Journal of Artificial Intelligence Research 16 (2002) 321–357 Apresentado em 01/09; publicado em 02/06

SMOTE: Técnica de sobreamostragem de minoria sintética

Nitesh V. Chawla chawla@csee.usf.edu
Departamento de Ciência da Computação e Engenharia, ENB 118
University of South Florida
4202 E. Fowler Ave. Tampa,
FL 33620-5399, EUA

Kevin W. Bowyer kwb@cse.nd.edu
Departamento de Ciência da Computação e
Engenharia 384 Fitzpatrick Hall
Universidade de Notre Dame
Notre Dame, IN 46556, EUA

Lawrence O. Hall hall@csee.usf.edu
Departamento de Ciência da Computação e Engenharia, ENB 118
University of South Florida
4202 E. Fowler Ave. Tampa,
FL 33620-5399, EUA

W. Philip Kegelmeyer wpk@california.sandia.gov
Laboratórios Nacionais de Sandia
Departamento de Pesquisa Biosystems, PO Box 969, MS 9951
Livermore, CA, 94551-0969, EUA

Abstrato

Uma abordagem para a construção de classificadores a partir de conjuntos de dados desbalanceados é descrita. Um conjunto de dados é desequilibrado se as categorias de classificação não forem representadas de forma aproximadamente igual. Muitas vezes, os conjuntos de dados do mundo real são predominantemente compostos de exemplos “normais” com apenas uma pequena porcentagem de exemplos “anormais” ou “interessantes”. Também é o caso que o custo de classificar erroneamente um exemplo anormal (interessante) como um exemplo normal é muitas vezes muito maior do que o custo do erro reverso. A subamostragem da classe majoritária (normal) foi proposta como um bom meio de aumentar a sensibilidade de um classificador à classe minoritária. Este artigo mostra que uma combinação de nosso método de superamostragem da classe minoritária (anormal) e subamostragem da classe majoritária (normal) pode alcançar melhor desempenho do classificador (no espaço ROC) do que apenas subamostrar a classe majoritária. Este artigo também mostra que uma combinação de nosso método de superamostragem da classe minoritária e subamostragem da classe majoritária pode obter melhor desempenho do classificador (no espaço ROC) do que variar as taxas de perda em Ripper ou classes prioritárias em Naive Bayes. Nosso método de superamostragem da classe minoritária envolve a criação de exemplos sintéticos de classe minoritária. Os experimentos são realizados usando C4.5, Ripper e um classificador Naive Bayes. O método é avaliado usando a área sob a curva Receiver Operating Characteristic (AUC) e a estratégia de casco convexo ROC. Este artigo também mostra que uma combinação de nosso método de superamostragem da classe minoritária e subamostragem da classe majoritária pode obter melhor desempenho do classificador (no espaço ROC) do que variar as taxas de perda em Ripper ou classes prioritárias em Naive Bayes. Nosso método de superamostragem da classe minoritária envolve a criação de exemplos sintéticos de classe minoritária. Os experimentos são realizados usando C4.5, Ripper e um classificador Naive Bayes. O método é avaliado usando a área sob a curva Receiver Operating Characteristic (AUC) e a estratégia de casco convexo ROC. Este artigo também mostra que uma combinação de nosso método de superamostragem da classe minoritária e subamostragem da classe majoritária pode obter melhor desempenho do classificador (no espaço ROC) do que variar as taxas de perda em Ripper ou classes prioritárias em Naive Bayes. Nosso método de superamostragem da classe minoritária envolve a criação de exemplos sintéticos de classe minoritária. Os experimentos são realizados usando C4.5, Ripper e um classificador Naive Bayes. O método é avaliado usando a área sob a curva Receiver Operating Characteristic (AUC) e a estratégia de casco convexo ROC. Nosso método de superamostragem da classe minoritária envolve a criação de exemplos sintéticos de classe minoritária. Os experimentos são realizados usando C4.5, Ripper e um classificador Naive Bayes. O método é avaliado usando a área sob a curva Receiver Operating Characteristic (AUC) e a estratégia de casco convexo ROC. Nosso método de superamostragem da classe minoritária envolve a criação de exemplos sintéticos de classe minoritária. Os experimentos são realizados usando C4.5, Ripper e um classificador Naive Bayes. O método é avaliado usando a área sob a curva Receiver Operating Characte

1. Introdução

Um conjunto de dados é desequilibrado se as classes não estiverem aproximadamente igualmente representadas. Desequilíbrio na ordem de 100 para 1 é predominante na detecção de fraudes e desequilíbrio de até 100.000 para

1 foi relatado em outras aplicações (Provost & Fawcett, 2001). Houve tentativas de lidar com conjuntos de dados desequilibrados em domínios como chamadas telefônicas fraudulentas (Fawcett & Provost, 1996), gerenciamento de telecomunicações (Ezawa, Singh, & Norton, 1996), classificação de texto (Lewis & Catlett, 1994; Dumais, Platt, Heckerman, & Sahami, 1998; Mladenić & Grobelnik, 1999; Lewis & Ringuette, 1994; Cohen, 1995a) e detecção de derramamentos de óleo em imagens de satélite (Kubat, Holte, & Matwin, 1998).

O desempenho dos algoritmos de aprendizado de máquina é normalmente avaliado usando precisão preditiva. No entanto, isso não é apropriado quando os dados são desequilibrados e/ou os custos de diferentes erros variam acentuadamente. Como exemplo, considere a classificação de pixels em imagens de mamografia como possivelmente cancerígenos (Woods, Doss, Bowyer, Solka, Priebe, & Kegelmeyer, 1993). Um conjunto de dados de mamografia típico pode conter 98% de pixels normais e 2% de pixels anormais. Uma estratégia padrão simples de adivinhar a classe majoritária daria uma precisão preditiva de 98%. No entanto, a natureza do aplicativo requer uma taxa bastante alta de detecção correta na classe minoritária e permite uma pequena taxa de erro na classe majoritária para conseguir isso. A precisão preditiva simples claramente não é apropriada em tais situações. A curva Receiver Operating Characteristic (ROC) é uma técnica padrão para resumir o desempenho do classificador em uma faixa de compensações entre as taxas de erro verdadeiro positivo e falso positivo (Swets, 1988). A Área Sob a Curva (AUC) é uma métrica de desempenho tradicional aceita para uma curva ROC (Duda, Hart e Stork, 2001; Bradley, 1997; Lee, 2000). O casco convexo ROC também pode ser usado como um método robusto de identificação de classificadores potencialmente ótimos (Provost & Fawcett, 2001). Se uma linha passa por um ponto no casco convexo, não há outra linha com a mesma inclinação passando por outro ponto com uma interceptação positiva (TP) maior. Assim, o classificador nesse ponto é ótimo sob quaisquer suposições de distribuição em conjunto com essa inclinação.

A comunidade de aprendizado de máquina abordou a questão do desequilíbrio de classes de duas maneiras. Uma é atribuir custos distintos aos exemplos de treinamento (Pazzani, Merz, Murphy, Ali, Hume, & Brunk, 1994; Domingos, 1999). A outra é reamostrar o conjunto de dados original, seja por superamostragem da classe minoritária e/ou subamostragem da classe majoritária (Kubat & Matwin, 1997; Japkowicz, 2000; Lewis & Catlett, 1994; Ling & Li, 1998). Nossa abordagem (Chawla, Bowyer, Hall e Kegelmeyer, 2000) combina subamostragem da classe majoritária com uma forma especial de superamostragem da classe minoritária. Experimentos com vários conjuntos de dados e o classificador de árvore de decisão C4.5 (Quinlan, 1992), Ripper (Cohen, 1995b) e um classificador Naive Bayes mostram que nossa abordagem melhora em relação a outras abordagens anteriores de reamostragem, modificação da taxa de perda e priorização de classe ,

A Seção 2 apresenta uma visão geral das medidas de desempenho. A Seção 3 revisa os trabalhos mais relacionados que lidam com conjuntos de dados desequilibrados. A Seção 4 apresenta os detalhes de nossa abordagem. A seção 5 apresenta resultados experimentais comparando nossa abordagem com outras abordagens de reamostragem. A Seção 6 discute os resultados e sugere direções para trabalhos futuros.

2. Medidas de Desempenho

O desempenho dos algoritmos de aprendizado de máquina é normalmente avaliado por uma matriz de confusão, conforme ilustrado na Figura 1 (para um problema de 2 classes). As colunas são a classe Predicted e as linhas são a classe Actual. Na matriz de confusão, TN é o número de exemplos negativos

	Previsto Negativo	Previsto Positivo
Real Negativo	TN	PF
Real Positivo	FN	TP

Figura 1: Matriz de confusão

corretamente classificados (verdadeiros negativos), *PF* é o número de exemplos negativos classificados incorretamente como positivos (Falsos Positivos), *FN* é o número de exemplos positivos incorretamente classificados como negativos (Falsos Negativos) e *TP* é o número de exemplos positivos classificados corretamente (Verdadeiros Positivos).

A precisão preditiva é a medida de desempenho geralmente associada a algoritmos de aprendizado de máquina e é definida como $Precisão = (TP + TN) / (TP + PF + TN + FN)$. No contexto de conjuntos de dados equilibrados e custos de erro iguais, é razoável usar a taxa de erro como uma métrica de desempenho. A taxa de erro é $1 - Precisão$. Na presença de conjuntos de dados desequilibrados com custos de erro desiguais, é mais apropriado usar a curva ROC ou outras técnicas semelhantes (Ling & Li, 1998; Drummond & Holte, 2000; Provost & Fawcett, 2001; Bradley, 1997; Turney, 1996).

As curvas ROC podem ser consideradas como representando a família dos melhores limites de decisão para os custos relativos de TP e FP. Em uma curva ROC, o eixo X representa $\%PF = PF / (TN + PF)$ e o eixo Y representa $\%TP = TP / (TP + FN)$. O ponto ideal na curva ROC seria (0,100), ou seja, todos os exemplos positivos são classificados corretamente e nenhum exemplo negativo é classificado erroneamente como positivo. Uma maneira pela qual uma curva ROC pode ser varrida é manipulando o equilíbrio das amostras de treinamento para cada classe no conjunto de treinamento. A Figura 2 mostra uma ilustração. A linha y = x representa o cenário de adivinhação aleatória da classe. A área sob a curva ROC (AUC) é uma métrica útil para o desempenho do classificador, pois é independente do critério de decisão selecionado e das probabilidades anteriores. A comparação AUC pode estabelecer uma relação de dominância entre classificadores. Se as curvas ROC estiverem se cruzando, a AUC total é uma comparação média entre os modelos (Lee, 2000). No entanto, para algumas distribuições específicas de custo e classe, o classificador com AUC máxima pode, de fato, ser subótimo. Por isso,

3. Trabalho anterior: conjuntos de dados desequilibrados

Kubat e Matwin (1997) subamostraram seletivamente a classe majoritária, mantendo a população original da classe minoritária. Eles usaram a média geométrica como medida de desempenho do classificador, que pode ser relacionada a um único ponto na curva ROC. Os exemplos minoritários foram divididos em quatro categorias: algum ruído sobrepondo a região de decisão de classe positiva, amostras limítrofes, amostras redundantes e amostras seguras. Os exemplos limítrofes foram detectados usando o conceito de links Tomek (Tomek, 1976). Outro

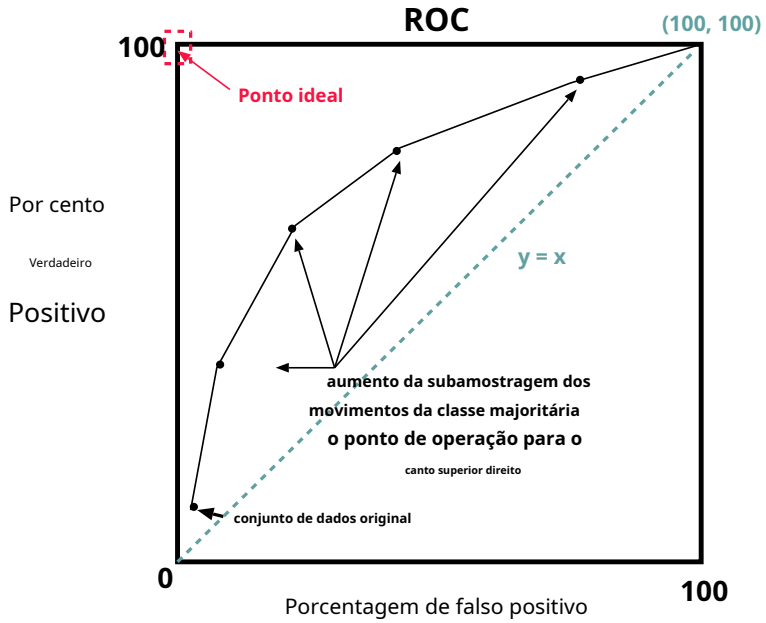

Figura 2: Ilustração da varredura de uma curva ROC por subamostragem. Aumentou a subamostragem da classe majoritária (negativa) moverá o desempenho do ponto inferior esquerdo para o superior direito.

trabalhos relacionados propuseram o sistema SHRINK que classifica uma região sobreposta de classes minoritárias (positivas) e majoritárias (negativas) como positivas; ele procura a "melhor região positiva" (Kubat et al., 1998).

