UNIVERSIDADE FEDERAL DE ALFENAS
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Vinícius Ferreira da Silva

SELEÇÃO DE PADRÕES PARA *RESAMPLING EM PROBLEMAS DE CLASSIFICAÇÃO*

Alfenas, 06 de julho de 2015.

UNIVERSIDADE FEDERAL DE ALFENAS
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

SELEÇÃO DE PADRÕES PARA *RESAMPLING EM PROBLEMAS DE CLASSIFICAÇÃO*

Vinícius Ferreira da Silva

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Alfenas como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Humberto César Brandão de Oliveira

Alfenas, 06 de julho de 2015.

Vinícius Ferreira da Silva

SELEÇÃO DE PADRÕES PARA *RESAMPLING EM PROBLEMAS DE CLASSIFICAÇÃO*

A Banca examinadora abaixo-assinada aprova a monografia apresentada como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação pela Universidade Federal de Alfenas.

Prof. Dr. Eric Batista Ferreira

Universidade Federal de Alfenas

Prof. Dr. Ricardo Menezes Salgado

Universidade Federal de Alfenas

Prof. Dr. Humberto César Brandão de Oliveira (Orientador)

Universidade Federal de Alfenas

Alfenas, 06 de julho de 2015.

Dedico este trabalho aos meus pais Carlos e Izanete, ao meu irmão Vitor e à minha namorada Isabela.

AGRADECIMENTO

Agradeço primeiramente aos meus queridos pais, Carlos Roberto Ribeiro da Silva e Izanete Aparecida Ferreira da Silva, por me prestarem todas as formas de amor, apoio e cuidados que uma pessoa pode receber, além de serem exemplos de caráter e dedicação. Sem seus ensinamentos e seu suporte, eu não conseguiria concluir esta etapa.

Ao meu irmão Vitor Ferreira da Silva pelo carinho e admiração que sempre mostrou por seu irmão mais velho, fator que sempre me motivou em horas complicadas.

À minha doce Isabela Cristina Pires Machado pela companhia, carinho e por ser uma grande fonte de inspiração e sorrisos.

Ao meu amigo Adenir Teixeira de Souza por esses 11 anos de amizade e companheirismo.

Aos amigos feitos na graduação, em especial os cativantes Eugênio Ferreira Cabral e Talysson Oliveira Cassiano.

Ao meu orientador Humberto César Brandão de Oliveira, por sempre acolher e nunca duvidar de ninguém que cruze a porta de seu laboratório em busca de conhecimento e oportunidades.

A todos os meus colegas do Laboratório de Pesquisa e Desenvolvimento (LP&D), os que estão lá e os que já não estão mais, pelas discussões sérias ou não.

Aos professores do BCC e da Matemática pelas lições, em especial ao professor Fabrício Goecking Avelar, pela paciência e interesse.

A todos, muito obrigado!

"O assunto mais importante do mundo pode ser simplificado até ao ponto em que todos possam apreciá-lo e compreendê-lo. Isso é, ou deveria ser, a mais elevada forma de arte."

Charles Chaplin

RESUMO

A classificação de padrões é um ramo da Inteligência Artificial que procura classificar objetos em categorias através de algoritmos inteligentes. Na criação de classificadores de padrões via treinamento supervisionado, o problema do desbalanceamento de dados se dá por conjuntos de treinamento que contém mais exemplos de uma determinada classe do que de outra, formando classes majoritárias e minoritárias. Esta situação pode levar o classificador a ignorar estas classes menos representadas, enquanto classifica eficientemente classes mais representadas. Existem várias técnicas que ajudam a lidar com este problema, sendo o *resampling* um tipo delas. Dentro do *resampling,* as abordagens vão desde a replicação de padrões da classe minoritária, até a criação de padrões artificiais, sendo comum a escolha de padrões aleatórios como entrada para tal. Com base nestas informações, este trabalho apresenta uma abordagem para a escolha de padrões, diferente da abordagem aleatória, tendo como base um ambiente desbalanceado. Testes foram realizados e através dos resultados podemos concluir que o método proposto é uma alternativa e que pode auxiliar na aplicação de técnicas de *resampling*.

Palavras-Chave: classificação, conjuntos desbalanceados, métodos de *resampling*.

ABSTRACT

The pattern classification is a branch of artificial intelligence that tries to classify objects into categories through the use of intelligent algorithms. In the pattern classification via supervised training, the imbalanced dataset problem happens when training sets containing many more examples of a given class than the other classes, forming majority and minority classes. This may lead the classifier to ignore these underrepresented classes, insofar efficiently classifies classes better represented. There are several techniques that help deal with this problem, and the resampling is a type of them. Within the resampling, approaches range from the replicate patterns of minority class, to create artificial patterns with the common choice of random patterns as input to such. Based on this information, this paper presents an approach for choosing patterns, unlike the random approach, based on an unbalanced environment. Tests were carried out and through the results we can conclude that the proposed method is an alternative that can facilitate the application that uses resampling techniques.

Keywords: classification, imbalanced databases, resampling methods.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

CP	Classificação de Padrões
RNA	Redes Neurais Artificiais
MLP	Multilayer Perceptron
SMOTE	*Synthetic Minority Oversampling Technique*
RUM	*Random Undersampling Method*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
FP	Falso Positivo
FN	Falso Negativo
ROC	*Receiver Operating Characteristic*
AUC	*Area Under Curve*
kNN	*k-Nearest Neighbors*
kRNN	*k-Reverse Nearest Neighbors*
VM	Vetor Médio
VDP	Vetor Desvio Padrão

SUMÁRIO

1
Introdução

Este capítulo apresenta a motivação e os objetivos deste trabalho. Na Seção 1.1 seguem a justificativa e motivação. A problematização é apresenta no item 1.2. Os itens 1.3.1 e 1.3.2 apresentam os objetivos gerais e específicos respectivamente. Finalmente, no item 4, é apresentada a organização desta monografia.

1.1Justificativa e Motivação

O avanço tecnológico disponibiliza cada vez maiores quantidades de dados brutos que, se analisados de uma maneira adequada, podem gerar conhecimento. A existência destes dados cria um ambiente propício para o desenvolvimento de algoritmos e técnicas de aprendizado de máquina, uma vez que estes dados são a alimentação para algoritmos, tais como algoritmos para classificação de padrões. Entretanto, alguns conjuntos de dados possuem um problema de distribuição de classes conhecido como dados desbalanceados.

O desbalanceamento acontece quando o conjunto de dados, que será utilizado para o treinamento do algoritmo de classificação de padrões, possui mais exemplos de um tipo do que de outra. Segundo os autores de (He, 2009) "O problema fundamental está na capacidade deste conjunto desbalanceado de comprometer o desempenho da maioria dos algoritmos de aprendizado de máquina".

Na literatura são encontradas técnicas para lidar com este problema que atuam durante o processo de treinamento e também técnicas de pré-processamento do conjunto, conhecido como *resampling*. A abordagem de *resampling* pode ser subdivida nas abordagens *oversampling* e *undersampling,* atuando para diminuir o desbalanceamento do conjunto. A abordagem *oversampling* se dá por técnicas que vão desde a réplica de padrões existentes até a criação de padrões artificiais da classe minoritária. Já a abordagem *undersampling* descreve técnicas para a remoção de padrões da classe majoritária. Geralmente a escolha dos padrões que serão replicados, alterados e removidos é feita de maneira aleatória.

Nesta conjuntura, este trabalho apresenta uma abordagem para a escolha de padrões para técnicas de *oversampling*, diferente da aleatória, que é utilizada em todos os trabalhos estudados. A abordagem apresentou resultados melhores do que as técnicas tradicionais nos testes realizados, sendo uma alternativa para a escolha aleatória e se mostrou uma contribuição para lidar com o problema do desbalanceamento de dados.

1.2 Problematização

A utilização de técnicas de *resampling* em conjuntos desbalanceados, geralmente, vem acompanhada de um acerto prévio de uma série de parâmetros. Somente aplicar uma técnica deste tipo em qualquer situação não garante melhoria no desempenho dos classificadores que tentam aprender naquele conjunto, pois o mesmo deve ser analisado e a técnica configurada de maneira adequada para que gere um resultado interessante.

A avaliação dos resultados deve ser feita de maneira adequada. É preciso escolher uma medida de desempenho adequada para os classificadores, escolher algoritmos para classificação que possuem aplicação real e realizar testes de maneira a evitar resultados tendenciosos. No caso da abordagem de escolha dos padrões proposta, é preciso comparar de maneira adequada o desempenho da abordagem sugerida com a técnica que é normalmente utilizada.

Portanto a questão principal é: “a escolha de padrões específicos para aplicação de técnicas de *oversampling* no conjunto de treinamento pode ser melhor que a escolha aleatória?”.

1.3 Objetivos

1.3.1 Gerais

Este trabalho possui como objetivo geral a proposta de uma abordagem para a escolha de padrões, que serão utilizados como entrada para técnicas de *oversampling,* que apresente desempenho mais alto do que a escolha aleatória.

1.3.2 Específicos

Os objetivos específicos deste trabalho são:

- Encontrar formas eficientes para determinar se um padrão é ou não apto a ser entrada para a aplicação de técnicas de *oversampling*;
- Aplicar técnicas de *resampling* para aumentar o desempenho de classificadores em ambientes desbalanceados;
- Desenvolver um *framework* para experimentação e testes, que facilite o desenvolvimento de novos classificadores e técnicas de seleção de padrões;
- Verificar, através de métodos estatísticos, qual a abordagem proposta tem um resultado melhor do que a abordagem aleatória;
- Testar a proposta para mais de um ambiente desbalanceado com diferentes graus de desbalanceamento;
- Defender este Trabalho de Conclusão de Curso (TCC), almejando a conclusão da graduação.

1.4 Organização da Monografia

No capítulo 2, são apresentados detalhes sobre a classificação de padrões e sobre o problema do desbalanceamento de dados. O capítulo contém introduções sobre os métodos utilizados e problemas tratados neste trabalho.

O capítulo 3 contém a revisão bibliográfica deste trabalho, com informações à respeito de métodos utilizados na literatura e formas de validação que foram utilizadas neste trabalho.

O capítulo 4 apresenta algumas definições e o método proposto por este trabalho para contribuir no enfrentamento do problema de desbalanceamento apresentado nas seções anteriores.

O capítulo 5 apresenta a metodologia utilizada para a realização deste trabalho. Informações sobre a implementação do ambiente de experimentação, os ciclos de teste realizados e informações sobre o ambiente desbalanceado estudado.

O capítulo 6 apresenta os resultados e testes de hipótese, para as duas bases utilizadas neste trabalho.

O capítulo 7 apresenta as conclusões, os trabalhos futuros que serão realizados e os trabalhos futuros sugeridos pelo autor para a continuidade do estudo.

2 Classificação de Padrões e o Desbalanceamento de Dados

Este capítulo tem como objetivo apresentar o ambiente deste trabalho. Apresenta na Seção 2.1 uma introdução à técnica de classificação de padrões e as técnicas e algoritmos utilizados neste trabalho, (Multilayer Perceptron e C4.5). Uma introdução ao problema de desbalanceamento de dados é apresentada na seção 2.2. A seção 2.3 contém uma breve introdução sobre validação de classificadores de padrões.

2.1 Classificação de padrões

A classificação de padrões é uma característica importante e está presente em toda a trajetória de evolução. Os seres que foram capazes de classificar um alimento como comestível sobreviveram, e hoje seus descendentes povoam nosso planeta.

Capacitar uma máquina para classificar padrões, ou seja, classificar objetos de interesse em uma categoria ou classe dentro de um número finito de categorias ou classes gerou uma série de aplicações, tais como reconhecimento de fala, análise de imagens, diagnósticos automatizados, entre outras. (Lippmann, 1989) cita diversas aplicações deste tipo de técnica. Para que a classificação aconteça, é necessário que existam medidas e características inerentes a cada classe que possibilite diferenciar os padrões. Tais características são chamadas de atributos.

Existem dois tipos de modelos para classificação, o supervisionado e o não supervisionado. No tipo supervisionado, que é o tipo estudado neste trabalho, se usa padrões já rotulados, ou seja, que já estão classificados, para criar um classificador (processo conhecido como treinamento) que seja capaz de classificar posteriormente objetos desconhecidos.

