0,690	66,000	0,640	0,700	68,000	0,600	0,710	68,000
	Boderline	0,280	0,870	71,000	0,530	0,740	68,000	0,610	0,730	70,000	0,500	0,780	70,000
	Smote	0,160	0,900	70,000	0,500	0,730	67,000	0,640	0,660	66,000	0,510	0,690	65,000
	Oversampling	0,250	0,860	70,000	0,550	0,740	69,000	0,580	0,720	68,000	0,590	0,680	66,000
	Undersampling	0,120	0,920	70,000	0,640	0,680	67,000	0,610	0,730	70,000	0,580	0,700	67,000
Heart	ADASYN	0,310	0,960	67,000	0,740	0,810	78,000	0,810	0,740	77,000	0,770	0,800	78,000
	Boderline	0,150	0,980	61,000	0,680	0,850	77,000	0,740	0,750	74,000	0,760	0,770	77,000
	Smote	0,200	0,950	61,000	0,680	0,840	77,000	0,800	0,770	78,000	0,730	0,790	76,000
	Oversampling	0,190	0,980	62,000	0,700	0,820	77,000	0,800	0,760	77,000	0,790	0,800	80,000
	Undersampling	0,100	0,960	57,000	0,680	0,790	74,000	0,800	0,760	77,000	0,800	0,740	77,000
Ionosphere	ADASYN	0,700	0,960	86,000	0,840	0,950	91,000	0,920	0,830	86,000	0,920	0,920	92,000
	Boderline	0,690	0,980	88,000	0,820	0,960	91,000	0,890	0,890	89,000	0,870	0,940	92,000
	Smote	0,690	0,980	88,000	0,800	0,960	90,000	0,900	0,880	89,000	0,840	0,960	92,000
	Oversampling	0,730	0,980	89,000	0,800	0,920	88,000	0,890	0,880	89,000	0,860	0,940	91,000
	Undersampling	0,750	0,930	86,000	0,810	0,900	87,000	0,880	0,890	89,000	0,840	0,960	92,000
Lettera	ADASYN	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00
	Boderline	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00
	Smote	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00
	Oversampling	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00	1,00	0,000	3,00
	Undersampling	0,860	0,990	98,000	0,970	0,970	97,000	0,990	0,920	92,000	0,980	0,980	98,000
NewThyroid	ADASYN	0,820	0,980	96,000	0,940	0,950	94,000	0,970	0,930	94,000	1,00	0,960	96,000
	Boderline	0,880	1,00	98,000	0,880	0,980	96,000	0,910	0,980	97,000	0,880	0,970	96,000
	Smote	0,850	0,980	96,000	0,880	0,980	96,000	0,910	0,970	96,000	0,850	0,980	96,000
	Oversampling	0,850	0,970	95,000	0,850	0,960	94,000	0,940	0,930	93,000	0,910	0,940	93,000
	Undersampling	0,940	0,920	93,000	0,940	0,890	90,000	0,910	0,910	91,000	0,910	0,930	93,000
Pima	ADASYN	0,240	0,950	49,000	0,640	0,760	68,000	0,770	0,620	72,000	0,660	0,790	70,000
	Boderline	0,390	0,920	57,000	0,710	0,760	73,000	0,830	0,570	74,000	0,730	0,720	72,000
	Smote	0,490	0,880	62,000	0,730	0,710	72,000	0,800	0,620	74,000	0,800	0,640	74,000
	Oversampling	0,360	0,940	56,000	0,680	0,730	70,000	0,810	0,600	74,000	0,690	0,750	71,000
	Undersampling	0,250	0,960	50,000	0,650	0,760	69,000	0,780	0,650	73,000	0,750	0,720	74,000
satimage	ADASYN	0,460	0,970	92,000	0,650	0,940	91,000	0,980	0,630	67,000	0,690	0,950	93,000
	Boderline	0,540	0,970	92,000	0,610	0,950	92,000	0,930	0,730	75,000	0,690	0,960	94,000
	Smote	0,490	0,970	92,000	0,600	0,960	92,000	0,960	0,680	71,000	0,610	0,970	94,000
	Oversampling	0,510	0,960	92,000	0,670	0,940	92,000	0,970	0,670	70,000	0,700	0,960	93,000
	Undersampling	0,470	0,950	91,000	0,800	0,870	86,000	0,930	0,730	75,000	0,900	0,840	84,000
Vehicle	ADASYN	0,840	0,960	93,000	0,920	0,940	94,000	0,980	0,840	88,000	0,960	0,950	95,000
	Boderline	0,550	0,980	88,000	0,900	0,960	94,000	0,970	0,890	91,000	0,950	0,970	97,000
	Smote	0,730	0,970	92,000	0,870	0,960	94,000	0,980	0,870	90,000	0,930	0,970	96,000
	Oversampling	0,820	0,960	93,000	0,910	0,940	93,000	0,990	0,820	86,000	0,950	0,960	96,000
	Undersampling	0,640	0,950	88,000	0,940	0,900	91,000	0,970	0,870	90,000	0,960	0,910	92,000

REFERÊNCIAS

[1] H. He, "Learning from imbalanced data," *and Data Engineering, IEEE Transactions*, vol. 21, 2009. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5128907

[2] M. Beckmann, "Algoritmos Genéticos como estratégia de Pré-Processamento em conjunto de dados Desbalanceados," Dissertação de Mestrado, UFRJ, 2010. [Online]. Available: http://objdig.ufrj.br/60/teses/coppe_m/MarceloBeckmann.pdf

[3] H. Han, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," *Advances in Intelligent Computing*, 2005. [Online]. Available: http://www.springerlink.com/index/5d3mfq0vd7wuh96g.pdf

[4] N. Chawla and A. Lazarevic, "SMOTEBoost: Improving prediction of the minority class in boosting," *Knowledge Discovery*, 2003. [Online]. Available: http://www.springerlink.com/index/mt1gbpar9akb02a2.pdf

[5] U. Fayyad and G. Piatetsky-Shapiro, "From data mining to knowledge discovery in databases," 1996. [Online]. Available: http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230

[6] G. E. A. P. A. Batista, "Pré-processamento de Dados em Aprendizado de Máquina Supervisionado," Tese de Doutorado, USP - São Carlos, 2003. [Online]. Available: http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/

[7] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *sci2s.ugr.es*, 2011. [Online]. Available: http://sci2s.ugr.es/publications/ficheros/2011-Galar-IEEE_TSMCc-Ensembles.pdf

[8] E. Machado, "Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes," Dissertação de Mestrado, Universidade de Brasília, 2009. [Online]. Available: http://repositorio.bce.unb.br/handle/10482/1397

[9] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning." *IEEE transactions on systems, man, and cybernetics.*, vol. 39, no. 2, Apr. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19095540

[10] K. Bowyer, N. Chawla, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Arxiv preprint arXiv*, vol. 16, 2011.

[11] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[12] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *International Joint Conference on Neural Networks*, 2008.