Japkowicz (2000) discutiu o efeito do desequilíbrio em um conjunto de dados. Ela avaliou três estratégias: subamostragem, reamostragem e um esquema de indução baseado em reconhecimento. Nós nos concentramos em suas abordagens de amostragem. Ela experimentou dados 1D artificiais para medir e construir facilmente a complexidade do conceito. Dois métodos de reamostragem foram considerados. A reamostragem aleatória consistia em reamostrar a classe menor aleatoriamente até que ela consistisse em tantas amostras quanto a classe majoritária e a "reamostragem focada" consistia em reamostrar apenas os exemplos minoritários que ocorreram na fronteira entre as classes minoritária e majoritária. Foi considerada a subamostragem aleatória, que envolveu a subamostragem das amostras da classe majoritária aleatoriamente até que seus números correspondessem ao número de amostras da classe minoritária; a sub-amostragem enfocada envolvia a sub-amostragem das amostras da classe majoritária situadas mais distantes. Ela observou que ambas as abordagens de amostragem eram eficazes e também observou que o uso de técnicas de amostragem sofisticadas não dava nenhuma vantagem clara no domínio considerado (Japkowicz, 2000).

Uma abordagem que é particularmente relevante para o nosso trabalho é a de Ling e Li (1998). Eles combinaram a superamostragem da classe minoritária com a subamostragem da classe majoritária. Eles usaram análise de elevação em vez de precisão para medir o desempenho de um classificador. Eles propuseram que os exemplos de teste fossem classificados por uma medida de confiança e então o elevador fosse usado como critério de avaliação. Uma curva de sustentação é semelhante a uma curva ROC, mas é mais adaptada para o

problema de análise de marketing (Ling & Li, 1998). Em um experimento, eles subamostraram a classe majoritária e notaram que o melhor índice de sustentação é obtido quando as classes são igualmente representadas (Ling & Li, 1998). Em outro experimento, eles superamostraram os exemplos positivos (minoritários) com substituição para combinar o número de exemplos negativos (maioria) com o número de exemplos positivos. A combinação de sobre-amostragem e sub-amostragem não proporcionou uma melhoria significativa no índice de elevação. No entanto, nossa abordagem à sobreamostragem difere da deles.

Solberg e Solberg (1996) consideraram o problema de conjuntos de dados desequilibrados na classificação de manchas de óleo a partir de imagens SAR. Eles usaram técnicas de sobreamostragem e subamostragem para melhorar a classificação de manchas de óleo. Seus dados de treinamento tinham uma distribuição de 42 manchas de óleo e 2.471 sósias, dando uma probabilidade anterior de 0,98 para sósias. Esse desequilíbrio levaria o aluno (sem nenhuma função de perda apropriada ou uma metodologia para modificar os anteriores) a classificar quase todos os sósias corretamente às custas de classificar erroneamente muitas das amostras de manchas de óleo (Solberg & Solberg, 1996). Para superar esse problema de desequilíbrio, eles sobre-amostraram (com reposição) 100 amostras da mancha de óleo e amostraram aleatoriamente 100 amostras da classe sem mancha de óleo para criar um novo conjunto de dados com probabilidades iguais. Eles aprenderam uma árvore classificadora neste conjunto de dados balanceado e alcançaram uma taxa de erro de 14% nas manchas de óleo em um método de exclusão para estimativa de erro; nos looklikes obtiveram uma taxa de erro de 4% (Solberg & Solberg, 1996).

Outra abordagem semelhante ao nosso trabalho é a de Domingos (1999). Ele compara a abordagem do "metacusto" a cada subamostragem majoritária e superamostragem minoritária. Ele descobre que o metacusto melhora em relação a ambos, e que a subamostragem é preferível à superamostragem minoritária. Os classificadores baseados em erros são sensíveis ao custo. A probabilidade de cada classe para cada exemplo é estimada e os exemplos são rotulados de forma otimizada em relação aos custos de classificação incorreta. A reetiquetagem dos exemplos expande o espaço de decisão, pois cria novas amostras a partir das quais o classificador pode aprender (Domingos, 1999).

Uma rede neural feed-forward treinada em um conjunto de dados desequilibrado pode não aprender a discriminar o suficiente entre as classes (DeRouin, Brown, Fausett e Schneider, 1991). Os autores propuseram que a taxa de aprendizado da rede neural fosse adaptada às estatísticas de representação de classes nos dados. Eles calcularam um fator de atenção a partir da proporção de amostras apresentadas à rede neural para treinamento. A taxa de aprendizado dos elementos da rede foi ajustada com base no fator atenção. Eles experimentaram em um conjunto de treinamento gerado artificialmente e em um conjunto de treinamento do mundo real, ambos com várias (mais de duas) classes. Eles compararam isso com a abordagem de replicar as amostras da classe minoritária para equilibrar o conjunto de dados usado para treinamento. A precisão da classificação na classe minoritária foi melhorada.

Lewis e Catlett (1994) examinaram amostragem de incerteza heterogênea para aprendizado supervisionado. Esse método é útil para treinar amostras com classes incertas. As amostras de treinamento são rotuladas incrementalmente em duas fases e as instâncias incertas são passadas para a próxima fase. Eles modificaram o C4.5 para incluir um índice de sinistralidade para determinar os valores de classe nas folhas. Os valores de classe foram determinados por comparação com um limite de probabilidade de$LR/(LR+1)$, ondeLRé a sinistralidade (Lewis & Catlett, 1994).

O domínio de recuperação de informação (RI) (Dumais et al., 1998; Mladenić & Grobelnik, 1999; Lewis & Ringuette, 1994; Cohen, 1995a) também enfrenta o problema do desequilíbrio de classes no conjunto de dados. Um documento ou página da Web é convertido em uma representação de pacote de palavras;

ou seja, é construído um vetor de características refletindo ocorrências de palavras na página. Normalmente, há muito poucas instâncias da categoria interessante na categorização de texto. Essa super-representação da classe negativa em problemas de recuperação de informações pode causar problemas na avaliação do desempenho dos classificadores. Como a taxa de erro não é uma boa métrica para conjuntos de dados assimétricos, o desempenho de classificação de algoritmos na recuperação de informações geralmente é medido por *precisão* e *lembrar*:

$$lembrar = \frac{TP}{TP + FN}$$

$$precisão = \frac{TP}{TP + PF}$$

Mladenić e Grobelnik (1999) propuseram uma abordagem de seleção de subconjunto de características para lidar com distribuição de classes desequilibrada no domínio IR. Eles experimentaram vários métodos de seleção de recursos e descobriram que o *razão de probabilidade* (van Rijsbergen, Harper, & Porter, 1981) quando combinado com um classificador Naive Bayes tem o melhor desempenho em seu domínio. *Razão de probabilidade* é uma medida probabilística usada para classificar documentos de acordo com sua relevância para a classe positiva (classe minoritária). *Ganho de informação* pois uma palavra, por outro lado, não presta atenção a uma determinada classe-alvo; é calculado por palavra para cada classe. Em um conjunto de dados de texto desequilibrado (assumindo que 98 a 99% é a classe negativa), a maioria dos recursos será associada à classe negativa. *Razão de probabilidade* incorpora as informações da classe alvo em sua métrica dando melhores resultados quando comparado com *ganho de informação* para categorização de texto.

Provost e Fawcett (1997) introduziram o método ROC convex hull para estimar o desempenho do classificador para conjuntos de dados desbalanceados. Eles observam que os problemas de distribuição de classe desigual e custos de erro desiguais estão relacionados e que pouco trabalho foi feito para resolver qualquer problema (Provost & Fawcett, 2001). No método de casco convexo ROC, o espaço ROC é usado para separar o desempenho de classificação das informações de distribuição de classe e custo.

Para resumir a literatura, a subamostragem da classe majoritária permite a construção de classificadores melhores do que a superamostragem da classe minoritária. Uma combinação dos dois como feito em trabalhos anteriores não leva a classificadores que superam aqueles construídos utilizando apenas subamostragem. No entanto, a sobreamostragem da classe minoritária foi feita por amostragem com substituição dos dados originais. Nossa abordagem usa um método diferente de sobreamostragem.

4. SMOTE: Técnica de sobreamostragem de minoria sintética

4.1 Sobreamostragem minoritária com substituição

Pesquisas anteriores (Ling & Li, 1998; Japkowicz, 2000) discutiram sobre-amostragem com substituição e notaram que isso não melhora significativamente o reconhecimento das classes minoritárias. Interpretamos o efeito subjacente em termos de regiões de decisão no espaço de características. Essencialmente, como a classe minoritária é superamostrada por quantidades crescentes, o efeito é identificar regiões semelhantes, porém mais específicas, no espaço de características como a região de decisão para a classe minoritária. Este efeito para as árvores de decisão pode ser entendido a partir dos gráficos na Figura 3.