O processo de treinamento consiste em apresentar para o modelo um conjunto já rotulado de padrões. O modelo irá aprender sobre aquele conjunto e o tempo de aprendizagem varia de acordo com o método adotado. Quando o processo de treinamento

termina, o modelo é testado com um novo conjunto, denominado conjunto de teste. Caso o modelo não apresente um desempenho satisfatório em classificar os padrões do conjunto de teste, o treinamento deve ser refeito. A seguir são apresentados dois tipos de algoritmos que podem ser usados para classificação de padrões, e foram adotados neste trabalho.

2.1.1 Redes Neurais Artificiais

Através da união de estudos neurobiológicos com lógica matemática, (McCulloch & Pitts, 1943) propuseram a criação de neurônios formais. Neste estudo, os autores interpretaram o funcionamento de um neurônio biológico como sendo um circuito binário que combina entradas e gera uma saída, através da soma ponderada dos valores de entrada. Baseando-se neste trabalho, (Rosenblatt, 1957) propôs o modelo *Perceptron,* que se dividia em três camadas que recebiam, processavam os estímulos e apresentavam a resposta.

As Redes Neurais Artificiais (RNAs) fazem parte do ramo da Inteligência Artificial, mais especificamente da parte da Inteligência Artificial Conexionista. "O interesse nas redes neurais artificiais está nas propriedades mais abstratas, como sua habilidade para executar computação distribuída, tolerar entradas ruidosas e aprender" (Norvig, 2004).

As RNAs são compostas de unidades, denominadas neurônios, conectadas por vínculos, que propagam a ativação. Cada vínculo possui um valor numérico denominado peso, que é ajustado durante o processo de treinamento. Portanto, cada vez que o conjunto de treinamento é apresentado para uma RNA, esta usa os padrões pertencentes ao conjunto para ajustar os pesos dos vínculos entre os neurônios.

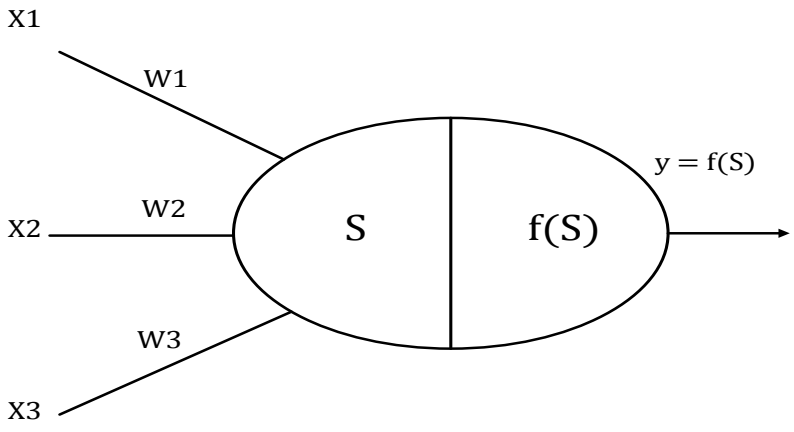

Figura 1 - Visão esquemática de um neurônio artificial.

Para o modelo simples de neurônio apresentado na Figura 1, considere o vetor de entradas $X = [X_1, X_2, \dots, X_n]$ com n dimensões, sendo que para cada X corresponde um peso

W_i. O valor S corresponde à soma ponderada das entradas X_i com os pesos correspondentes W, $S = \sum_i W_i X_i$. Então, é aplicada uma função de ativação f na soma S para obter a saída y. Uma configuração de pesos bem ajustada, permite a generalização eficiente para padrões desconhecidos. As redes neurais artificiais utilizadas neste trabalho são do tipo *Multilayer Perceptron* (MLP). A seguir é apresentado o outro tipo de algoritmo usado neste trabalho.

2.1.2 Árvores de decisão

Uma árvore de decisão toma como entrada um padrão e retorna uma decisão sobre aquele padrão, com base nos atributos do mesmo, sendo que esses atributos podem ser dados discretos ou dados contínuos. "Uma árvore desse tipo alcança sua decisão executando uma sequência de testes." (Norvig & Russell, 2004).

Cada nó interno na árvore corresponde a um teste realizado em um dos atributos do padrão sendo classificado, e as ramificações a partir do nó são os valores possíveis para o teste. As folhas da árvore de decisão correspondem às possíveis classes que o padrão pode pertencer.

É possível criar um caminho na árvore de decisões para cada padrão, sendo que da próxima vez que aquele padrão for apresentado, basta percorrer o caminho construído que a solução será encontrada, porém, a árvore não extrairá qualquer aprendizado e só será capaz de classificar padrões que já foram apresentados outrora, não sendo capaz de generalizar. Os algoritmos mais utilizados para construção de árvores de decisão para classificação escolhem os atributos que fazem mais diferença na classificação, tentando diminuir o número de testes necessários para se chegar à folha. Para este trabalho, o algoritmo de árvore de decisão utilizado foi o C4.5, proposto em (Quinlan, 1993), que constrói a árvore de decisão selecionando os atributos mais significativos através da entropia dos dados. A Figura 2 apresenta um exemplo de representação de uma árvore C4.5, construída para classificar dias em aptos ou não para a ocorrência de jogos de tênis.

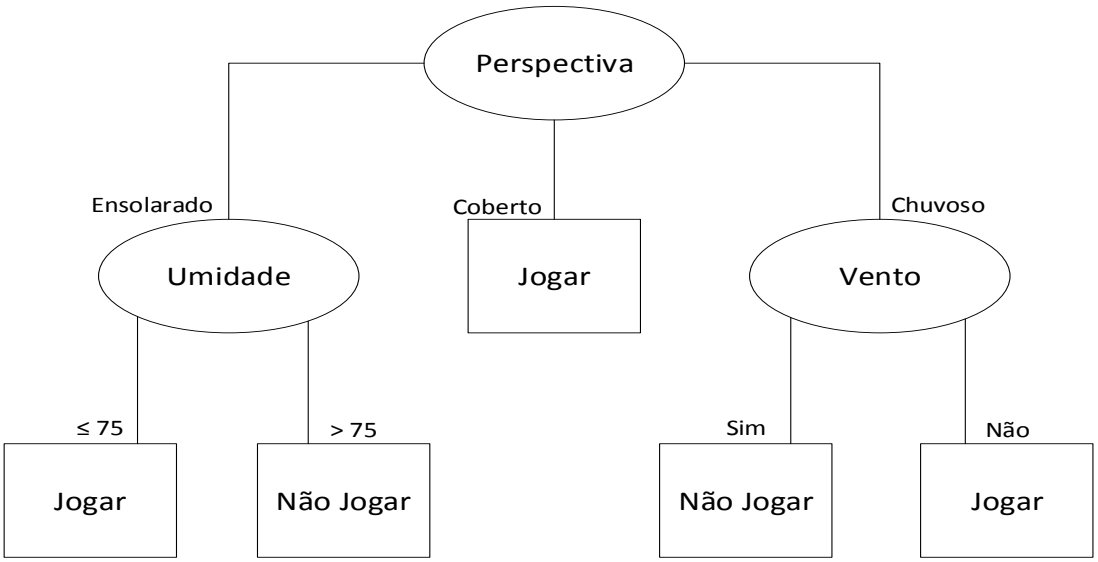

Figura 2 - Exemplo de uma árvore de decisão C4.5.

Neste exemplo, ao receber os atributos de um dia (perspectiva, umidade, vento), é possível classifica-lo como apto ou não para ocorrer o jogo de tênis, percorrendo o caminho na árvore. Seja um dia com as seguintes características: Ensolarado, com umidade menor do que 50 e com vento. Percorrendo o caminho na árvore para este exemplo, classificamos o mesmo como apto para jogo.

2.2 Desbalanceamento de dados

Qualquer conjunto de dados que apresente uma distribuição desigual de dados entre as classes pode ser considerado como desbalanceado. Entretanto, segundo os autores de (He, 2009) "O entendimento comum de desbalanceamento de dados na comunidade científica corresponde aos conjuntos que demonstram desequilíbrios significantes e em alguns casos extremos".

O desbalanceamento pode ser intrínseco ou extrínseco. O desbalanceamento intrínseco está relacionado à natureza dos dados, como no ambiente de fraudes de cartão de crédito, por exemplo, na qual existem muito mais clientes verídicos do que falsários. O desbalanceamento extrínseco não está relacionado a fatores naturais, e sim fatores variáveis, tais como tempo e armazenamento. Além disso, o desbalanceamento pode ser relativo ou consequente de instâncias raras. É relativo quando a quantidade de padrões da classe minoritária é baixa em comparação ao número de padrões da classe majoritária. É consequente de instâncias raras quando existem poucos padrões.

Para lidar com o problema de desbalanceamento de dados, duas categorias de métodos são utilizados, os métodos de *resampling* e os métodos *cost-sensitive.* Os métodos de *resampling* consistem em mecanismos para prover balanceamento ao conjunto, seja retirando padrões da classe majoritária ou inserindo padrões da classe minoritária. De acordo com os resultados de (Provost & Weiss, 2001), (Laurikkala, 2001) e (Japkowicz, Estabrooks, & Jo, 2004), para diversos classificadores, um conjunto de treinamento balanceado provê um desempenho na classificação superior em relação a conjuntos desbalanceados. Estes resultados justificam o uso de métodos de *resampling* em conjuntos desbalanceados.

Os métodos *cost-sensitive*, por sua vez, consideram os custos associados às classificações incorretas, ou seja, ao invés de criar distribuições balanceadas de dados, estes métodos criam matrizes que descrevem custos para classificações erradas de cada padrão.

Os métodos de *resampling* são o objetivo de estudo deste trabalho e podem ser divididos em *undersampling* e *oversampling*.

2.2.1 Oversampling

As técnicas da categoria *oversampling* proveem balanceamento ao conjunto de dados através da inserção de padrões da classe minoritária. Os padrões que serão inseridos podem vir de réplica de padrões existentes ou da criação de novos padrões artificiais, como é apresentado em (He, 2009).

De acordo com (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) "O método SMOTE (*Synthetic Minority Oversampling Technique*) para geração de padrões sintéticos é um método poderoso que tem mostrado um grande sucesso em várias aplicações". Este é o método de *oversampling* estudado neste trabalho.

As desvantagens dos métodos de *oversampling* são basicamente duas: a possibilidade de *overfitting* no treinamento e a presença de dados sintéticos no conjunto. O *overfitting* ocorre quando o modelo atinge um bom desempenho para classificar os padrões de treinamento, mas tem um desempenho ruim para classificar padrões inéditos. Quando um padrão é replicado por uma técnica de *oversampling*, ele pode ser apresentado para o modelo mais vezes do que os outros padrões, fazendo com o que o modelo se especialize.

2.2.2 Undersampling

As técnicas da categoria *undersampling* também proveem balanceamento ao conjunto de dados, porém através da remoção de padrões da classe majoritária. Os métodos variam na escolha de quais padrões serão removidos.

A desvantagem dos métodos de *undersampling* está na possibilidade de remover padrões fundamentais, o que pode levar o classificador a perder importantes conceitos pertencentes à classe majoritária. Isso ocorre principalmente em métodos como o RUM (*Random Undersampling Method*), cujo critério de remoção é a escolha aleatória.

2.3 Validação

Um ponto importante do processo de construção de um classificador de padrões é a validação do mesmo. É preciso, através de uma métrica, saber se o classificador está tendo um bom desempenho em aprender do conjunto de treinamento e classificar o conjunto de teste. Supondo um ambiente de classificação binária, ou seja, onde existam apenas duas classes possíveis às quais os padrões podem pertencer e sendo as classes possíveis 1 e 0, temos a matriz situação demonstrada pela Figura 3:

		Classe Predita	
		1	0
Classe Real	1	VP	FN
	0	FP	VN

Figura 3 - Matriz de confusão genérica.

A Figura 3 apresenta a matriz de confusão, para um ambiente de classificação binária, e é interessante para o entendimento de processos de validação. Essa matriz mostra o número de classificações reais em oposição às classificações preditas, pelo classificador, para cada classe.

Quando possuímos um classificador binário já treinado e recebemos um padrão desconhecido para classificação, existem quatro possibilidades. VP (Verdadeiro Positivo), quando o modelo classifica como 1 e o padrão é de fato um 1. VN (Verdadeiro Negativo), quando o modelo classifica como um 0 e o padrão é de fato um 0. FP (Falso Positivo), quando o modelo classifica como 1, mas na verdade o padrão é um 0. FN (Falso Negativo), quando o modelo classifica como um 0, mas na verdade é um 1. Duas possibilidades de acerto, representadas na diagonal principal, e duas possibilidades de erro, representadas na diagonal secundária da matriz.