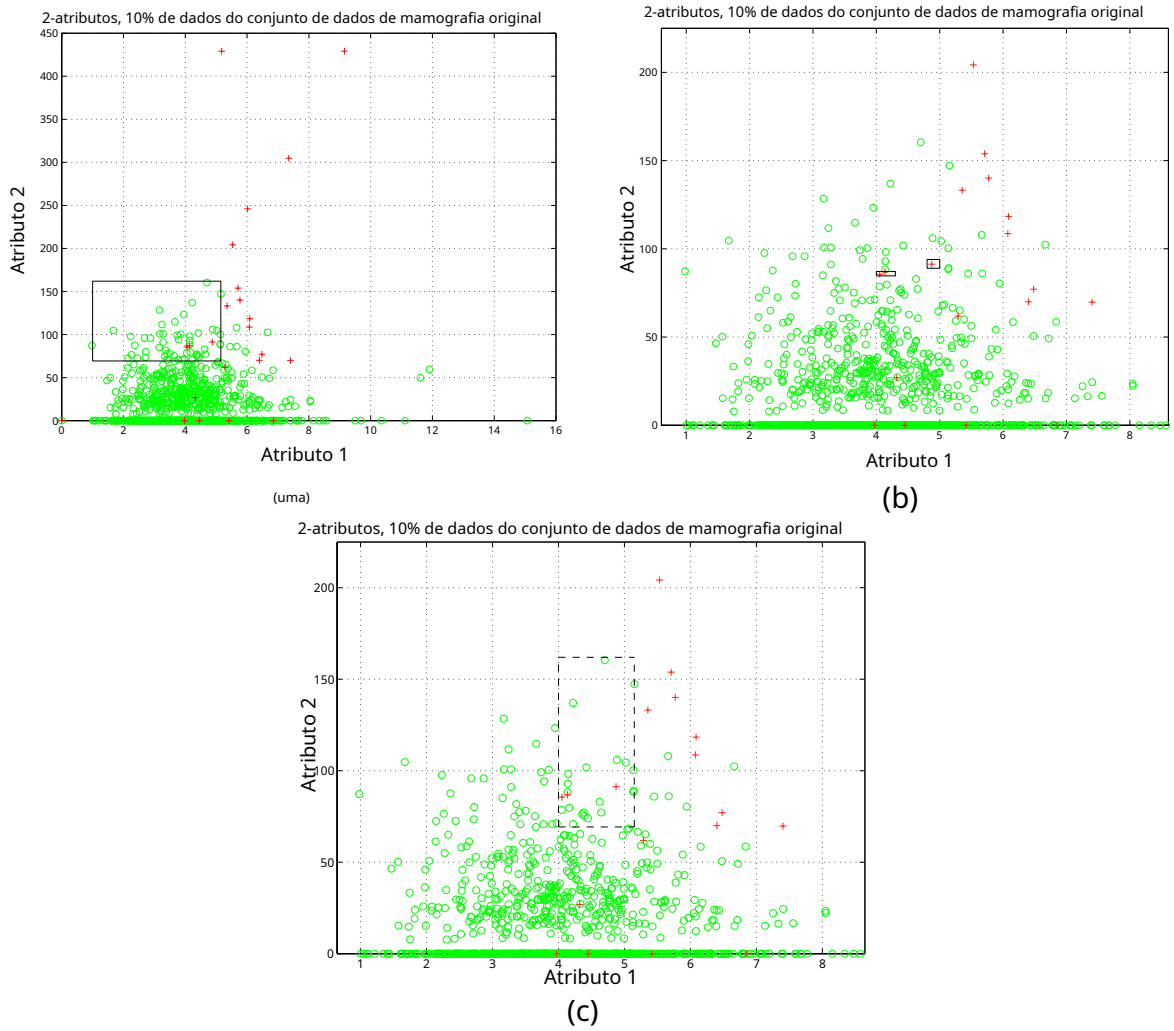

Figura 3: a) Região de decisão na qual residem as três amostras de classe minoritária (mostradas por '+') depois de construir uma árvore de decisão. Esta região de decisão é indicada pelo retângulo de linha sólida. b) Uma visão ampliada das amostras de classe minoritária escolhidas para o mesmo conjunto de dados. Pequenos retângulos de linha sólida mostram as regiões de decisão como resultado da superamostragem da classe minoritária com replicação. c) Uma visão ampliada das amostras de classe minoritária escolhidas para o mesmo conjunto de dados. As linhas tracejadas mostram a região de decisão após superamostragem da classe minoritária com geração sintética.

Os dados para o gráfico na Figura 3 foram extraídos de um conjunto de dados de mamografia[1](Woods et ai., 1993). As amostras de classe minoritária são mostradas por + e as amostras de classe majoritária são mostradas por*o*na trama. Na Figura 3(a), a região indicada pelo retângulo de linha contínua é uma região de decisão de classe majoritária. No entanto, contém três amostras de classe minoritária mostradas por '+' como falsos negativos. Se replicarmos a classe minoritária, a região de decisão para a classe minoritária se torna muito específica e causará novas divisões na árvore de decisão. Isso levará a mais nós terminais (folhas) à medida que o algoritmo de aprendizado tenta aprender cada vez mais regiões específicas da classe minoritária; em essência, overfitting. A replicação da classe minoritária não faz com que seu limite de decisão se espalhe para a região da classe majoritária. Assim, na Figura 3(b), as três amostras anteriormente na região de decisão da classe majoritária agora possuem regiões de decisão muito específicas.

4.2 SMOTE

Propomos uma abordagem de superamostragem na qual a classe minoritária é superamostrada criando exemplos “sintéticos” em vez de superamostragem com substituição. Essa abordagem é inspirada em uma técnica que provou ser bem-sucedida no reconhecimento de caracteres manuscritos (Ha & Bunke, 1997). Eles criaram dados extras de treinamento realizando certas operações em dados reais. No caso deles, operações como rotação e inclinação eram formas naturais de perturbar os dados de treinamento. Geramos exemplos sintéticos de uma maneira menos específica do aplicativo, operando em “espaço de recursos” em vez de “espaço de dados”. A classe minoritária é superamostrada tomando cada amostra de classe minoritária e introduzindo exemplos sintéticos ao longo dos segmentos de linha que unem qualquer/todos os*k*vizinhos mais próximos da classe minoritária. Dependendo da quantidade de sobreamostragem necessária, os vizinhos do*k*vizinhos mais próximos são escolhidos aleatoriamente. Nossa implementação atualmente usa cinco vizinhos mais próximos. Por exemplo, se a quantidade de sobreamostragem necessária for 200%, apenas dois vizinhos dos cinco vizinhos mais próximos são escolhidos e uma amostra é gerada na direção de cada um. Amostras sintéticas são geradas da seguinte maneira: Tome a diferença entre o vetor de características (amostra) em consideração e seu vizinho mais próximo. Multiplique essa diferença por um número aleatório entre 0 e 1 e adicione-o ao vetor de recursos em consideração. Isso causa a seleção de um ponto aleatório ao longo do segmento de linha entre dois recursos específicos. Essa abordagem efetivamente força a região de decisão da classe minoritária a se tornar mais geral.

Algoritmo*SMOTE*, na próxima página, é o pseudocódigo para SMOTE. A Tabela 4.2 mostra um exemplo de cálculo de amostras sintéticas aleatórias. A quantidade de sobreamostragem é um parâmetro do sistema, e uma série de curvas ROC podem ser geradas para diferentes populações e análises ROC realizadas.

Os exemplos sintéticos fazem com que o classificador crie regiões de decisão maiores e menos específicas, conforme mostrado pelas linhas tracejadas na Figura 3(c), em vez de regiões menores e mais específicas. Regiões mais gerais agora são aprendidas para as amostras de classes minoritárias, em vez daquelas que estão sendo incluídas nas amostras de classes majoritárias ao seu redor. O efeito é que as árvores de decisão generalizam melhor. As Figuras 4 e 5 comparam a sobreamostragem minoritária com reposição e SMOTE. Os experimentos foram conduzidos no conjunto de dados de mamografia. Havia 10.923 exemplos na classe majoritária e 260 exemplos na classe minoritária originalmente. Temos aproximadamente 9831 exemplos na classe majoritária e 233 exemplos

1. Os dados estão disponíveis no USF Intelligent Systems Lab, http://morden.csee.usf.edu/~chawla.

na classe minoritária para o conjunto de treinamento usado na validação cruzada de 10 vezes. A classe minoritária foi superamostrada em 100%, 200%, 300%, 400% e 500% de seu tamanho original. Os gráficos mostram que os tamanhos das árvores para superamostragem minoritária com substituição em graus mais altos de replicação são muito maiores do que para SMOTE, e o reconhecimento de classe minoritária da técnica de superamostragem minoritária com substituição em graus mais altos de replicação não é tão bom como SMOTE.

Algoritmo *SMOTE*(T, N, k)
Entrada: Número de amostras de classe minoritária *T*; Quantidade de SMOTE *N*%; Número de mais próximos vizinhos *k*
Resultado: (*N*/100) * *T* amostras sintéticas de classe minoritária
1. (**Se N for inferior a 100%, aleatorize as amostras da classe minoritária, pois apenas uma porcentagem aleatória delas será SMOTEd.**)
2. **E se** *N* <100
3. **então** Randomizar o *T* amostras de classe minoritária
4. *T* = (*N*/100) * *TN* =
5. 100
6. **fim se**
7. *N* = (*int*) (*N*/100) (**A quantidade de SMOTE é assumida em múltiplos integrais de 100.**)
8. *k* = Número de vizinhos mais próximos
9. *numattrs* = Número de atributos
10. *Amostra*[ ][ ]: array para amostras originais de classe minoritária
11. *novo índice*: mantém uma contagem do número de amostras sintéticas geradas, inicializadas em 0
12. *Sintético*[ ][ ]: array para amostras sintéticas
(**Calcular k vizinhos mais próximos apenas para cada amostra de classe minoritária.**)
13. **por** *eu* ← 1 **para** *T*
14. Calcular *k* vizinhos mais próximos para *eu*, e salve os índices no *narrar*
15. Preencher(*N*, *eu*, *narrar*)
16. **fim**

Preencher(*N, eu, narrar*) (**Função para gerar as amostras sintéticas.**)
17. **enquanto** *N* =0
18. Escolha um número aleatório entre 1 e *k*, chame-o *nn*. Esta etapa escolhe um dos *k* vizinhos mais próximos de *eu*. **por** *atr* ← 1 **para** *numattrs*
19.
20. Calcular: *dif* = *Amostra*[*narrar*[*nn*]][*atr*] − *Amostra*[*eu*][*atr*]
21. Calcular: *Gap = Vão* = número aleatório entre 0 e 1 *Sintético*[
22. *novo índice*][*atr*] = *Amostra*[*eu*][*atr*] + *Gap = Vão* * *dif*
23. **fim**
24. *novo índice*++
25. *N* = *N* -1
26. **sem fim**
27. **Retorna** (**Fim de povoar.**) Fim do Pseudo-Código.

Considere uma amostra (6,4) e seja (4,3) seu vizinho mais próximo. (6,4) é a amostra para a qual k-vizinhos mais próximos estão sendo identificados. (4,3) é um de seus k-vizinhos mais próximos. Deixar: f1_1 = 6 f2_1 = 4 f2_1 - f1_1 = -2 f1_2 = 4 f2_2 = 3 f2_2 - f1_2 = -1 As novas amostras serão geradas como (f1',f2') = (6,4) + rand(0-1) * (-2,-1) rand(0-1) gera um número aleatório entre 0 e 1.

Tabela 1: Exemplo de geração de exemplos sintéticos (SMOTE).

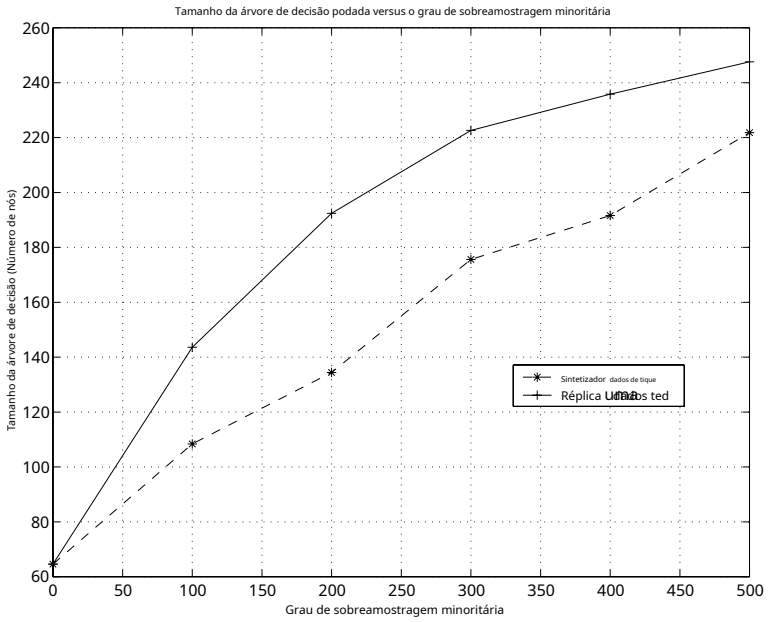

Figura 4: Comparação dos tamanhos das árvores de decisão para superamostragem replicada e SMOTE para o conjunto de dados de mamografia

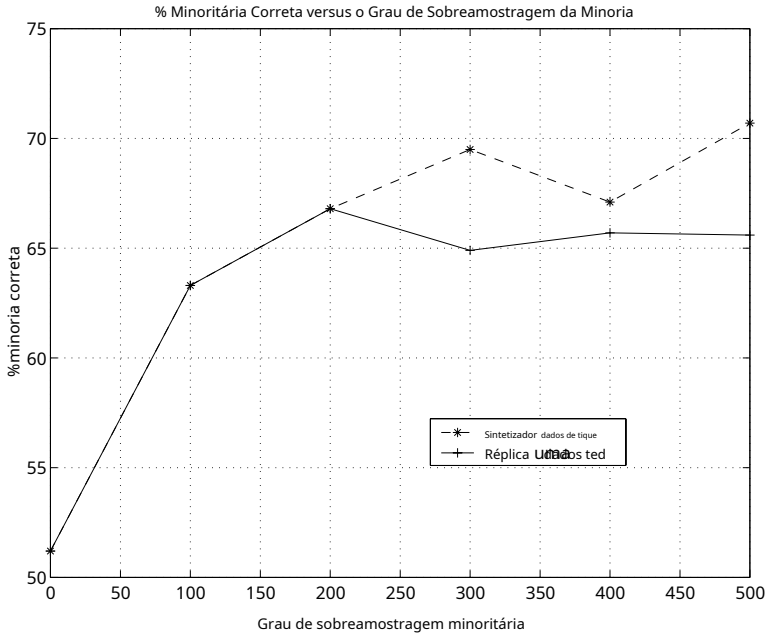

Figura 5: Comparação da % Minoria correta para superamostragem replicada e SMOTE para o conjunto de dados de mamografia

4.3 Subamostragem e Combinação SMOTE

A classe majoritária é sub-amostrada pela remoção aleatória de amostras da população da classe majoritária até que a classe minoritária se torne uma porcentagem especificada da classe majoritária. Isso força o aluno a experimentar vários graus de subamostragem e, em graus mais altos de subamostragem, a classe minoritária tem uma presença maior no conjunto de treinamento. Ao descrever nossos experimentos, nossa terminologia será tal que, se*subamostra a classe majoritária em 200%*, isso significaria que o conjunto de dados modificado conterá*duas vezes mais elementos da classe minoritária do que da classe majoritária*; isto é, se a classe minoritária tivesse 50 amostras e a classe majoritária tivesse 200 amostras e nós subamostramos a maioria em 200%, a classe majoritária acabaria tendo 25 amostras. Ao aplicar uma combinação de sub-amostragem e sobre-amostragem, o viés inicial do aluno para a classe negativa (maioria) é revertido em favor da classe positiva (minoria). Os classificadores são aprendidos no conjunto de dados perturbados por “SMOTING” da classe minoritária e subamostragem da classe majoritária.