Se um conjunto de teste é apresentado ao modelo treinado, este classificará os padrões e as saídas obtidas estarão representadas na matriz de confusão. A acurácia preditiva é maneira simples para avaliar o desempenho de classificadores, e é dada por:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

Entretanto, a acurácia preditiva pode provocar certas ilusões em algumas situações, além de ser extremamente sensível a mudanças no conjunto de dados. Em ambientes desbalanceados, um classificador que apresenta um desempenho ruim para classe minoritária, pode apresentar um bom desempenho para classe majoritária, terá um valor de acurácia preditiva alto, devido à distribuição desigual dos dados. De acordo com (He, 2009) "A métrica de acurácia preditiva não proporciona informações adequadas para avaliar a funcionalidade dos classificadores no que diz respeito ao tipo de classificação requerida".

Para validação dos experimentos neste trabalho, uma métrica adequada ao ambiente foi adotada. Para exemplificar o motivo de não se usar acurácia preditiva neste trabalho, considere a seguinte matriz de confusão de exemplo (a matriz foi construída através da base de dados que será apresentada na Seção 5.4.1):

		Classe Predita	
		1	0
Classe Real	1	7	76
	0	20	415

Figura 4 - Matriz de confusão de exemplo.

No caso acima temos um exemplo de medida ilusória. Temos um ambiente com o total de 518 classificações, sendo que o classificador tem uma taxa alta de FN e FP, falhando em classificar exemplos da classe 1. Entretanto, o classificador apresentou uma acurácia de 0,815 (acertou 81,5% das classificações). Esta base de dados apresenta um alto grau de desbalanceamento e para medir o desempenho de classificadores treinados nela, é necessário utilizar outras métricas.

Portanto, no capítulo seguinte será apresentada a alternativa que foi utilizada neste trabalho para a validação dos classificadores.

3 Revisão Bibliográfica

Este capítulo apresenta algumas técnicas conhecidas na literatura para o problema do desbalanceamento de dados. A Seção 3.1 apresenta a técnica de SMOTE e algumas variações da mesma. A Seção 3.2 apresenta abordagens para validar classificadores em ambientes desbalanceados.

Este trabalho propõe uma técnica para seleção de padrões para *resampling*. A revisão bibliográfica apresenta os principais trabalhos que serviram de base para a pesquisa, e a base para a maneira como a experimentação foi feita. A revisão bibliográfica pode ser dividida em duas partes.

A primeira parte da revisão apresenta alguns trabalhos que discutem os efeitos do desbalanceamento no conjunto de dados, além de trabalhos que utilizaram a técnica SMOTE e outros que propuseram variações na maneira que ela é aplicada em ambientes desbalanceados.

A segunda parte da revisão apresenta trabalhos que abordaram a validação de classificadores em ambientes desbalanceados, através do método ROC (Receiver Operating Characteristic), tanto no gráfico quanto na curva ROC. A revisão termina com a métrica utilizada para a experimentação deste trabalho, a AUC (Area Under Curve).

3.1 SMOTE e a remoção de *outliers* extremos

(Japcowicz, 2000) discutiu o efeito do desbalanceamento de dados em conjuntos de treinamento para classificadores de padrões, utilizando técnicas de oversampling e undersampling. Para o estudo, um conjunto unidimensional de dados artificiais foi utilizado. As técnicas utilizadas consistiam em igualar a distribuição dos conjuntos, primeiramente retirando padrões aleatórios da classe majoritária (método conhecido como *Random Undersampling Method*) e depois replicando padrões da classe minoritária (*Random Oversampling Method*). O estudo mostrou que as duas abordagens foram efetivas, e que a utilização de técnicas sofisticadas não apresentavam vantagens claras no ambiente testado.

(Chawla, Bowyer, Hall, & Kegelmeyer, 2002) propuseram a técnica conhecida como SMOTE para melhorar o desempenho de classificadores em classes minoritárias, criando padrões sintéticos. Os padrões sintéticos são criados separando (aleatoriamente) algumas amostras da classe minoritária e utilizando os vizinhos mais próximos de cada padrão da amostra, calculados a partir do algoritmo kNN (k-Nearest Neighbors). Os atributos dos novos padrões são calculados a partir da subtração de um padrão da amostra com um de seus vizinhos próximos, calculado com o kNN. A diferença é multiplicada por um número aleatório gap $\{\text{gap} \in \mathrm{R} \mid 1 \geq \text{gap} \geq 0\}$ e somada ao atributo do padrão da

amostra. Fazendo isso para todos os atributos, tem-se um padrão sintético. O SMOTE funciona ainda melhor quando é combinado com uma técnica de undersampling.

Existem dois parâmetros de configuração para o algoritmo SMOTE, o primeiro é a proporção de oversampling, que indica quantos padrões artificiais serão gerados (25%, por exemplo, aumentaria os padrões da classe minoritária em 25%) e o segundo é o número de vizinhos próximos que o algoritmo considera na geração dos padrões artificiais.

(Padmaja, Dhulipalla, Bapi, & Krishna, 2007) realizaram um estudo para o problema de detecção de fraudes em seguro. Uma fraude deste tipo é caracterizada por qualquer ato enganoso deliberado perpetrado contra ou pela companhia seguradora, corretor, prestador de serviço ou segurado com o propósito de obter ganho financeiro não garantido, podendo ocorrer durante o processo de contratação e utilização do seguro. Em seu estudo, utilizaram uma combinação de SMOTE e RUM para prover balanceamento ao conjunto. Além disso, utilizaram o algoritmo kRNN (k-Reverse Nearest Neighbors) para detectar e remover outliers extremos, pertencentes à classe minoritária, do conjunto de treinamento. Definindo outliers extremos como os pontos mais distantes em relação aos vizinhos, eles notaram um aumento no desempenho do classificador C4.5. Os autores sugeriram o uso da técnica em outros ambientes de desbalanceados, como em fraudes de cartão de crédito, entre outras.

Com base nos trabalhos citados, o intuito deste trabalho foi analisar a escolha dos padrões que são aplicados na técnica SMOTE para oversampling, através de uma abordagem alternativa.

3.2 Validação em ambientes desbalanceados

Como citado na Seção 2.3, a validação comum pode gerar análises ilusórias a respeito do desempenho do classificador que é treinado em um ambiente desbalanceado. Portanto, uma análise alternativa foi utilizada e será descrita nesta seção.

3.2.1 Análise ROC

Em (Fawcett, 2005) é apresentado o argumento de que o aumento na utilização da análise ROC se deve à pobreza de informação das métricas de acurácia padrão. Além disso, a análise ROC permite a visualização gráfica e também possui propriedades que a fazem especialmente útil em domínios desbalanceados ou com dados escassos. A análise ROC pode ser dividida em basicamente duas categorias, o gráfico ROC e a curva ROC. O gráfico ROC é um gráfico bidimensional em que a taxa de VP (Verdadeiros Positivos) é mostrada no eixo y e a taxa de FP (Falsos Positivos) é mostrada no eixo x. Cada classificador produz um ponto no gráfico ROC, o que permite dispor a relação de vários entre benefícios (Taxa

de VP) e os custos (Taxa de FP). A Figura 4 mostra dois pontos, A e B, que representam dois classificadores.

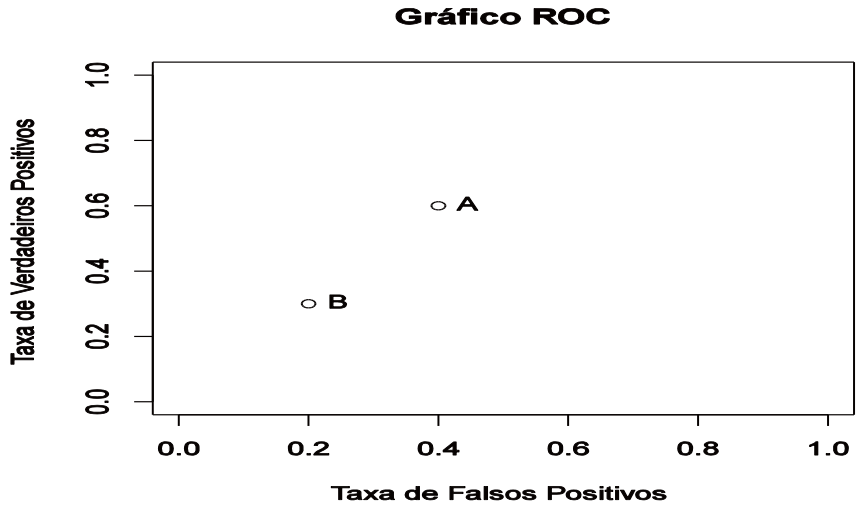

Figura 5 - Exemplo de gráfico ROC para dois classificadores..

O ponto B representa um classificador de desempenho abaixo do que o representado pelo ponto A. É interessante notar que o eixo $x = y$ representa classificadores aleatórios, uma vez que a taxa de VP é igual à taxa de FP. Temos o ponto ótimo, denominado ponto de classificação perfeita, em $(0,1)$.

3.2.2 Curva ROC e AUC

A curva ROC é uma variação do gráfico ROC. Alguns classificadores, como RNAs, produzem scores ou probabilidades que representam o grau de pertinência de um padrão para determinada classe. Através do uso de um limiar, é possível produzir uma saída discreta, sendo que cada valor de limiar produz um ponto na curva ROC, logo cada classificador gera uma curva. Portanto, toda curva ROC é na verdade uma função de etapa, que se aproxima de uma curva à medida que o número de padrões se aproxima do infinito. Esta curva tem a característica de não ser sensível à distribuição das classes do conjunto do treinamento do classificador.

De acordo com (Bradley, 1996), “análise ROC é uma boa maneira de visualizar o desempenho de classificadores, porém às vezes é necessário reduzir a análise a uma única métrica". No trabalho citado, o autor discute a utilização da medida AUC (Area Under Curve) como medida de desempenho para classificadores. A medida de AUC de um classificador corresponde à área da curva ROC que o mesmo construiu. Através de experimentos com diversos classificadores, o autor demonstrou que a medida de AUC demonstra boas propriedades para avaliar o desempenho de um classificador. Algumas propriedades são:

- Não é dependente da escolha de limiares (como a curva ROC);

- É invariante a distribuição das classes (fator importante para ambientes desbalanceados);
- Dá indicativas do desempenho do classificador em separar classes positivas e negativas.
- Reduz a análise toda a um único valor.

Para avaliar todos os classificadores neste trabalho, a métrica de AUC foi adotada.

4
Método Proposto

Este capítulo apresenta o método proposto neste trabalho. Na Seção 4.1 é apresentada a proposta para seleção de padrões utilizada neste trabalho. As hipóteses formuladas neste trabalho são apresentadas na Seção 4.2.

4.1 Seleção de Padrões

A proposta deste trabalho é uma abordagem diferente para escolher os padrões que serão replicados, ou que serão base para a criação de novos padrões artificiais. A abordagem proposta é a seleção através de um fator, definido neste trabalho como tipicidade.

O ambiente escolhido para este trabalho é um ambiente multidimensional, portanto cada padrão é composto por diversos atributos. A abordagem de seleção por tipicidade foi dividida em duas frentes, a seleção de padrões típicos e a seleção de padrões atípicos.

Mediante a passagem de um parâmetro que representa o número de atributos mínimos, o algoritmo seleciona os padrões considerados típicos e atípicos. Este parâmetro se trata de um número de atributos específicos que o padrão precisa ter para ser selecionado. Se for uma seleção de padrões típicos, este número de atributos representa o número mínimo de atributos típicos que o padrão deve ter. Para seleção de padrões atípicos, este número representa o número mínimo de atributos atípicos.