5. Experimentos

Usamos três algoritmos de aprendizado de máquina diferentes para nossos experimentos. A Figura 6 fornece uma visão geral de nossos experimentos.

1. **C4.5:** Comparamos várias combinações de SMOTE e subamostragem com subamostragem simples usando C4.5 versão 8 (Quinlan, 1992) como classificador base.

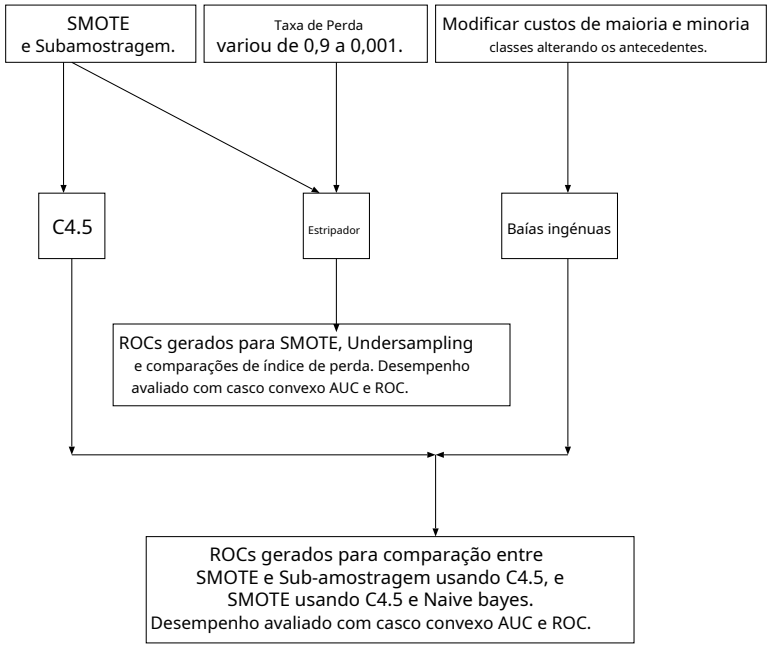

Figura 6: Visão geral dos experimentos

2.**Estripador:**Comparamos várias combinações de SMOTE e subamostragem com subamostragem simples usando Ripper (Cohen, 1995b) como classificador base. Também variamos a sinistralidade de Ripper (Cohen & Singer, 1996; Lewis & Catlett, 1994) de 0,9 a 0,001 (como forma de variar o custo de classificação incorreta) e comparamos o efeito dessa variação com a combinação de SMOTE e subamostragem. Ao reduzir a sinistralidade de 0,9 para 0,001 conseguimos construir um conjunto de regras para a classe minoritária.

3.**Classificador Naive Bayes:**O classificador Naive Bayes[2]pode tornar-se sensível ao custo variando os antecedentes da classe minoritária. Variamos os antecedentes da classe minoritária de 1 a 50 vezes a classe majoritária e comparamos com a combinação SMOTE e subamostragem do C4.5.

Esses diferentes algoritmos de aprendizado permitiram que o SMOTE fosse comparado a alguns métodos que podem lidar diretamente com os custos de classificação incorreta. A média de %FP e %TP foi calculada em 10 vezes de execuções de validação cruzada para cada uma das combinações de dados. Os exemplos de classe minoritária foram superamostrados calculando os cinco vizinhos mais próximos e gerando exemplos sintéticos. A AUC foi calculada usando a regra trapezoidal. Extrapolamos um ponto extra de TP = 100% e FP = 100% para cada curva ROC. Também calculamos o casco convexo ROC para identificar os classificadores ótimos, pois os pontos situados no casco são classificadores potencialmente ótimos (Provost & Fawcett, 2001).

2. O código fonte foi baixado de http://fuzzy.cs.uni-magdeburg.de/~borgelt/software.html.

5.1 Conjuntos de dados

Experimentamos em nove conjuntos de dados diferentes. Esses conjuntos de dados estão resumidos na Tabela 5.2. Esses conjuntos de dados variam amplamente em tamanho e proporções de classe, oferecendo diferentes domínios para o SMOTE. Em ordem crescente de desequilíbrio, eles são:

1. O Pima Indian Diabetes (Blake & Merz, 1998) possui 2 classes e 768 amostras. Os dados são usados para identificar os casos positivos de diabetes em uma população perto de Phoenix, Arizona. O número de amostras de classe positivas é de apenas 268. Boa sensibilidade para detecção de casos de diabetes será um atributo desejável do classificador.

2. O conjunto de dados Phoneme é do projeto ELENA[3]. O objetivo do conjunto de dados é distinguir entre sons nasais (classe 0) e orais (classe 1). Existem 5 funcionalidades. A distribuição de classes é de 3.818 amostras na classe 0 e 1.586 amostras na classe 1.

3. O conjunto de dados Adult (Blake & Merz, 1998) possui 48.842 amostras com 11.687 amostras pertencentes à classe minoritária. Este conjunto de dados tem 6 características contínuas e 8 características nominais. Os algoritmos SMOTE e SMOTE-NC (consulte a Seção 6.1) foram avaliados neste conjunto de dados. Para o SMOTE, extraímos os recursos contínuos e geramos um novo conjunto de dados apenas com recursos contínuos.

4. Os dados do estado E[4](Hall, Mohney, & Kier, 1991) consiste em descritores de estado eletrotopológico para uma série de compostos da triagem de drogas anti-câncer do Instituto Nacional do Câncer. Descritores de estado E do NCI Yeast AntiCancer Drug Screen foram gerados por Tripos, Inc. Resumidamente, uma série de cerca de 60.000 compostos foi testada contra uma série de 6 cepas de levedura em uma determinada concentração. O teste foi uma tela de alto rendimento em apenas uma concentração, de modo que os resultados estão sujeitos a contaminação, etc. A inibição do crescimento da cepa de levedura quando exposta ao determinado composto (em relação ao crescimento da levedura em um solvente neutro) foi medida . As classes de atividade são ativas — pelo menos uma única cepa de levedura foi inibida em mais de 70%, ou inativa — nenhuma cepa de levedura foi inibida em mais de 70%. O conjunto de dados tem 53.220 amostras com 6,

5. O conjunto de dados Satimage (Blake & Merz, 1998) possui originalmente 6 classes. Escolhemos a menor classe como a classe minoritária e colapsamos o resto das classes em uma, como foi feito em (Provost et al., 1998). Isso nos deu um conjunto de dados de 2 classes distorcidas, com 5.809 amostras de classe majoritária e 626 amostras de classe minoritária.

6. O conjunto de dados Forest Cover é do repositório UCI (Blake & Merz, 1998). Este conjunto de dados tem 7 classes e 581.012 amostras. Este conjunto de dados é para a previsão do tipo de cobertura florestal com base em variáveis cartográficas. Como nosso sistema atualmente funciona para classes binárias, extraímos dados para duas classes desse conjunto de dados e ignoramos o resto. A maioria das outras abordagens só funciona para apenas duas classes (Ling & Li, 1998; Japkowicz, 2000; Kubat & Matwin, 1997; Provost & Fawcett, 2001). As duas classes que consideramos são Pinho Ponderosa com 35.754 amostras e Cottonwood/Willow com 2.747

3. ftp.dice.ucl.ac.be no diretório pub/neural-nets/ELENA/databases.

4. Gostaríamos de agradecer a Steven Eschrich por nos fornecer o conjunto de dados e a descrição.

Conjunto de dados	Classe majoritária	Classe Minoritária
Pima	500	268
Fonema	3818	1586
Adulto	37155	11687
Estado	46869	6351
Satimagem	5809	626
Cobertura florestal	35754	2747
Óleo	896	41
Mamografia	10923	260
Posso	435512	8360

Tabela 2: Distribuição do conjunto de dados

amostras. No entanto, a técnica SMOTE pode ser aplicada a um problema de múltiplas classes, especificando para qual classe SMOTE. No entanto, neste artigo, focamos em problemas de 2 classes, para representar explicitamente classes positivas e negativas.

7. O conjunto de dados Oil foi fornecido por Robert Holte e é usado em seu artigo (Kubat et al., 1998). Este conjunto de dados tem 41 amostras de manchas de óleo e 896 amostras de não manchas de óleo.

8. O conjunto de dados de mamografia (Woods et al., 1993) tem 11.183 amostras com 260 calcificações. Se olharmos para a precisão preditiva como uma medida de qualidade do classificador para este caso, a precisão padrão seria de 97,68% quando cada amostra é rotulada como não calcificada. Mas, é desejável que o classificador preveja a maioria das calcificações corretamente.

9. O conjunto de dados Can foi gerado a partir dos dados Can ExodusII usando a versão AVATAR (Chawla & Hall, 1999) da ferramenta Mustafa Visualization[5]. A parte da lata sendo esmagada foi marcada como “muito interessante” e o restante da lata foi marcado como “desconhecido”. Um conjunto de dados de tamanho 443.872 amostras com 8.360 amostras marcadas como “muito interessantes” foi gerado.

5.2 Criação de ROC

Uma curva ROC para SMOTE é produzida usando C4.5 ou Ripper para criar um classificador para cada um de uma série de conjuntos de dados de treinamento modificados. Uma determinada curva ROC é produzida pela primeira superamostragem da classe minoritária em um grau especificado e, em seguida, subamostragem da classe majoritária em graus crescentes para gerar os pontos sucessivos na curva. A quantidade de subamostragem é idêntica à subamostragem simples. Assim, cada ponto correspondente em cada curva ROC para um conjunto de dados representa o mesmo número de amostras de classe majoritária. Diferentes curvas ROC são produzidas começando com diferentes níveis de sobreamostragem minoritária. As curvas ROC também foram geradas variando a taxa de perda no Ripper de 0,9 a 0,001 e variando as prioris da classe minoritária da distribuição original para até 50 vezes a classe majoritária para um classificador Naive Bayes.