4.1.1 Tipicidade de atributos

Seja I_{P_xA} uma matriz de padrões de entrada, onde P é o número de padrões e A o conjunto de atributos em cada padrão e o atributo é um valor real, podemos calcular o VM (Vetor Médio) e o VDP (Vetor de Desvio Padrão) dos atributos através de:

$$VM = \frac{\sum_{i=1}^{P} I_{[p_i][a_j]}}{P}, \forall\, j \in A$$

$$\mathrm{VDP} = \sqrt{\frac{\sum_{i=1}^{P}(I_{[p_i][a_j]} - VM)}{\mathrm{P} - 1}}, \forall\, j \in A$$

Nessas condições, um atributo at é considerado normal se o seu valor satisfaz a seguinte condição:

$$VM - DP \leq at \leq DP + VM$$

Caso não satisfaça a condição acima, o atributo é considerado anormal. É importante frisar que dois algoritmos foram implementados, um que seleciona apenas padrões típicos do conjunto e outro que seleciona apenas padrões atípicos.

Se, por exemplo, o padrão contém 19 atributos no total, o parâmetro de atributos típicos mínimos for 10, e se trata da seleção de padrões típicos, a condição acima precisa ser satisfeita para pelo menos 10 atributos, caso contrário o padrão não é selecionado para o *resampling*. Usando o mesmo exemplo, porém com uma seleção de padrões atípicos, a condição acima não pode ser satisfeita para pelo menos dez atributos, caso contrário o padrão não é selecionado para *resampling*.

Conforme o número de atributos mínimos aumenta, menos padrões são selecionados para ambos os algoritmos. As tabelas abaixo exemplificam esta ocorrência em uma das bases de dados utilizada, que será descrita na Seção 5.4.2, através da passagem de diferentes valores para os atributos mínimos, sendo a primeira tabela para seleção de padrões atípicos e a segunda para seleção de padrões típicos:

Tabela 1 – Exemplos de proporções da seleção de padrões atípicos.

Atributos Atípicos Mínimos Passados	Proporção de Padrões Atípicos Selecionados
5	*15,90%*
7	*6,28%*
9	*0,62%*

Tabela 2 - Exemplos de proporções da seleção de padrões típicos.

Atributos Típicos Mínimos Passados	Proporção de Padrões Típicos Selecionados
5	*21,75%*
7	*10,66%*
9	*1,88%*

4.2 Padrões escolhidos em métodos de *resampling*

Uma vez que os padrões foram escolhidos, a proposta foi utilizá-los como entrada para o método SMOTE e comparar os resultados com a abordagem clássica, aleatória. Tanto a seleção de padrões típicos quanto a de padrões atípicos foram comparados à abordagem clássica. Uma vez que os testes foram definidos, as seguintes hipóteses se formaram:

$$\begin{cases} \text{H0: A seleção de padrões típicos não apresenta diferenças para a seleção aleatória.} \\ H1: \textit{A seleção de padrões normais tem melhores resultados do que a seleção aleatória.} \end{cases}$$

$$\begin{cases} \text{H0: A seleção de padrões atípicos não apresenta diferenças para a seleção aleatória.} \\ H1: \textit{A seleção de padrões anormais tem melhores resultados do que a seleção aleatória.} \end{cases}$$

Caso alguma das hipóteses nulas seja rejeitada para os casos acima, a utilização de seleção de padrões por tipicidade pode ser uma alternativa à abordagem aleatória.

5 Metodologia Utilizada

Este capítulo apresenta a metodologia utilizada neste trabalho. Na Seção 5.1 é apresentado o framework que foi desenvolvido para auxiliar nos testes. A explicação sobre a utilização dos padrões escolhidos como entrada para métodos de resampling e a base de dados é feita na Seção 5.2. Uma biblioteca que foi utilizada é apresentada na Seção 5.3. A seção 5.4 apresenta as bases de dados utilizadas no estudo.

5.1 Framework

Neste trabalho, um *framework* foi desenvolvido para facilitar a implementação de um ambiente que permitisse fazer experimentos de maneira metodológica e eficiente. Segundo os autores de (FAYAD & SCHMIDT, 1997), "*framework* é um conjunto de classes que colaboram para realizar uma responsabilidade para um domínio de um subsistema da aplicação". Portanto, através de um conjunto de classes e interfaces, foi possível decompor os processos de aprendizado e teste, e criar um conjunto flexível e extensível de objetos para resolver os problemas especificando apenas as particularidades de cada aplicação futura. Abaixo estão listadas algumas características do framework:

- Implementação de tipos variados de classificadores;
- Implementação de tipos variados de *resampling;*
- Implementação de tipos variados de técnicas para seleção de padrões;
- Combinação de variados tipos de *resampling* para um mesmo teste e testes diferentes;
- Combinação de variados tipos de seletores de padrões para um mesmo teste e testes diferentes;
- Implementação de modelos de validação com métricas.

Com o uso do *framework* foi possível criar um ambiente de experimentação que fosse distribuído e que viabilizasse o teste com classificadores diferentes configurados com parâmetros diferentes.

5.2 Experimentação

Para avaliar a abordagem da seleção em relação à abordagem aleatória, dois conjuntos de experimento foram feitos;

- Experimento com SMOTE;
- Experimento com SMOTE e RUM.

Para cada teste, os classificadores foram treinados com três tipos diferentes de conjuntos:

- Conjunto com SMOTE usando abordagem aleatória;
- Conjunto com SMOTE usando a seleção de padrões típicos;
- Conjunto com SMOTE usando a seleção de padrões atípicos.

Após a fase de treinamento, os classificadores foram usados para classificar um mesmo conjunto de teste.

5.2.1 Ciclos de testes

Para cada conjunto citado, vinte classificadores, sendo dezenove RNAs e um C4.5, eram treinados. Sendo que a semente geradora que atribuía os pesos iniciais para as RNAs variava para cada uma, o classificador gera resultados diferentes após cada treinamento. Assim que cada conjunto era treinado e testado, a sua medida de AUC era calculada. Ao final da execução. As sementes usadas para iniciar a matriz de pesos da RNA também foram geradas aleatoriamente. Existe apenas um C4.5 pois apresentava o mesmo resultado para todos os treinamentos, uma vez que não necessita de parâmetros para ser construído.

É interessante observar que cada teste é montado a partir dos parâmetros para o SMOTE, para a seleção de padrões e para o RUM, gerando assim três conjuntos. Sendo assim, é preciso variar esses parâmetros de configuração para obter um número maior de resultados e aumentar a confiabilidade do experimento. Para este trabalho, cem conjuntos diferentes de parâmetros foram usados para configurar os algoritmos de pré-processamento. Cada conjunto de parâmetros de configuração contém os seguintes parâmetros:

- Proporção de *oversampling* do SMOTE;
- Número de vizinhos próximos utilizados pelo SMOTE;

- Proporção de *oversampling* do SMOTE seletivo;
- Número de atributos típicos mínimos (Parâmetro para a seleção de padrões);
- Proporção de undersampling (Parâmetro para o RUM).

A parametrização é um ponto negativo neste experimento, pois os resultados são dependentes dela. É possível gerar combinações de padrões a fim de aperfeiçoar o processo de treinamento e obter um melhor resultado no teste. Como o objetivo deste trabalho é analisar a seleção de padrões através da tipicidade para algoritmos de *oversampling,* esta opção foi descartada. Portanto, um número grande de conjuntos de padrões foi gerado para analisar a média de desempenho dos classificadores.

Embora este trabalho analise apenas o comportamento de algoritmos de *oversampling,* o teste combinando a técnica SMOTE (de *oversampling*) com a técnica RUM (de *undersampling*) foi também realizado. O intuito era analisar as diferenças de desempenho do SMOTE aplicado unicamente e aplicado em conjunto com outras técnicas.

5.3 Biblioteca Weka

Para implementar os classificadores, o *software* (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009) foi utilizado como base. O *software* possui várias aplicações em trabalhos científicos e sua utilização é muito recomendada. O Weka possui uma coleção estável de algoritmos de aprendizado de máquina, além de ferramentas para processamento de dados de entrada, como filtros, operadores de normalização, entre outras.

Neste trabalho, a API do Weka para linguagem Java foi utilizada para implementar os classificadores, a leitura dos conjuntos de treinamento e teste, os métodos de seleção de padrões, os métodos de *resampling* e os validadores de desempenho.

5.4 Bases de dados

Duas bases de dados foram escolhidas para a realização deste estudo. A primeira base é uma base de solavancos sísmicos e apresenta um alto grau de desbalanceamento e uma grande dificuldade para classificação. A segunda base é uma base de câncer de mama, que apesar de apresentar um grau de desbalanceamento, não é tão difícil de classificar.

5.4.1 Base de Solavancos Sísmicos

A primeira base de dados é uma base de solavancos sísmicos e de medições em minas de carvão. Um solavanco é definido como uma liberação de energia de deformação biológica, resultando na expulsão de carvão a partir de um pilar ou costela da mina. Segundo os autores de (Ellenberger & Heasley), "mineiros que trabalham em condições como estas, devem ser evacuados caso o solavanco gere abalos sísmicos".

Na base temos duas classes possíveis para os padrões: caso de evacuação e caso de não evacuação dos operários. Considere a classe 0 como solavancos não tão perigosos e a classes 1 como solavancos em que é preciso evacuar a mina. Caso seja notado que o solavanco pode gerar um abalo sísmico de fato, a mina deve ser evacuada imediatamente para evitar catástrofes. Se nos basearmos em uma matriz de confusão, introduzida na seção 4.3, os casos de FN podem gerar uma catástrofe, pois o classificador não classificou corretamente um momento de perigo. Os casos de FP causam prejuízos, pelo atraso ocasionado pela evacuação.

Tendo o conjunto total de abalos, é preciso dividí-lo em conjuntos de treinamento e teste do modelo. A tabela a seguir mostra a divisão que foi feita.

Tabela 3 - Divisão dos conjuntos da base de solavancos sísmicos.

Conjuntos	Proporção	Número de Padrões	Número de Atributos
Inicial	*100%*	*2584*	*19*
Treinamento	*79,95%*	*2066*	*19*
Teste	*20,05%*	*518*	*19*

Esta base de dados foi escolhida para o estudo devido ao alto grau de desbalanceamento dos dados. A tabela a seguir mostra as proporções entre as classes para o conjunto de treinamento:

Tabela 4 - Distribuição de classes no conjunto de treinamento da base de solavancos sísmicos.

Classes	Proporção	Número de Padrões	Número de Atributos
Classe 0	*95,78%*	*1979*	*19*
Classe 1	*4,22%*	*87*	*19*

Portanto, o conjunto de treinamento tem um alto nível desbalanceamento relativo, sendo assim escolhido para os experimentos deste trabalho. A primeira base utilizada pode ser encontrada no seguinte endereço eletrônico:

https://archive.ics.uci.edu/ml/datasets/seismic-bumps

5.4.2 Base de Câncer de Mama

A segunda base de dados é uma base de características de biópsias feitas em massas mamárias. A base de dados é bastante conhecida e utilizada em trabalhos científicos, vide (Wolberg, Street, & Mangasarian, 1999), e foi disponibilizada gratuitamente para estudos de aprendizado de máquina. Os autores identificaram visualmente algumas características de biópsias consideradas relevantes para o diagnóstico. Através da utilização de um algoritmo classificador, os padrões foram separados e a base de dados, conhecida como *Winsconsin Breast Cancer Data,* foi criada.

Na base temos duas classes possíveis para os padrões: caso de tumor maligno e caso de tumor benigno. Considere a classe 0 como tumores benignos e a classe 1 como tumores malignos. Caso seja notado que o tumor é de fato maligno, o paciente deve receber tratamento. Se nos basearmos em uma matriz de confusão, introduzida na seção 4.3, os casos de FN podem gerar um problema grave ao paciente, pois o classificador não classificou corretamente um tumor maligno. Os casos de FP causam problemas menores.

Tendo o conjunto total de abalos, é preciso dividí-lo em conjuntos de treinamento e teste do modelo. A tabela a seguir mostra a divisão que foi feita.

Tabela 5 - Divisão dos conjuntos da base de câncer de mama.

Conjuntos	Proporção	Número de Padrões	Número de Atributos
Inicial	*100%*	*683*	*11*
Treinamento	*69,98%*	*478*	*11*
Teste	*30,01%*	*205*	*11*

Esta base possui um desbalanceamento de dados, porém o grau é menor do que a base de solavancos. A tabela a seguir mostra as proporções entre as classes para o conjunto de treinamento:

Tabela 6 - Distribuição de classes no conjunto de treinamento da base de câncer de mama.

Classes	Proporção	Número de Padrões	Número de Atributos
Classe 0	*68,41%*	*327*	*11*
Classe 1	*31,59%*	*151*	*11*

Portanto, o conjunto de treinamento apresenta um nível desbalanceamento relativo. A segunda base utilizada pode ser encontrada no seguinte endereço eletrônico:

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

6 Resultados

Este capítulo apresenta os resultados deste trabalho. Na Seção 6.1 é feita a apresentação dos resultados para a base de solavancos sísmicos. A validação da base de solavancos sísmicos é apresentada na Seção 6.2. Os resultados para a base de câncer de mama são apresentados na Seção 6.3. E finalmente, a validação para a base de câncer de mama é apresentada na Seção 6.4.

Para testar as hipóteses formuladas na Seção 4.2, é necessário um teste de comparação de medianas. Inicialmente, verifica-se se a distribuição dos dados é uma distribuição normal, através do teste de Shapiro-Wilk (Shapiro & Wilk, 1965). Caso não seja, é preciso utilizar um teste não paramétrico para amostras não pareadas, o teste de Wilcoxon (Wilcoxon, 1945).

A comparação entre as medianas tem por objetivo escolher o método com melhor desempenho. Caso o método proposto neste trabalho seja o escolhido, há indícios que apontam que ele é melhor do que a abordagem de escolha aleatória, abordagem mais utilizada atualmente.

6.1 Apresentação dos Resultados para a Base de Solavancos Sísmicos

A seguir os resultados dos testes de SMOTE e SMOTE combinado com RUM para a base de solavancos sísmicos são apresentados. Para cada teste temos uma tabela que apresenta os índices de média de todos os classificadores, desvio padrão e mediana para valores de AUC.

6.1.1 Teste SMOTE para a base de solavancos sísmicos

O teste SMOTE foi realizado treinando vinte classificadores para cada conjunto de parâmetros de configuração. Como temos cem conjuntos diferentes de parâmetros de configuração, dois mil classificadores ao todo foram treinados. O primeiro teste é dividido em quatro experimentos, o treinamento dos dez mil classificadores com o conjunto bruto, o treinamento para o SMOTE comum, o treinamento para o SMOTE com padrões típicos e o

treinamento para o SMOTE com padrões atípicos. Para a exibição dos dados, foi adotado um arredondamento de três casas decimais. Em tais condições, foram obtidos os seguintes dados:

Tabela 7 - Resultados do teste SMOTE para a base de solavancos sísmicos

Experimento	Média AUC	Máximo	Mínimo
Conjunto Bruto	*0,549*	*0,719*	*0,344*
SMOTE	*0,522*	*0,692*	*0,364*
SMOTE Típicos	*0,560*	*0,719*	*0,344*
SMOTE Atípicos	*0,559*	*0,719*	*0,344*

É interessante reparar que devido à não otimização de parâmetros do SMOTE, a média de desempenho dos classificadores foi maior no conjunto bruto do que com o SMOTE comum. As aplicações do SMOTE com seleções de padrões, típicos e atípicos, tiveram uma medida de AUC média maior do que o SMOTE com a abordagem clássica de seleção aleatória. Para ilustrar os resultados, foram gerados histogramas dos valores obtidos pelos experimentos. O eixo x apresenta a média de AUC dos classificadores e o eixo y apresenta a frequência, ou o número de classificadores, com as respectivas médias. Abaixo estão os histogramas dos conjuntos de treinamento do primeiro teste:

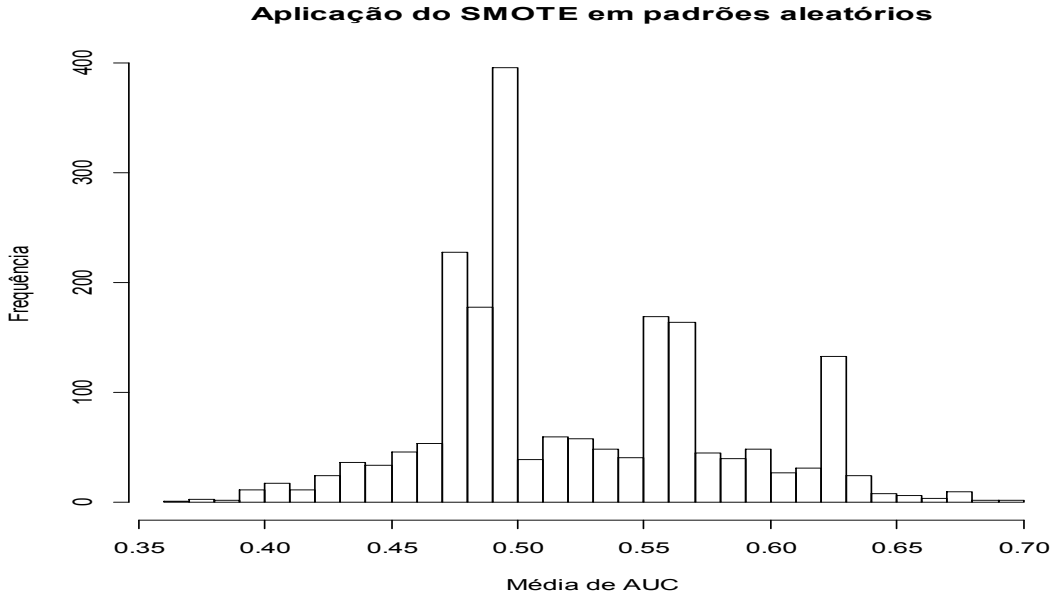

Figura 6 - Histograma do conjunto de treinamento com SMOTE para o primeiro teste para a base de solavancos sísmicos.

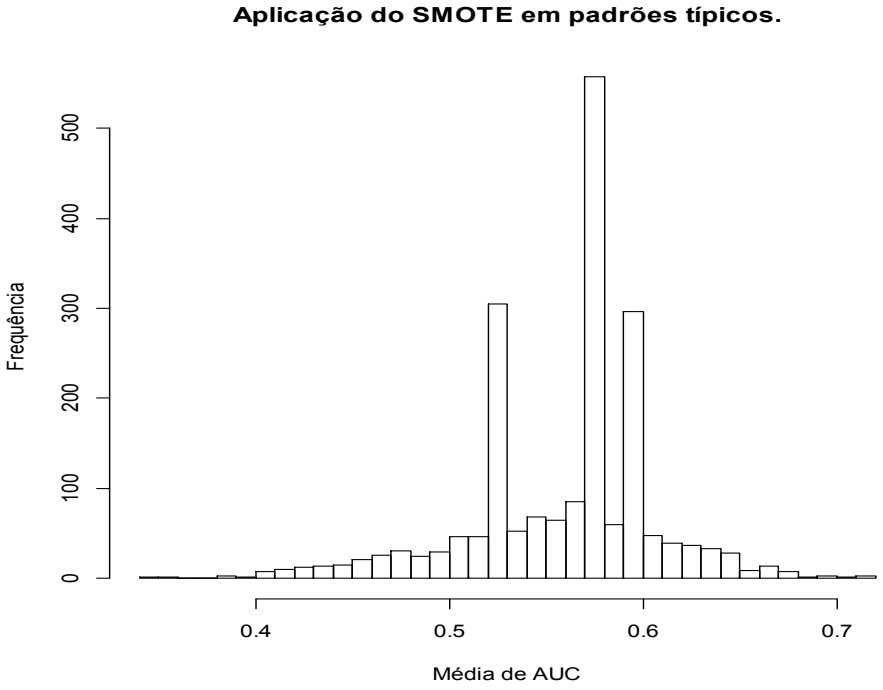

Figura 7 - Histograma do conjunto de treinamento com SMOTE em padrões típicos para o primeiro teste para a base de solavancos sísmicos.

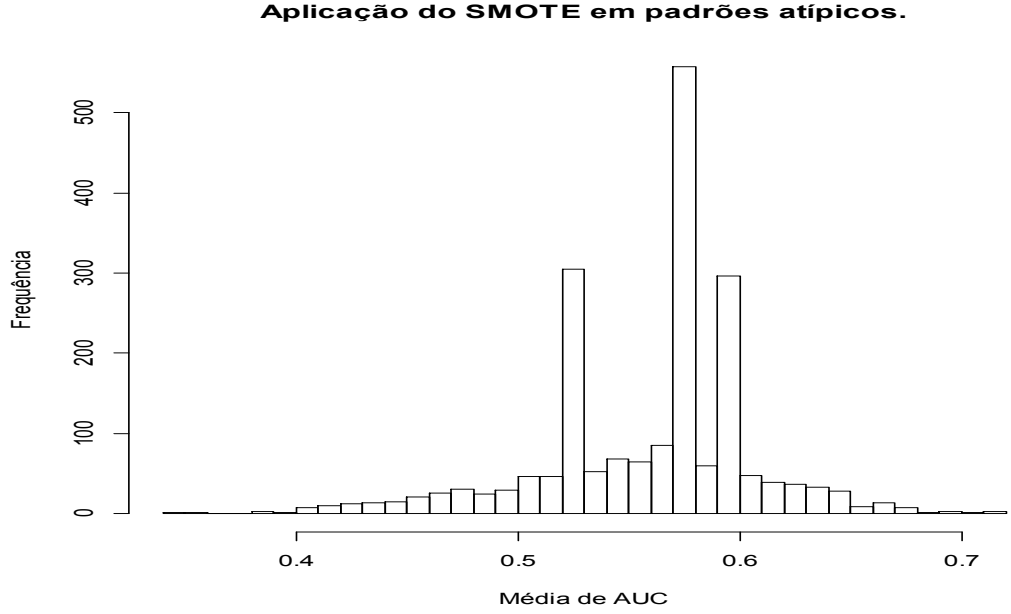

Figura 8 - Histograma do conjunto de treinamento com SMOTE em padrões atípicos para o primeiro teste para a base de solavancos sísmicos.

6.1.2 Teste de SMOTE e RUM para a base de solavancos sísmicos

O teste com SMOTE e RUM, foi realizado com o mesmo número de classificadores e conjuntos de parâmetros de configuração do primeiro teste. A diferença é que agora os conjuntos de dados passaram também pelo método RUM de *undersampling*. Em tais condições, foram obtidos os seguintes dados:

Tabela 8 - Resultados do teste SMOTE e RUM para a base de solavancos sísmicos.

Experimento	Média AUC	Máximo	Mínimo
Conjunto Bruto	*0,549*	*0,711*	*0,391*
SMOTE RUM	*0,565*	*0,713*	*0,378*
SMOTE RUM Típicos	*0,578*	*0,739*	*0,390*
SMOTE RUM Atípicos	*0,576*	*0,759*	*0,368*

Desta vez, mesmo não utilizando parâmetros de configuração otimizados, a técnica SMOTE RUM com abordagem clássica teve uma medida de AUC superior ao treinamento com conjunto bruto. Embora tenham resultados muito parecidos, as seleções de padrões, típicos e atípicos, se saíram melhor do que o SMOTE com a abordagem clássica de seleção aleatória. Assim como o teste passado, as aplicações do SMOTE RUM com seleções de padrões, típicos e atípicos, tiveram uma medida de AUC média maior do que o SMOTE com a abordagem clássica de seleção aleatória. Abaixo estão os histogramas dos conjuntos de treinamento do primeiro teste:

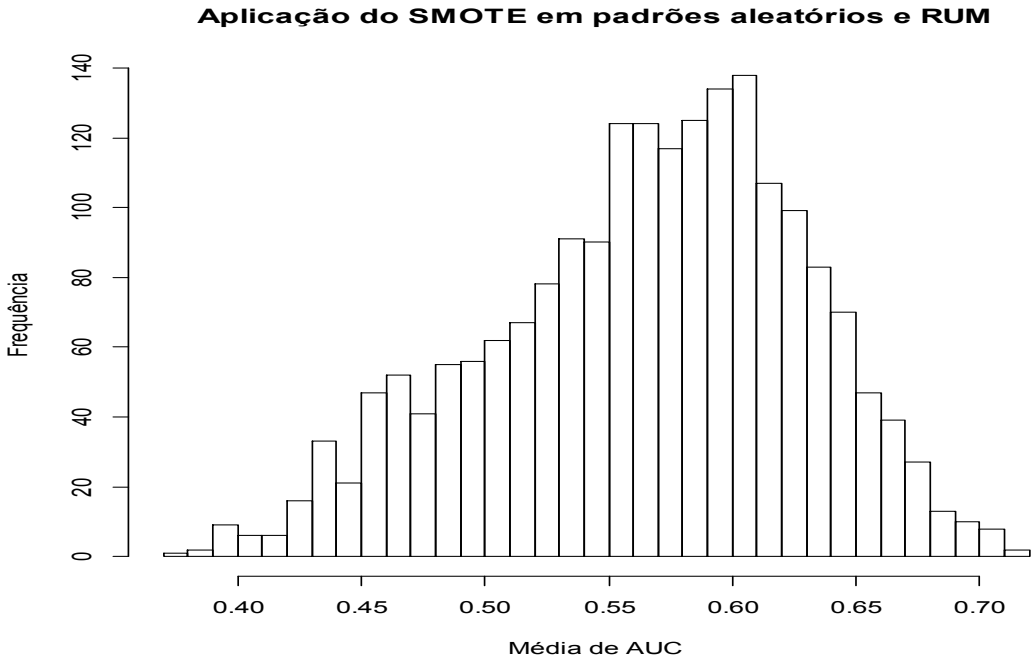

Figura 9 - Histograma do conjunto de treinamento com SMOTE e RUM para o segundo teste para a base de solavancos sísmicos.