5. A ferramenta de visualização Mustafa foi desenvolvida por Mike Glass do Sandia National Labs.

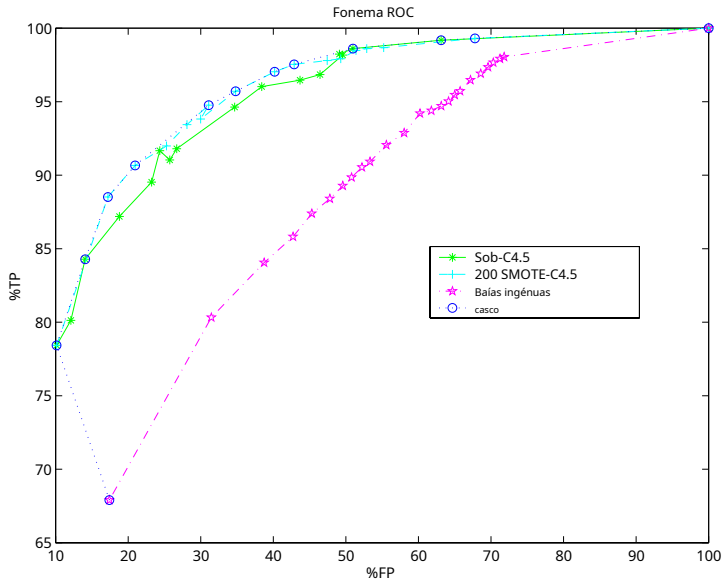

Figura 7: Fonema. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. SMOTE-C4.5 domina sobre Naive Bayes e Under-C4.5 no espaço ROC. Classificadores SMOTE-C4.5 são classificadores potencialmente ótimos.

As Figuras 9 a 23 mostram as curvas ROC experimentais obtidas para os nove conjuntos de dados com os três classificadores. A curva ROC para subamostragem simples da classe majoritária (Ling & Li, 1998; Japkowicz, 2000; Kubat & Matwin, 1997; Provost & Fawcett, 2001) é comparada com nossa abordagem de combinar superamostragem sintética da classe minoritária (SMOTE) com subamostragem de classe majoritária. A curva de subamostragem simples é rotulada como “Sob”, e a curva ROC de combinação SMOTE e subamostragem é rotulada como “SMOTE”. Dependendo do tamanho e do desequilíbrio relativo do conjunto de dados, são criadas de uma a cinco curvas SMOTE e de subamostragem. Mostramos apenas os melhores resultados do SMOTE combinados com sub-amostragem e a curva de sub-amostragem simples nos gráficos. A curva SMOTE ROC de C4. 5 também é comparada com a curva ROC obtida a partir da variação das prioris da classe minoritária usando um classificador Naive Bayes — rotulado como “Naive Bayes”. As curvas ROC “SMOTE”, “Under” e “Loss Ratio”, geradas usando o Ripper, também são comparadas. Para uma dada família de curvas ROC, é gerado um casco convexo ROC (Provost & Fawcett, 2001). O casco convexo ROC é gerado usando o algoritmo de Graham (O'Rourke, 1998). Para referência, mostramos a curva ROC que seria obtida usando a sobreamostragem minoritária por replicação na Figura 19. O casco convexo ROC é gerado usando o algoritmo de Graham (O'Rourke, 1998). Para referência, mostramos a curva ROC que seria obtida usando a sobreamostragem minoritária por replicação na Figura 19. O casco convexo ROC é gerado usando o algoritmo de Graham (O'Rourke, 1998). Para referência, mostramos a curva ROC que seria obtida usando a sobreamostragem minoritária por replicação na Figura 19.

Cada ponto na curva ROC é o resultado de um classificador (C4.5 ou Ripper) aprendido para uma combinação particular de subamostragem e SMOTE, um classificador (C4.5 ou Ripper) aprendido com subamostragem simples ou um classificador (Ripper) aprendido usando alguma taxa de perda ou um classificador (Naive Bayes) aprendido para um prior diferente para a classe minoritária. Cada ponto representa o resultado médio (%TP e %FP) da validação cruzada 10 vezes. O ponto inferior esquerdo para uma determinada curva ROC é do conjunto de dados brutos, sem qualquer subclasse de classe majoritária.

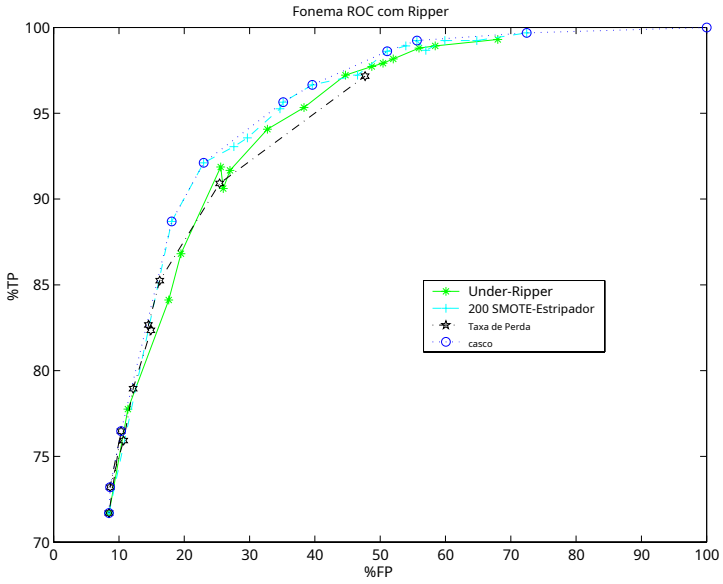

Figura 8: Fonema. Comparação de SMOTE-Ripper, Under-Ripper e perda de modificação Relação no Estripador. SMOTE-Ripper domina sobre Under-Ripper e Loss Ratio no espaço ROC. Mais classificadores SMOTE-Ripper estão no casco convexo do ROC.

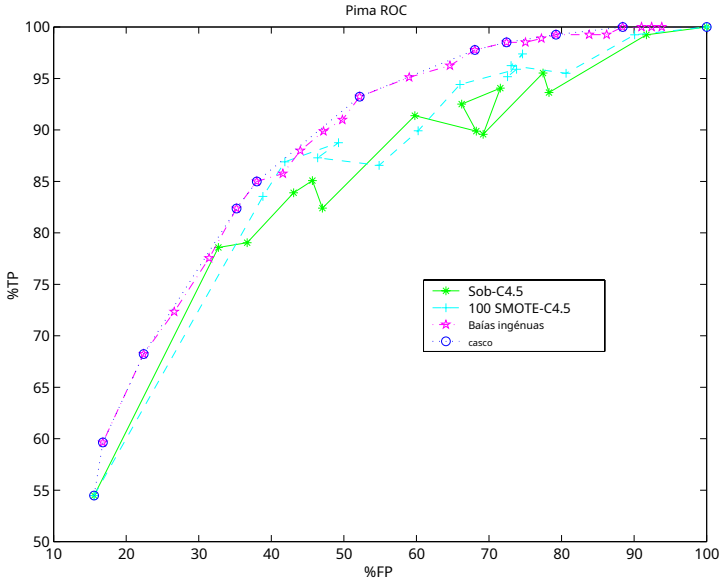

Figura 9: Diabetes dos índios Pima. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. Naive Bayes domina o SMOTE-C4.5 no espaço ROC.

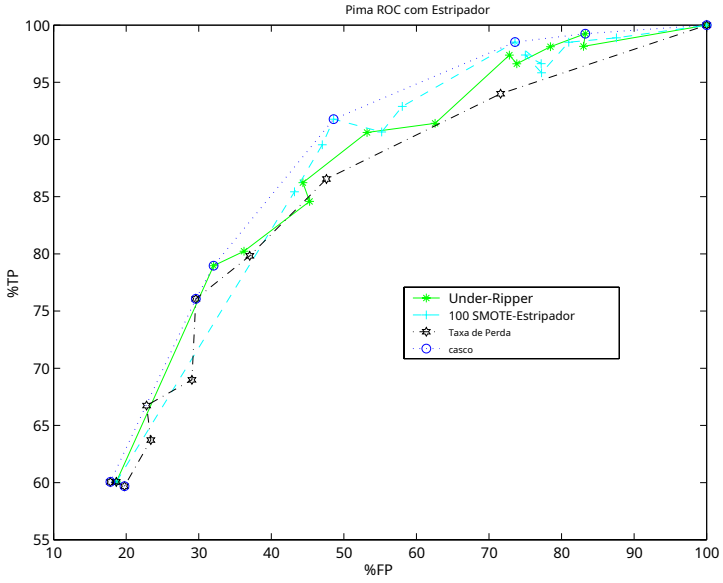

Figura 10: Diabetes dos índios Pima. Comparação de SMOTE-Ripper, Under-Ripper e modificando a Taxa de Perda no Estripador. SMOTE-Ripper domina sobre Under-Ripper e Loss Ratio no espaço ROC.

amostragem ou superamostragem de classe minoritária. A classe minoritária foi superamostrada em 50%, 100%, 200%, 300%, 400%, 500%. A classe majoritária foi sub-amostrada em 10%, 15%, 25%, 50%, 75%, 100%, 125%, 150%, 175%, 200%, 300%, 400%, 500%, 600%, 700%, 800%, 1000% e 2000%. A quantidade de subamostragem de classe majoritária e superamostragem de classe minoritária dependia do tamanho do conjunto de dados e das proporções de classe. Por exemplo, considere as curvas ROC na Figura 17 para o conjunto de dados de mamografia. Existem três curvas - uma para subamostragem de classe majoritária simples na qual a faixa de subamostragem varia entre 5% e 2000% em intervalos diferentes, uma para uma combinação de subamostragem SMOTE e classe majoritária e uma para Naive Bayes — e uma curva de casco convexa ROC. A curva ROC mostrada na Figura 17 é para a classe minoritária sobreamostrada em 400%. Cada ponto nas curvas SMOTE ROC representa uma combinação de superamostragem (sintética) e subamostragem, a quantidade de subamostragem segue a mesma faixa da subamostragem simples. Para uma melhor compreensão dos gráficos ROC, mostramos diferentes conjuntos de curvas ROC para um de nossos conjuntos de dados no Apêndice A.

Para o conjunto de dados Can, tivemos que SMOTE em menor grau do que para os outros conjuntos de dados devido à natureza estrutural do conjunto de dados. Para o conjunto de dados Can existe uma vizinhança estrutural já estabelecida na geometria da malha, então o SMOTE pode levar à criação de vizinhos que estão sob a superfície (e, portanto, não são interessantes), já que estamos olhando para o espaço de características das variáveis físicas e não para as variáveis estruturais. em formação.

As curvas ROC mostram uma tendência de que, à medida que aumentamos a quantidade de subamostragem associada à sobreamostragem, nossa precisão de classificação minoritária aumenta, é claro, às custas de mais erros de classe majoritária. Para quase todas as curvas ROC, a abordagem SMOTE

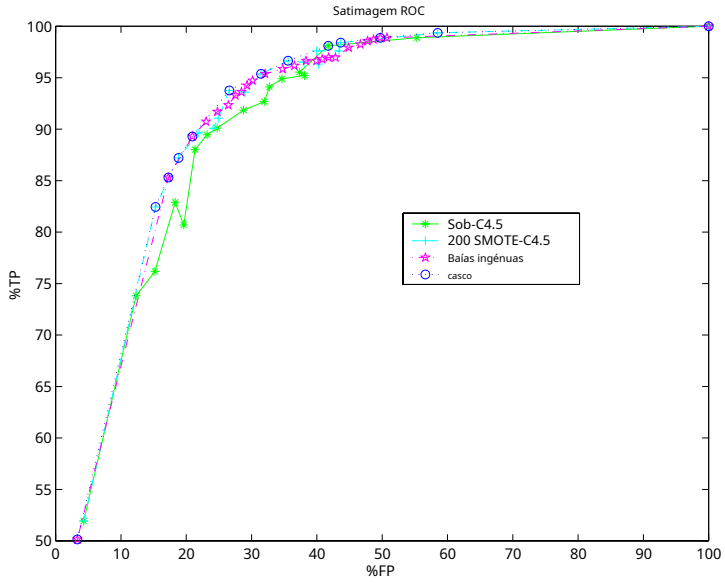

Figura 11: Satimagem. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. o As curvas ROC de Naive Bayes e SMOTE-C4.5 mostram uma sobreposição; no entanto, em TPs mais altos, mais pontos do SMOTE-C4.5 estão no casco convexo do ROC.