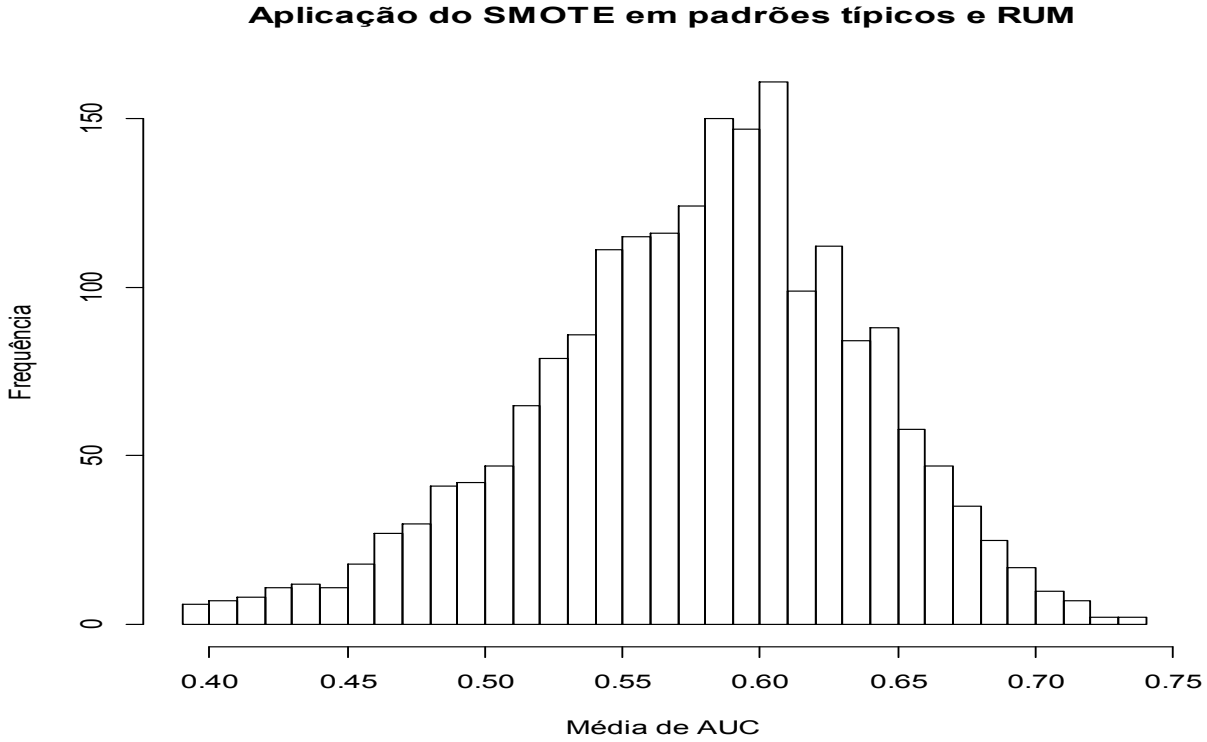

Figura 10- Histograma do conjunto de treinamento com SMOTE e RUM em padrões típicos para o segundo teste para a base de solavancos sísmicos.

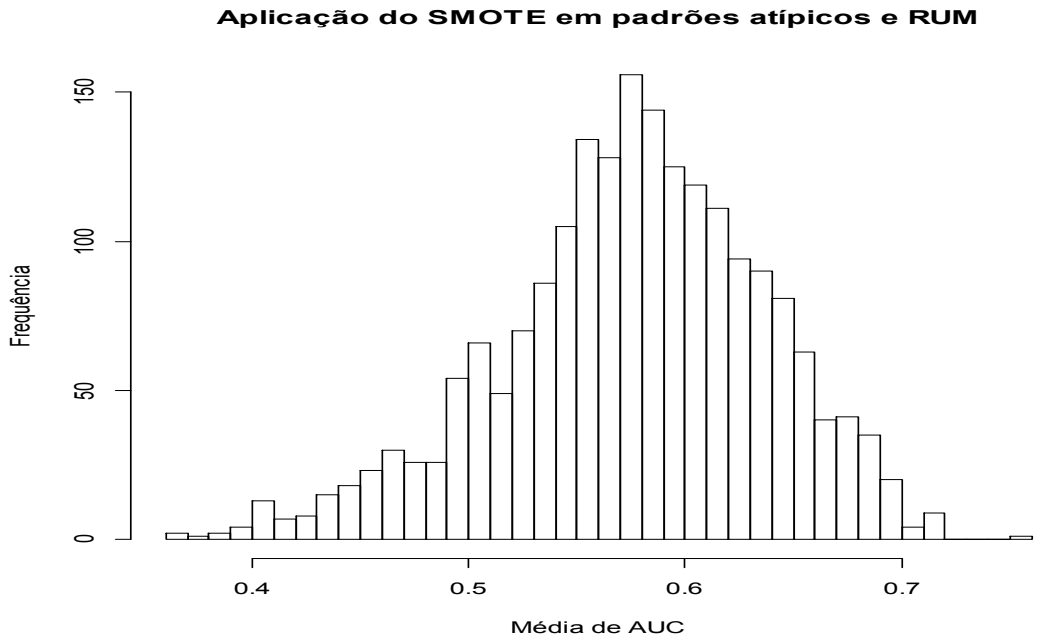

Figura 11 - Histograma do conjunto de treinamento com SMOTE e RUM em padrões atípicos para o segundo teste para a base de solavancos sísmicos.

Apesar das visualizações, é necessário testar as hipóteses introduzidas na seção 4.2. Os testes de hipóteses serão mostrados na próxima seção.

6.2 Validação dos Testes da Base de Solavancos Sísmicos

Para validar as hipóteses, é necessário fazer uma comparação de duas amostras pareadas, pois aplicamos técnicas diferentes de *resampling* (SMOTE aleatório, SMOTE com seleção de padrões típicos e SMOTE com seleção de padrões atípicos) no mesmo conjunto original. Os testes devem ser feitos aos pares, comparando cada SMOTE de seleção por tipicidade com o SMOTE aleatório. O primeiro teste considerado foi o teste t, porém para tal teste, teríamos que ter amostras provenientes de uma população normal.

Para verificar se os dados provêm de uma distribuição normal, o teste de Shapiro-Wilk (Shapiro & Wilk, 1965) foi utilizado. Este é um teste de normalidade que utiliza a hipótese nula para verificar se uma amostra provém de uma população normal. Através deste teste, é possível escolher um teste viável para os testes de hipótese. Os resultados para o primeiro teste (SMOTE) estão na tabela abaixo:

Tabela 9 - Teste de Shapiro-Wilk para o SMOTE para a base de solavancos sísmicos

Conjunto	P-Value
Bruto	*0*
SMOTE	*0*
SMOTE Típicos	*0*
SMOTE Atípicos	*0*

Em todos os conjuntos, deve-se rejeitar a hipótese nula, ou seja, rejeita-se a hipótese de que os dados provêm de uma população normal. A tabela a seguir mostra os resultados para o segundo teste (SMOTE RUM):

Tabela 10 - Teste de Shapiro-Wilk para o SMOTE e RUM para a base de solavancos sísmicos

Conjunto	P-Value
SMOTE RUM	*0*
SMOTE RUM Típicos	*0*
SMOTE RUM Atípicos	*0*

Da mesma forma que os conjuntos do primeiro teste, deve-se rejeitar a hipótese nula, ou seja, rejeita-se a hipótese de que os dados provêm de uma população normal.

A solução encontrada foi utilizar um teste não paramétrico para comparação de médias. O teste escolhido foi o teste de Wilcoxon para experimentos não pareados, (Wilcoxon, 1945). Através de uma variação deste teste, foi possível reescrever as hipóteses introduzidas na seção 4.2 e testar a superioridade ou inferioridade das médias sem fazer comparações das mesmas. Considerando μ_0 a média de AUC do SMOTE, μ_1 a média de AUC do SMOTE em padrões típicos e μ_2 média de AUC do SMOTE em padrões atípicos, a seguir são apresentadas as hipóteses e os resultados da aplicação do teste de Wilcoxon para as respectivas médias:

$$C_1 \begin{cases} H0: \mu_0 \geq \mu_1 \\ H1: \mu_1 > \mu_0 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e do SMOTE em padrões típicos, caso a hipótese nula seja rejeitada, há indícios que o SMOTE em padrões típicos tem média de AUC maior que SMOTE em padrões aleatórios.

$$C_2 \begin{cases} H0: \mu_0 \geq \mu_2 \\ H1: \mu_2 > \mu_0 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e do SMOTE em padrões atípicos, caso a hipótese nula seja rejeitada, há indícios que o SMOTE em padrões atípicos tem média de AUC maior que SMOTE em padrões aleatórios. A tabela abaixo mostra os resultados do teste de Wilcoxon nas hipóteses formuladas:

Tabela 11 - Resultados do teste de Wilcoxon para o primeiro teste para a base de solavancos sísmicos.

Conjunto de Hipóteses	P-Value
C_1	*0*
C_2	*0*

Para o segundo teste (com a utilização RUM), temos as hipóteses parecidas, as diferenças são as médias, pois agora temos que μ_3 é a média de AUC do SMOTE e RUM, μ_4 é a média de AUC do SMOTE em padrões típicos e RUM e μ_5 é a média de AUC do SMOTE em padrões atípicos. Para tais, temos os seguintes conjuntos de hipóteses:

$$C_3 \begin{cases} H0: \mu_3 \geq u_4 \\ H1: \mu_4 > \mu_3 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e RUM e do SMOTE em padrões típicos e RUM. Caso a hipótese nula seja rejeitada, há indícios que o SMOTE em padrões típicos e RUM têm média de AUC maior que SMOTE e RUM.

$$C_4 \begin{cases} H0: \mu_3 \geq \mu_5 \\ H1: \mu_5 > \mu_3 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e RUM e do SMOTE em padrões atípicos e RUM. Caso a hipótese nula seja rejeitada, há indícios que o SMOTE em padrões atípicos e RUM têm média de AUC maior que SMOTE em padrões aleatórios e RUM. A tabela abaixo mostra os resultados do teste de Wilcoxon nas hipóteses formuladas:

Tabela 12 - Resultados do teste de Wilcoxon para o segundo teste para a base de solavancos sísmicos.

Conjunto de Hipóteses	P-Value
C_3	*0*
C_4	*0*

6.3 Apresentação dos Resultados para a Base de Câncer de Mama

A seguir os resultados dos testes de SMOTE e SMOTE combinado com RUM para a base de câncer de mama são apresentados. Para a exibição dos resultados, foi adotado um arredondamento de quatro casas decimais, pelo motivo das variações serem mais singelas. Do restante, as apresentações são feitas da mesma maneira que a dos testes da base de solavancos sísmicos.