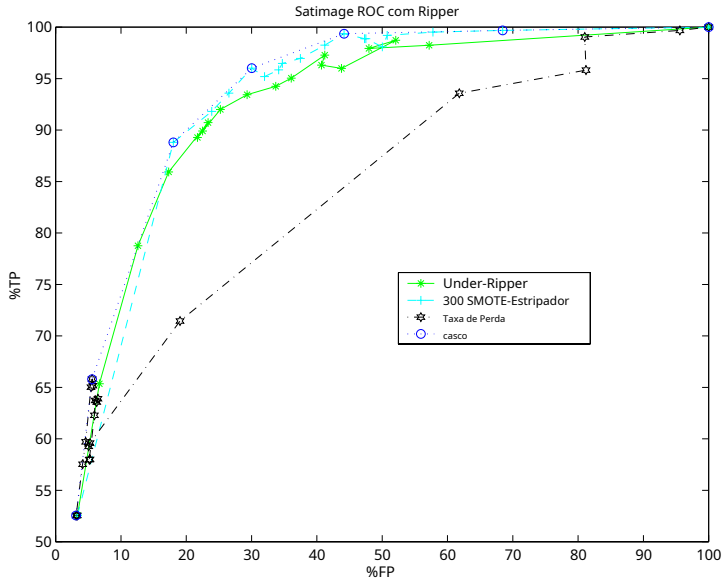

Figura 12: Satimagem. Comparação de SMOTE-Ripper, Under-Ripper e perda de modificação Relação no Estripador. SMOTE-Ripper domina o espaço ROC. O casco convexo ROC é construído principalmente com pontas de SMOTE-Ripper.

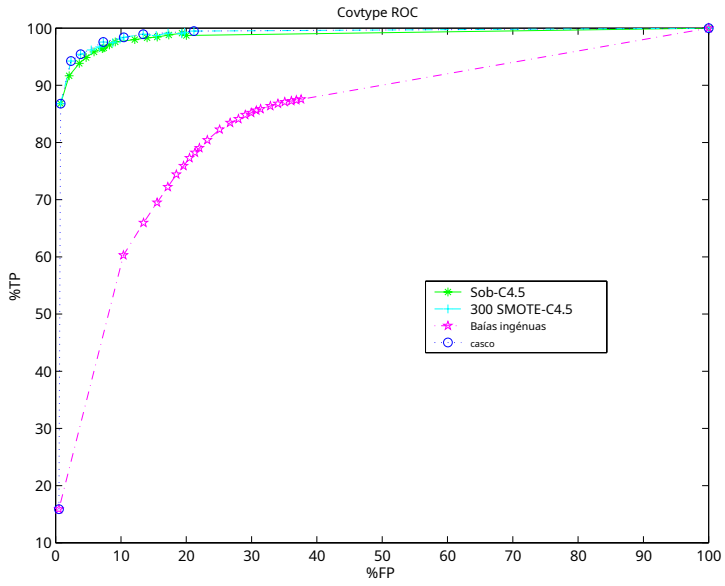

Figura 13: Cobertura Florestal. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. As curvas ROC SMOTE-C4.5 e Under-C4.5 são muito próximas umas das outras. No entanto, mais pontos da curva ROC SMOTE-C4.5 estão no casco convexo ROC, estabelecendo assim uma dominância.

inatos. Aderindo à definição de casco convexo ROC, a maioria dos classificadores potencialmente ótimos são aqueles gerados com SMOTE.

5.3 Cálculo da AUC

A área sob a curva ROC (AUC) é calculada usando uma forma da regra do trapézio. O ponto inferior esquerdo para uma determinada curva ROC é o desempenho de um classificador nos dados brutos. O ponto superior direito é sempre (100%, 100%). Se a curva não terminar naturalmente neste ponto, o ponto é adicionado. Isso é necessário para que as AUCs sejam comparadas no mesmo intervalo de %FP.

As AUCs listadas na Tabela 5.3 mostram que, para todos os conjuntos de dados, a sobreamostragem de minoria sintética combinada e a sobreamostragem de maioria são capazes de melhorar em relação à subamostragem de maioria simples com C4.5 como classificador de base. Assim, nossa abordagem SMOTE proporciona uma melhoria na classificação correta dos dados na classe sub-representada. A mesma conclusão vale a partir de um exame dos cascos convexos ROC. Algumas das entradas estão faltando na tabela, pois o SMOTE não foi aplicado nas mesmas quantidades a todos os conjuntos de dados. A quantidade de SMOTE foi menor para conjuntos de dados menos distorcidos. Além disso, não incluímos AUCs para Ripper/Naive Bayes. O casco convexo ROC identifica classificadores SMOTE como potencialmente ótimos em comparação com subamostragem simples ou outros tratamentos de custos de classificação incorreta, em geral. As exceções são as seguintes: para o conjunto de dados Pima, Naive Bayes domina sobre SMOTE-C4.5; para o conjunto de dados de petróleo, Under-Ripper domina sobre SMOTE-Ripper. Para o conjunto de dados Can, SMOTE-*classificador*(*classificador*=C4.5 ou Estripador) e*classificador*ROC

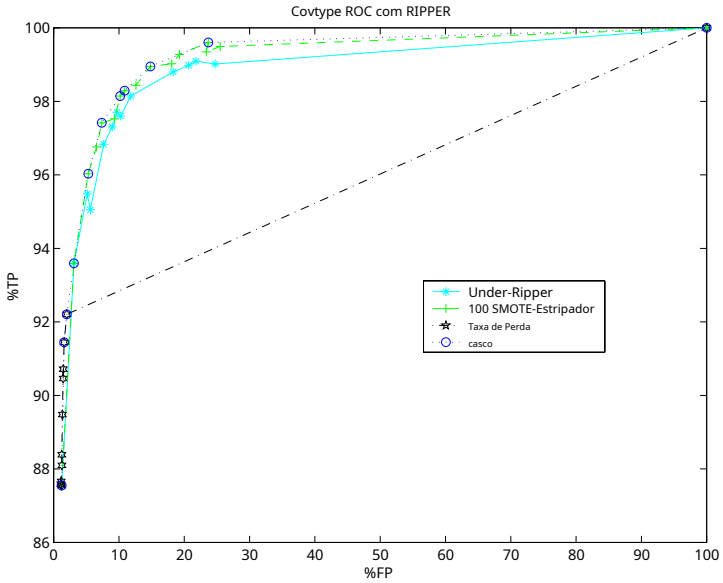

Figura 14: Cobertura Florestal. Comparação de SMOTE-Ripper, Under-Ripper e modificação Taxa de perda no Ripper. SMOTE-Ripper mostra uma dominação no espaço ROC. Mais pontos da curva SMOTE-Ripper estão no casco convexo ROC.

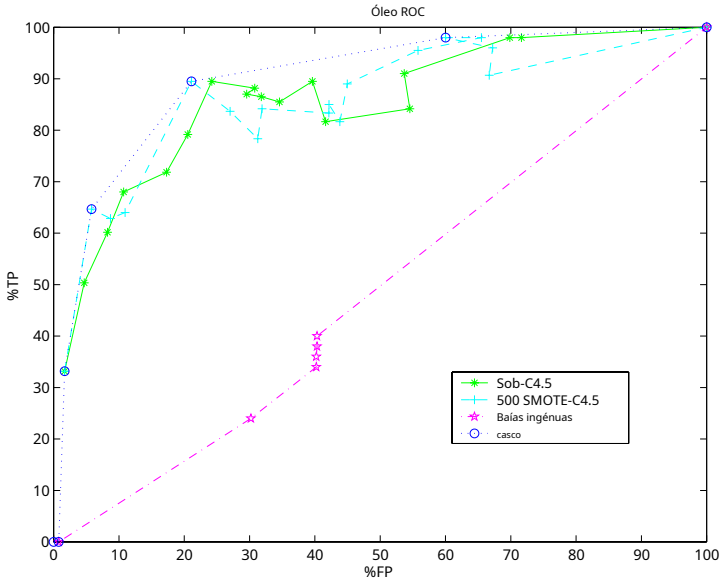

Figura 15: Óleo. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. Embora, As curvas SMOTE-C4.5 e Under-C4.5 ROC se cruzam em pontos, mais pontos da curva SMOTE-C4.5 estão no casco convexo ROC.

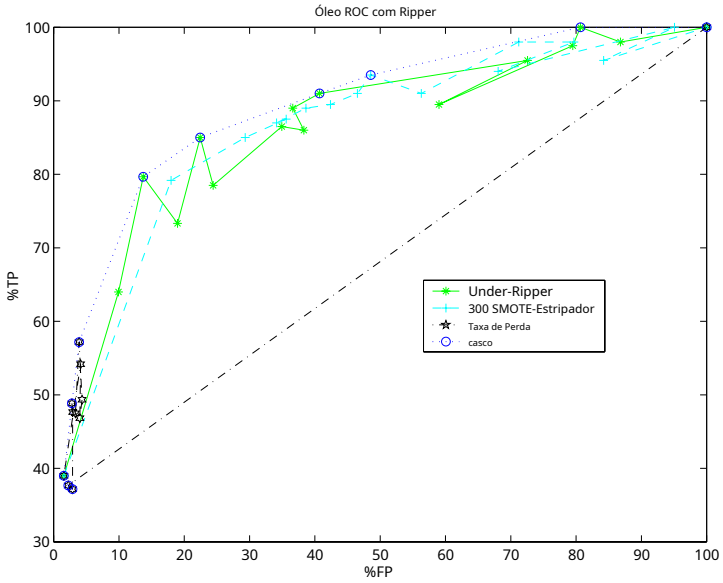

Figura 16: Óleo. Comparação de SMOTE-Ripper, Under-Ripper e modificação da taxa de perda em Ripper. As curvas Under-Ripper e SMOTE-Ripper se cruzam, e mais pontos da curva Under-Ripper estão no casco convexo do ROC.

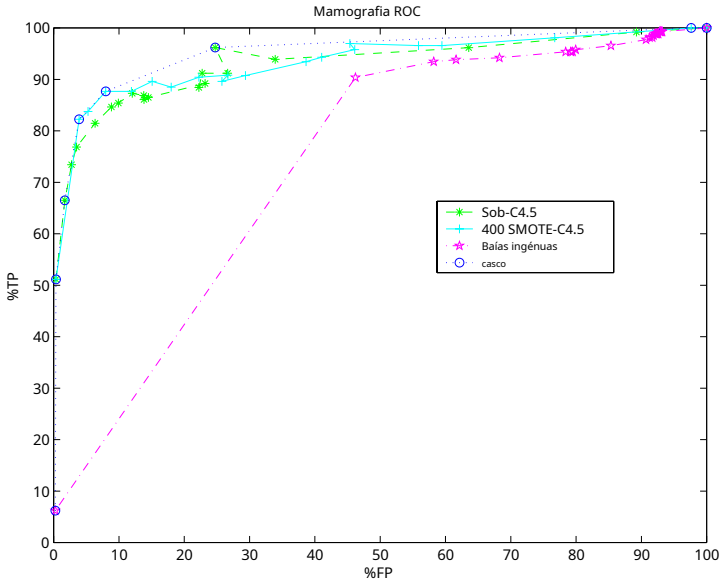

Figura 17: Mamografia. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. As curvas SMOTE-C4.5 e Under-C4.5 se cruzam no espaço ROC; no entanto, em virtude do número de pontos no casco convexo do ROC, o SMOTE-C4.5 possui classificadores potencialmente ótimos.

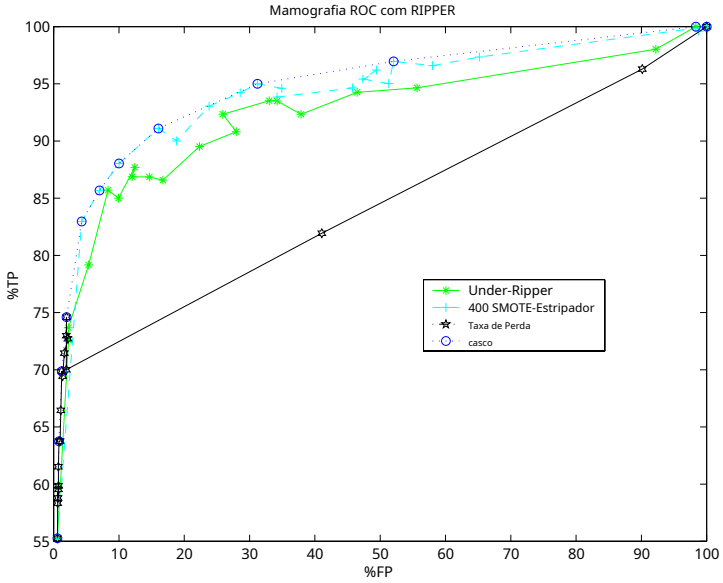

Figura 18: Mamografia. Comparação de SMOTE-Ripper, Under-Ripper e modificação Taxa de perda no Ripper. SMOTE-Ripper domina o espaço ROC para TP>75%.

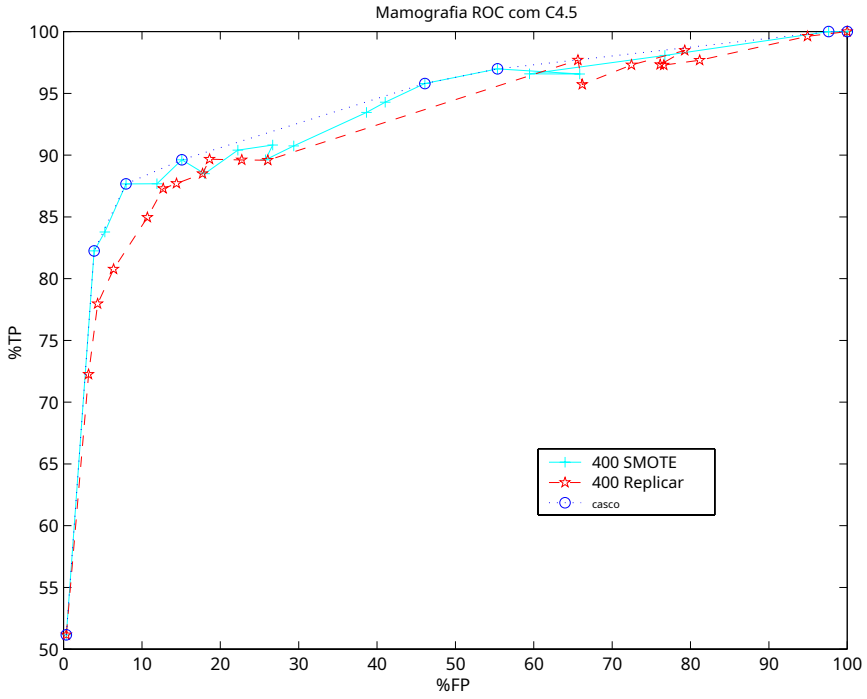

Figura 19: Uma comparação de exemplos de classe minoritária de sobreamostragem por SMOTE e sobre-amostragem dos exemplos de classe minoritária por replicação para o conjunto de dados de mamografia.

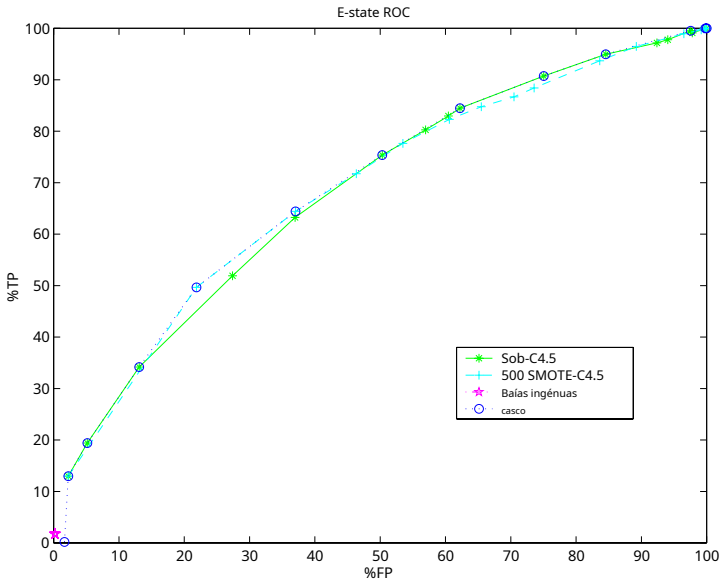

Figura 20: Estado eletrônico. (a) Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. As curvas SMOTE-C4.5 e Under-C4.5 se cruzam no espaço ROC; no entanto, o SMOTE-C4.5 possui classificadores mais potencialmente ótimos, com base no número de pontos no casco convexo do ROC.

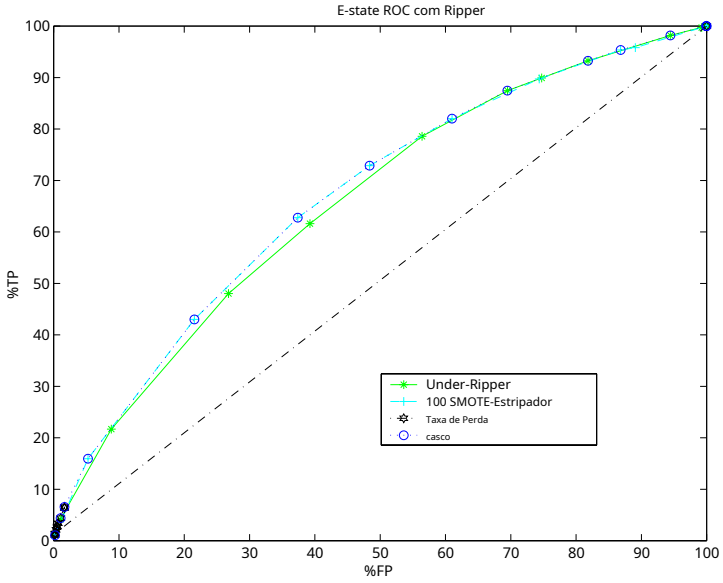

Figura 21: Estado eletrônico. Comparação de SMOTE-Ripper, Under-Ripper e perda de modificação Relação no Estripador. O SMOTE-Ripper possui classificadores potencialmente ótimos, com base no número de pontos no casco convexo do ROC.

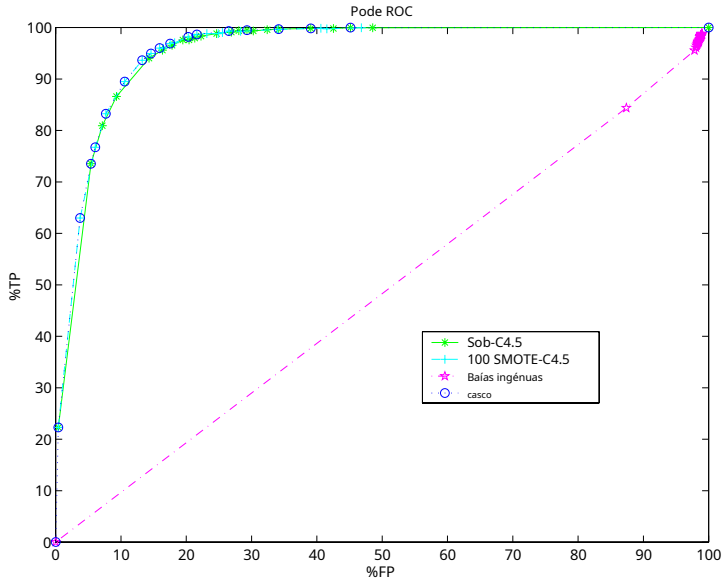

Figura 22: Lata. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. SMOTE-
As curvas ROC C4.5 e Under-C4.5 se sobrepõem na maior parte do espaço ROC.

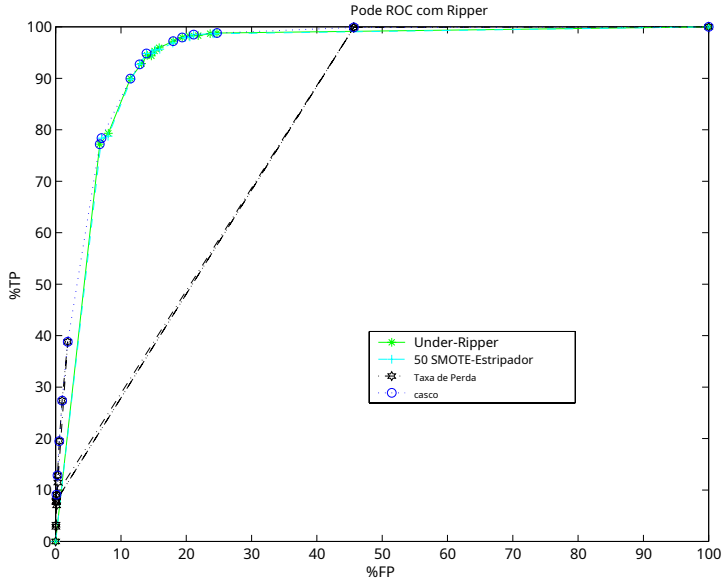

Figura 23: Lata. Comparação de SMOTE-Ripper, Under-Ripper e modificação da taxa de perda em Ripper. As curvas ROC SMOTE-Ripper e Under-Ripper se sobrepõem na maior parte do espaço ROC.

Conjunto de dados	Debaixo	50 SMOTE	100 SMOTE	200 SMOTE	300 SMOTE	400 SMOTE	500 SMOTE
Pima	7242		**7307**				
Fonema	8622		8644	**8661**			
Satimagem	8900		8957	**8979**	8963	8975	8960
Cobertura florestal	9807		9832	9834	**9849**	9841	9842
Óleo	8524		8523	8368	8161	8339	**8537**
Mamografia	9260		9250	9265	9311	**9330**	9304
Estado	6811		6792	**6828**	6784	6788	6779
Posso	9535	**9560**	9505	9505	9494	9472	9470

Tabela 3: AUCs [C4.5 como classificador base] com o melhor destacado em negrito.

curvas se sobrepõem no espaço ROC. Para todos os outros conjuntos de dados, SMOTE-*classificador*tem mais classificadores potencialmente ótimos do que qualquer outra abordagem.

5.4 Comparação adicional para alterar os limites de decisão

Provost (2000) sugeriu que a simples mudança do limiar de decisão deve sempre ser considerada como uma alternativa a abordagens mais sofisticadas. No caso de C4.5, isso significaria alterar o limite de decisão nas folhas das árvores de decisão. Por exemplo, uma folha pode classificar exemplos como classe minoritária mesmo que mais de 50% dos exemplos de treinamento na folha representem a classe majoritária. Experimentamos definindo os limites de decisão nas folhas para o aluno da árvore de decisão C4.5 em 0,5, 0,45, 0,42, 0,4, 0,35, 0,32, 0,3, 0,27, 0,25, 0,22, 0,2, 0,17, 0,15, 0,12, 0,1, 0,05 , 0,0. Nós experimentamos no conjunto de dados Phoneme. A Figura 24 mostra a comparação da combinação SMOTE e subamostragem com o aprendizado C4.5 ajustando o viés para a classe minoritária.