6.3.1 Teste SMOTE para a base de câncer de mama

O mesmo número de classificadores e conjuntos de parâmetros de configuração dos testes da primeira base foi utilizado. Em tais condições, foram obtidos os seguintes dados:

Tabela 13 - Resultados do teste SMOTE para a base de solavancos sísmicos

Experimento	Média AUC	Máximo	Mínimo
Conjunto Bruto	*0,9848*	*0,9890*	*0,9702*
SMOTE	*0,9841*	*0,9909*	*0,9640*
SMOTE Típicos	*0,9844*	*0,9898*	*0,9727*
SMOTE Atípicos	*0,9844*	*0,9898*	*0,9727*

Da mesma forma que a primeira base de dados, a média de desempenho dos classificadores foi maior no conjunto bruto do que com o SMOTE comum. As aplicações do SMOTE com seleções de padrões, típicos e atípicos, tiveram uma medida de AUC média maior do que o SMOTE com a abordagem clássica de seleção aleatória. A seleção de padrões típicos e atípicos obtiveram os mesmos resultados. Abaixo estão os histogramas dos conjuntos de treinamento do primeiro teste:

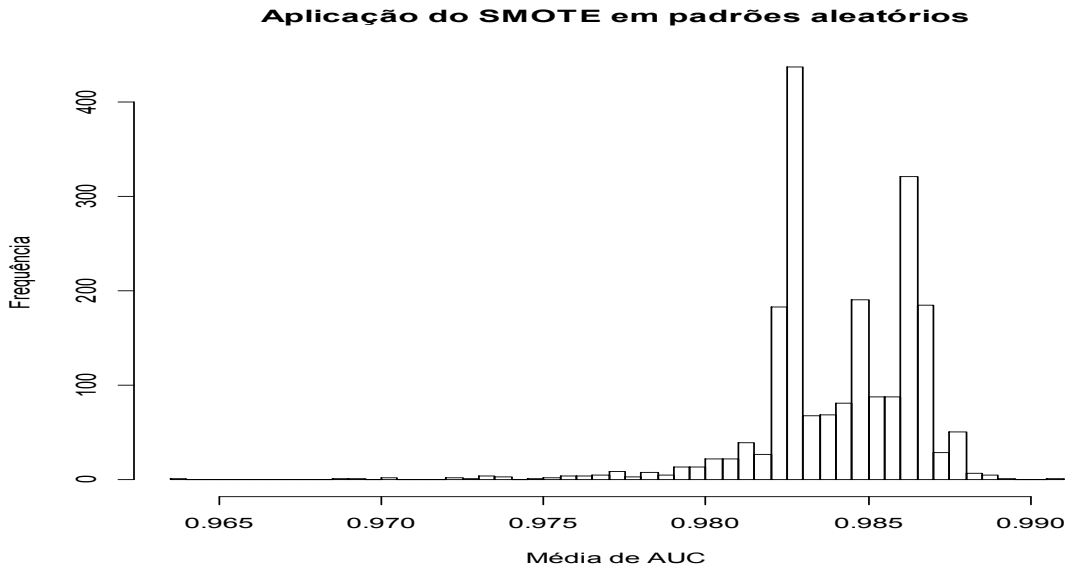

Figura 12 - Histograma do conjunto de treinamento com SMOTE para o primeiro teste para a base de câncer de mama.

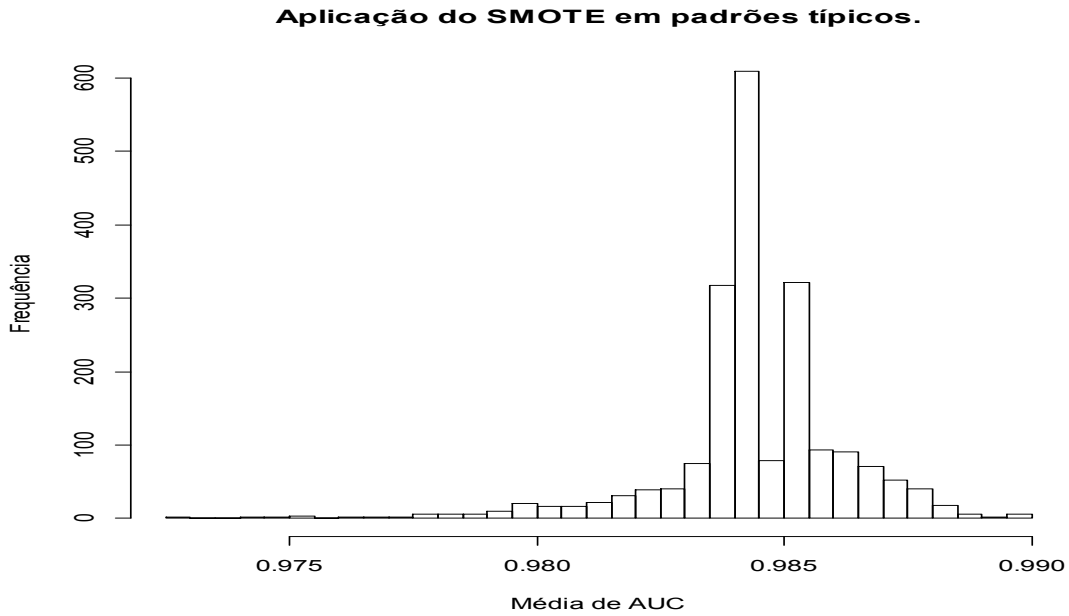

Figura 13 - Histograma do conjunto de treinamento com SMOTE em padrões típicos para o primeiro teste para a base de câncer de mama

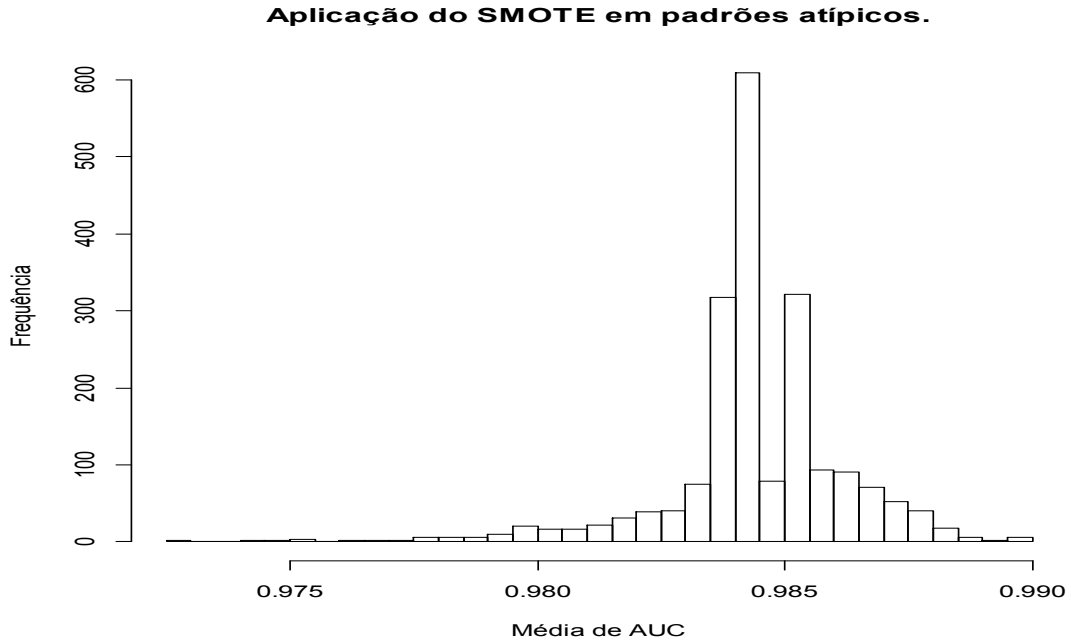

Figura 14 - Histograma do conjunto de treinamento com SMOTE em padrões atípicos para o primeiro teste para a base de câncer de mama.

Pelos histogramas apresentados, é possível também perceber a falta de padrão em relação à distribuição dos dados.

6.3.2 Teste de SMOTE e RUM para a base câncer de mama

O mesmo número de classificadores e conjuntos de parâmetros de configuração dos testes da primeira base foi utilizado. Em tais condições, foram obtidos os seguintes dados:

Tabela 14 - Resultados do teste SMOTE e RUM para a base de câncer de mama.

Experimento	Média AUC	Máximo	Mínimo
Conjunto Bruto	*0,9848*	*0,9890*	*0,9702*
SMOTE RUM	*0,9845*	*0,9931*	*0,9544*
SMOTE RUM Típicos	*0,9855*	*0,9934*	*0,9608*
SMOTE RUM Atípicos	*0,9855*	*0,9929*	*0,9409*

Embora tenham resultados muito parecidos, as seleções de padrões, típicos e atípicos, se saíram melhor do que o SMOTE com a abordagem clássica de seleção aleatória. Abaixo estão os histogramas dos conjuntos de treinamento do primeiro teste para a base de câncer de mama:

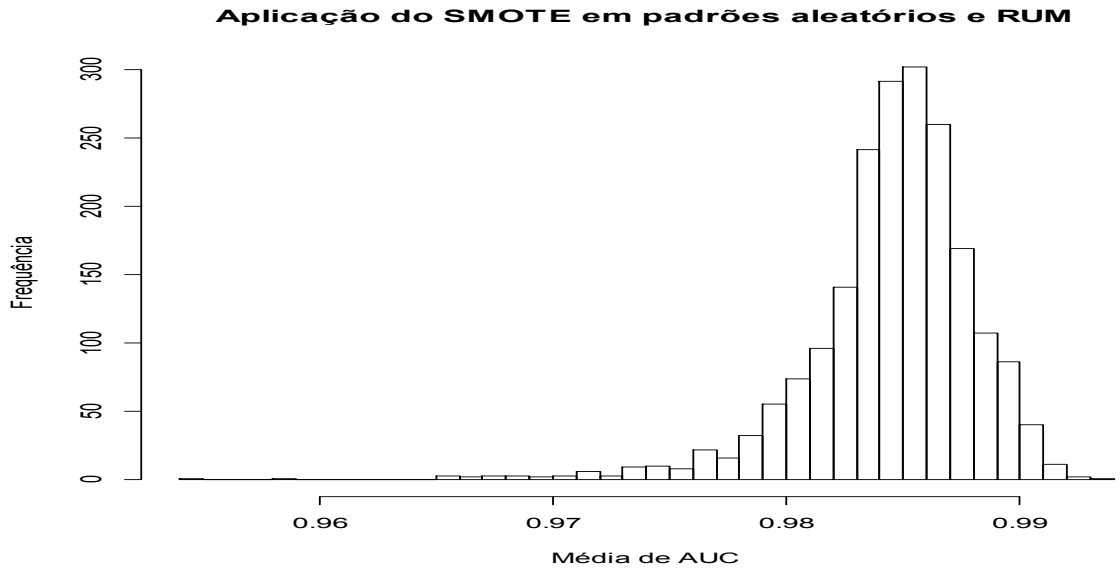

Figura 15 - Histograma do conjunto de treinamento com SMOTE e RUM para o segundo teste para a base de câncer de mama.

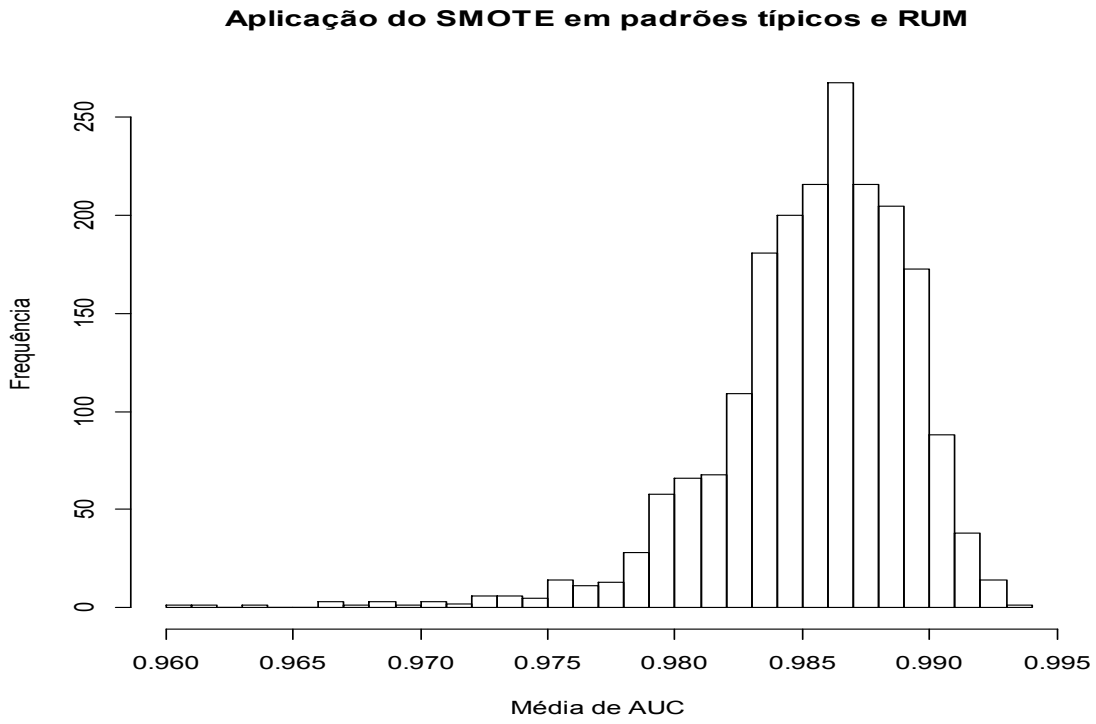

Figura 16 - Histograma do conjunto de treinamento com SMOTE e RUM em padrões típicos para o segundo teste para a base de câncer de mama.

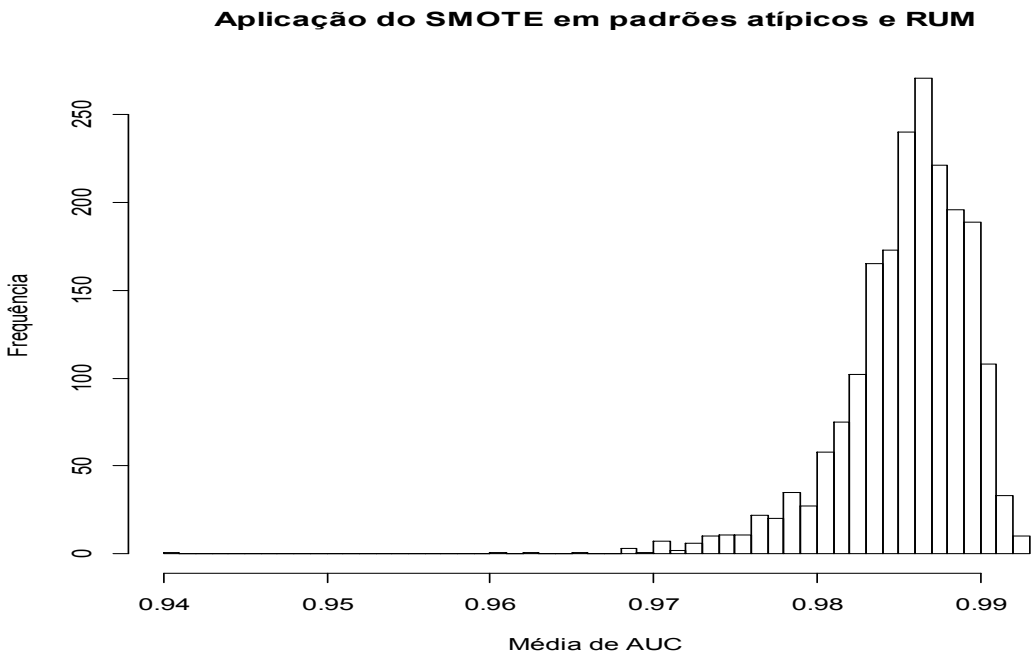

Figura 17 - Histograma do conjunto de treinamento com SMOTE e RUM em padrões atípicos para o segundo teste para a base de câncer de mama.

Apesar das visualizações, é necessário também testar as hipóteses introduzidas na seção 4.2 para a base de câncer de mama. Os testes de hipóteses serão mostrados na próxima seção.

6.4 Validação dos Testes Para a Base de Câncer de Mama

Da mesma maneira que a base de solavancos sísmicos, para verificar se os dados provêm de uma distribuição normal, o teste de Shapiro-Wilk (Shapiro & Wilk, 1965) foi utilizado. Os resultados para o primeiro teste (SMOTE) estão na tabela abaixo:

Tabela 15 - Teste de Shapiro-Wilk para o SMOTE para a base de câncer de mama.

Conjunto	P-Value
Bruto	*0*
SMOTE	*0*
SMOTE Típicos	*0*
SMOTE Atípicos	*0*

Em todos os conjuntos, deve-se rejeitar a hipótese nula, ou seja, rejeita-se a hipótese de que os dados provêm de uma população normal. A tabela a seguir mostra os resultados para o segundo teste (SMOTE RUM):

Tabela 16 - Teste de Shapiro-Wilk para o SMOTE e RUM para a base de câncer de mama.

Conjunto	P-Value
SMOTE RUM	*0*
SMOTE RUM Típicos	*0*
SMOTE RUM Atípicos	*0*

Da mesma forma que os conjuntos do primeiro teste, deve-se rejeitar a hipótese nula, ou seja, rejeita-se a hipótese de que os dados provêm de uma população normal.

Da mesma maneira que para a base de solavancos sísmicos, o teste de Wilcoxon para experimentos não pareados, (Wilcoxon, 1945) foi adotado. Considerando agora, para a base de câncer de mama, μ_0 a média de AUC do SMOTE, μ_1 a média de AUC do SMOTE em padrões típicos e μ_2 média de AUC do SMOTE em padrões atípicos, a seguir são apresentadas as hipóteses e os resultados da aplicação do teste de Wilcoxon para as respectivas médias:

$$C_1 \begin{cases} H0: \mu_0 \geq \mu_1 \\ H1: \mu_1 > \mu_0 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e do SMOTE em padrões típicos. Caso a hipótese nula rejeitada, há indícios que o SMOTE em padrões típicos tem média de AUC maior que SMOTE em padrões aleatórios.

$$C_2 \begin{cases} H0: \mu_0 \geq \mu_2 \\ H1: \mu_2 > \mu_0 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e do SMOTE em padrões atípicos. Caso a hipótese nula seja rejeitada, há indícios que o SMOTE em padrões atípicos

tem média de AUC maior que SMOTE em padrões aleatórios. A tabela abaixo mostra os resultados do teste de Wilcoxon nas hipóteses formuladas para a base de câncer de mama:

Tabela 17 - Resultados do teste de Wilcoxon para o primeiro teste para a base de câncer de mama.

Conjunto de Hipóteses	P-Value
C_1	*0*
C_2	*0*

Da mesma maneira que para a base de solavancos sísmicos, gora temos que μ_3 é a média de AUC do SMOTE e RUM, μ_4 é a média de AUC do SMOTE em padrões típicos e RUM e μ_5 é a média de AUC do SMOTE em padrões atípicos. Para tais, temos as hipóteses:

$$C_3 \begin{cases} H0: \mu_3 \geq \mu_4 \\ H1: \mu_4 > \mu_3 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e RUM e do SMOTE em padrões típicos e RUM, caso a hipótese nula seja rejeitada, há indícios que o SMOTE em padrões típicos e RUM têm média de AUC maior que SMOTE e RUM.

$$C_4 \begin{cases} H0: \mu_3 \geq \mu_5 \\ H1: \mu_5 > \mu_3 \end{cases}$$

A hipótese acima diz respeito às médias do SMOTE e RUM e do SMOTE em padrões atípicos e RUM, caso a hipótese nula seja rejeitada, há indícios que o SMOTE em padrões atípicos e RUM têm média de AUC maior que SMOTE em padrões aleatórios e RUM. A tabela abaixo mostra os resultados do teste de Wilcoxon nas hipóteses formuladas:

Tabela 18 - Resultados do teste de Wilcoxon para o segundo teste para a base de solavancos sísmicos.

Conjunto de Hipóteses	P-Value
C_3	*0*
C_4	*0*

7 Conclusões e Trabalhos Futuros

Este capítulo apresenta as conclusões do trabalho. A Seção 7.1 apresenta as conclusões tiradas a partir dos experimentos realizados. A Seção 7.2 apresenta algumas sugestões de trabalhos futuros.

7.1 Conclusões

O objetivo geral foi cumprido e corroborado pelos testes de hipótese da Seção 6.2. Em relação aos objetivos específicos:

- A seleção por tipicidade se mostrou uma maneira mais eficiente do que a aleatória de seleção de padrões para o método SMOTE;
- As técnicas de *resampling* aplicadas e geraram resultados diferenciados;
- O *framework* foi desenvolvido com sucesso e, aliado a uma aplicação, permitiu realizar os testes de maneira eficiente e distribuída;
- A proposta se mostrou interessante para as duas bases estudadas, com diferentes graus de desbalanceamento e de dificuldade de classificação;
- Através de testes de normalidade e não paramétricos, as hipóteses foram analisadas.

De acordo com os testes, existe um grande indício a favor da seleção de padrões pela tipicidade em relação à seleção de padrões aleatórios. Tanto na seleção de padrões típicos, quanto na seleção de padrões atípicos, os modelos conseguiram classificar padrões com melhores desempenhos. Portanto, sugere-se o uso deste tipo de abordagem em relação à abordagem aleatória.

A implementação desta abordagem não requer muitas alterações no experimento e não é complexa, demandando pouco tempo para implementação, além de ser independente de qual classificador está sendo usado no experimento. Além disso, é o *framework* desenvolvido possibilita a criação de novas abordagens para seleção de padrões. Os resultados obtidos neste trabalho serão publicados.

Este trabalho também sugere que todos os experimentos de classificação de padrões que atuam em ambientes desbalanceados, devem ser realizados com uma certa preocupação em relação à métrica de validação para o sistema. Métodos comuns e sensíveis à distribuição dos dados podem gerar resultados ilusórios e comprometer o experimento. Este trabalho sugere a utilização de curvas ROC, devido à sua insensibilidade à distribuição dos dados.

7.2 Trabalhos Futuros

Sugestões de trabalhos futuros:

- Implementar outras maneiras de selecionar padrões e verificar seu desempenho em relação à abordagem proposta neste trabalho e à abordagem aleatória;
- Verificar a seleção de padrões para métodos de *undersampling;*
- Verificar a seleção de padrões para métodos híbridos;
- Aplicar o método em outras bases de dados;
- Verificar a seleção de padrões em outros meios de operação, além da classificação de padrões.

8 Bibliografia

Bradley, A. P. (1996). The Use of Area Under Curve The ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 1145 - 1159.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *J. Artificial Intelligence Research, vol 16*, 321-357.

Ellenberger, J. L., & Heasley, A. K. (s.d.). *COAL MINE SEISMICITY AND BUMPS: HISTORICAL CASE STUDIES AND CURRENT FIELD ACTIVITY.* Pittsburgh, PA: National Institute for Occupational Safety and Health .

Fawcett, T. (19 de Dezembro de 2005). An introduction to ROC analysis. *Pattern Recognition Letters 27*, pp. 861 - 874.

FAYAD, M. E., & SCHMIDT, D. C. (1997). Object-oriented Application frameworks. *Communications of the ACM*, 10.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, H. I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*.

Haykin, S. (2000). *Redes Neurais - Princípios e Prática.* Bookman Editora .

He, H. (Setembro de 2009). Learning From Imbalanced Data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 1263 - 1283.

Japcowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence*.

Japkowicz, N., Estabrooks, A., & Jo, T. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*, 18-36.

Laurikkala, J. (2001). Improving Identification of Difficult Small Classes by Balancing Class Distribution. *Artificial Intelligence Medicine*, 63-66.

Lippmann, R. P. (1989). Pattern Classification Using Neural Networks. *IEEE Communications Magazine*, 47-64.

McCulloch, W. S., & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 115 - 133.

Norvig, P., & Russell, S. (2004). *Inteligência Artificial.* Prentice Hall.

Padmaja, T. M., Dhulipalla, N., Bapi, S. R., & Krishna, P. R. (2007). Unbalanced Data Classification Using extreme outlier Elimination and Sampling Techniques for

Fraud Detection. *15th International Conference on Advanced Computing and Communications*, (pp. 511-516).

Provost, G., & Weiss, F. (2001). The Effect of Class Distribution on Classifier Learning: An Empirical Study. *Technical Report ML-TR*.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*.

Rosenblatt, F. (1957). *The Perceptron: a perceiving and recognizing automaton.* Cornell Aeronautical Laboratory.

Shapiro, S., & Wilk, B. M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 591 - 611.

Triola, M. F. (2008). *Introdução à Estatística.* Rio de Janeiro: LTC Editora.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 80 - 83.

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1999). *A comparison of computer-based nuclear analysis versus lymph node status for staging breast cancer.* Clinical Cancer Research.