5.5 Comparação adicional com seleção unilateral e SHRINK

Para o conjunto de dados de óleo, também seguimos uma linha de experimentos ligeiramente diferente para obter resultados comparáveis a (Kubat et al., 1998). Para aliviar o problema de conjuntos de dados desequilibrados, os autores propuseram (a) seleção unilateral para sub-amostragem da classe majoritária (Kubat & Matwin, 1997) e (b) o sistema SHRINK (Kubat et al., 1998). A Tabela 5.5 contém os resultados de (Kubat et al., 1998). Acc+ é a precisão em exemplos positivos (minoritários) e Acc−é a precisão nos exemplos negativos (maioria). A Figura 25 mostra a tendência para Acc+ e Acc−para uma combinação da estratégia SMOTE e vários graus de subamostragem da classe majoritária. O eixo Y representa a precisão e o eixo X representa a porcentagem da classe majoritária subamostrada. Os gráficos indicam que na faixa de subamostragem entre 50% e 125% os resultados são comparáveis aos alcançados pelo SHRINK e melhores que o SHRINK em alguns casos. A Tabela 5.5 resume os resultados para o SMOTE a 500% e combinação de subamostragem. Também tentamos combinações de SMOTE em 100-400% e graus variados de subamostragem e obtivemos resultados comparáveis. o

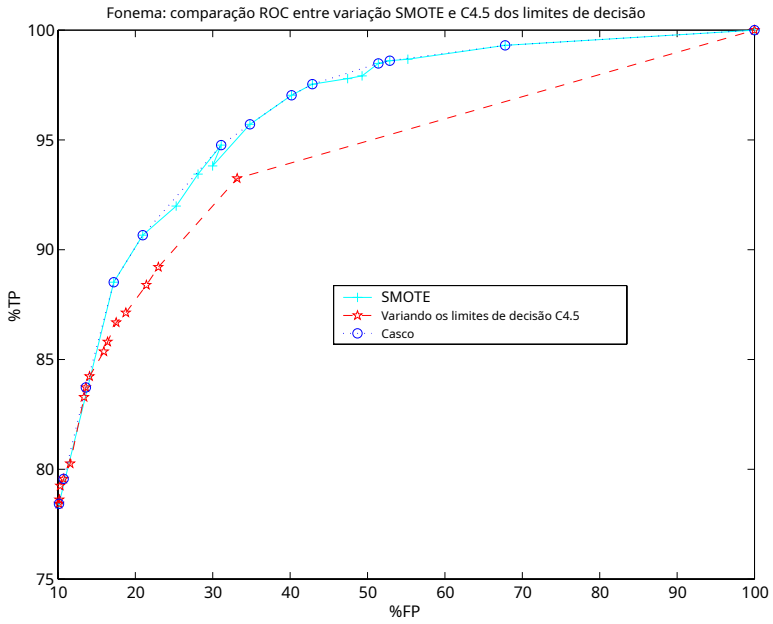

Figura 24: Combinação SMOTE e Subamostragem em relação ao aprendizado C4.5 ajustando o preconceito contra a classe minoritária

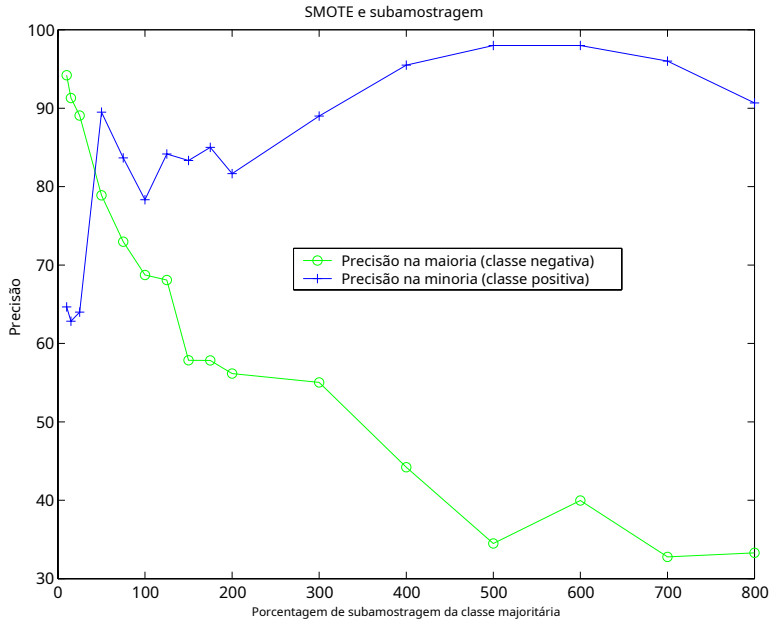

Figura 25: Desempenho da combinação SMOTE (500 OU) e subamostragem

A abordagem SHRINK e nossa abordagem SMOTE não são diretamente comparáveis, pois veem pontos de dados diferentes. O SMOTE não oferece nenhuma melhoria clara em relação à seleção unilateral.

Método	Ace+	Conta−
PSIQUIATRA	82,5%	60,9%
Seleção unilateral	76,0%	86,6%

Tabela 4: Resultados da validação cruzada (Kubat et al., 1998)

% de subamostragem	Ace+	Conta−
10%	64,7%	94,2%
15%	62,8%	91,3%
25%	64,0%	89,1%
50%	89,5%	78,9%
75%	83,7%	73,0%
100%	78,3%	68,7%
125%	84,2%	68,1%
150%	83,3%	57,8%
175%	85,0%	57,8%
200%	81,7%	56,7%
300%	89,0%	55,0%
400%	95,5%	44,2%
500%	98,0%	35,5%
600%	98,0%	40,0%
700%	96,0%	32,8%
800%	90,7%	33,3%

Tabela 5: Resultados de validação cruzada para SMOTE em 500% SMOTE no conjunto de dados de óleo.

6. Trabalho Futuro

Há vários tópicos a serem considerados nesta linha de pesquisa. A seleção adaptativa automatizada do número de vizinhos mais próximos seria valiosa. Diferentes estratégias para criar os vizinhos sintéticos podem melhorar o desempenho. Além disso, selecionar os vizinhos mais próximos com foco em exemplos classificados incorretamente pode melhorar o desempenho. Uma amostra de classe minoritária poderia ter uma amostra de classe majoritária como seu vizinho mais próximo, em vez de uma amostra de classe minoritária. Essa aglomeração provavelmente contribuirá para o redesenho das superfícies de decisão em favor da classe minoritária. Além desses tópicos, as subseções a seguir discutem duas extensões possíveis do SMOTE e uma aplicação do SMOTE à recuperação de informações.

6.1 SMOTE-NC

Embora nossa abordagem SMOTE atualmente não lide com conjuntos de dados com todos os recursos nominais, ela foi generalizada para lidar com conjuntos de dados mistos de recursos contínuos e nominais. Chamamos essa abordagem Synthetic Minority Over-sampling TEchnique-Nominal Continuous [SMOTE-NC]. Testamos essa abordagem no conjunto de dados Adult do repositório UCI. O algoritmo SMOTE-NC é descrito abaixo.

1. Cálculo da mediana: Calcule a mediana dos desvios padrão de todas as características contínuas para a classe minoritária. Se as características nominais diferem entre uma amostra e seus potenciais vizinhos mais próximos, então esta mediana é incluída no cálculo da distância euclidiana. Usamos a mediana para penalizar a diferença de características nominais por um valor que está relacionado à diferença típica em valores de características contínuas.

2. Computação do vizinho mais próximo: Calcule a distância euclidiana entre o vetor de características para o qual os k vizinhos mais próximos estão sendo identificados (amostra de classe minoritária) e os outros vetores de característica (amostra de classe minoritária) usando o espaço de característica contínua. Para cada característica nominal diferente entre o vetor de características considerado e seu potencial vizinho mais próximo, inclua a mediana dos desvios padrão calculados anteriormente, no cálculo da distância euclidiana. A Tabela 2 mostra um exemplo.

F1 = 1 2 3 ABC [Seja esta a amostra para a qual estamos computando os vizinhos mais próximos]
F2 = 4 6 5 ADE F3
= 3 5 6 ABK
Então, a distância euclidiana entre F2 e F1 seria: Eucl
= sqrt[(4-1)$_2$+ (6-2)$_2$+ (5-3)$_2$+Med$_2$+Med$_2$]
Medé a mediana dos desvios padrão das características contínuas da classe minoritária.
O termo mediano é incluído duas vezes para os números de recurso 5: B→D e 6: C→E, que diferem para os dois vetores de características: F1 e F2.

Tabela 6: Exemplo de cálculo do vizinho mais próximo para SMOTE-NC.

3. Preencha a amostra sintética: Os recursos contínuos da nova amostra sintética da classe minoritária são criados usando a mesma abordagem do SMOTE descrita anteriormente. A característica nominal recebe o valor que ocorre na maioria dos k vizinhos mais próximos.

Os experimentos SMOTE-NC relatados aqui são configurados da mesma forma que aqueles com SMOTE, exceto pelo fato de examinarmos apenas um conjunto de dados. O SMOTE-NC com o conjunto de dados Adulto difere do nosso resultado típico: ele tem um desempenho pior do que a subamostragem simples com base na AUC, conforme mostrado nas Figuras 26 e 27. Extraímos apenas recursos contínuos para separar o efeito do SMOTE e do SMOTE-NC neste conjunto de dados e para determinar se essa estranheza foi devido ao nosso manuseio de recursos nominais. Conforme mostrado na Figura 28, mesmo o SMOTE com apenas recursos contínuos aplicados ao conjunto de dados Adulto não alcança nenhum desempenho melhor do que a subamostragem simples. Algumas das características contínuas da classe minoritária têm uma variância muito alta, portanto, a geração sintética de amostras da classe minoritária pode se sobrepor ao espaço da classe majoritária, levando assim a mais falsos positivos do que subamostragem simples. Esta hipótese também é apoiada pela medida de AUC diminuída à medida que SMOTE em graus superiores a 50%. Os graus mais altos de SMOTE levam a mais amostras de classes minoritárias no conjunto de dados e, portanto, uma maior sobreposição com o espaço de decisão da classe majoritária.

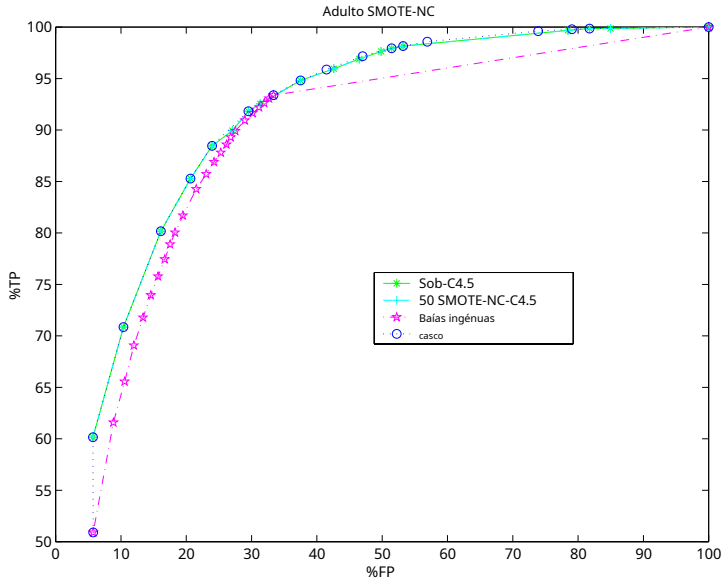

Figura 26: Adulto. Comparação de SMOTE-C4.5, Under-C4.5 e Naive Bayes. SMOTE-
As curvas ROC C4.5 e Under-C4.5 se sobrepõem na maior parte do espaço ROC.

6.2 SMOTE-N

Potencialmente, o SMOTE também pode ser estendido para características nominais — SMOTE-N — com os vizinhos mais próximos computados usando a versão modificada do Value Difference Metric (Stanfill & Waltz, 1986) proposto por Cost e Salzberg (1993). A Métrica de Diferença de Valor (VDM) analisa a sobreposição de valores de recursos em todos os vetores de recursos. Uma matriz que define a distância

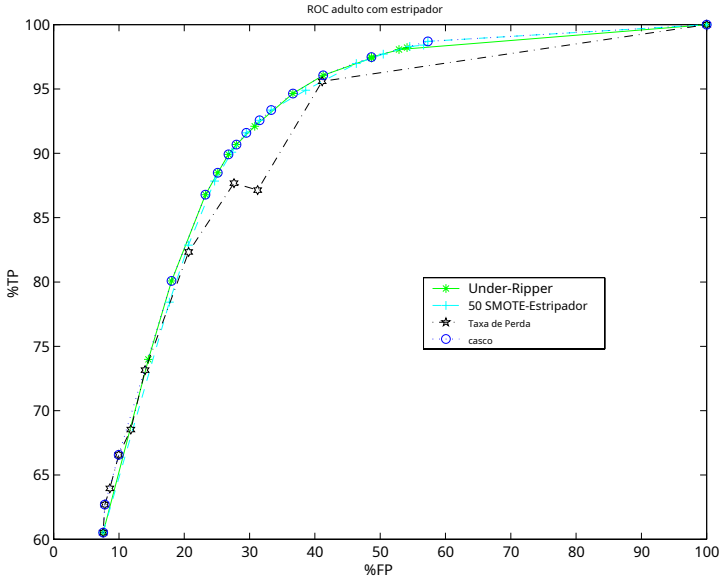

Figura 27: Adulto. Comparação de SMOTE-Ripper, Under-Ripper e modificação da taxa de perda tio em Ripper. As curvas ROC SMOTE-Ripper e Under-Ripper se sobrepõem na maior parte do espaço ROC.

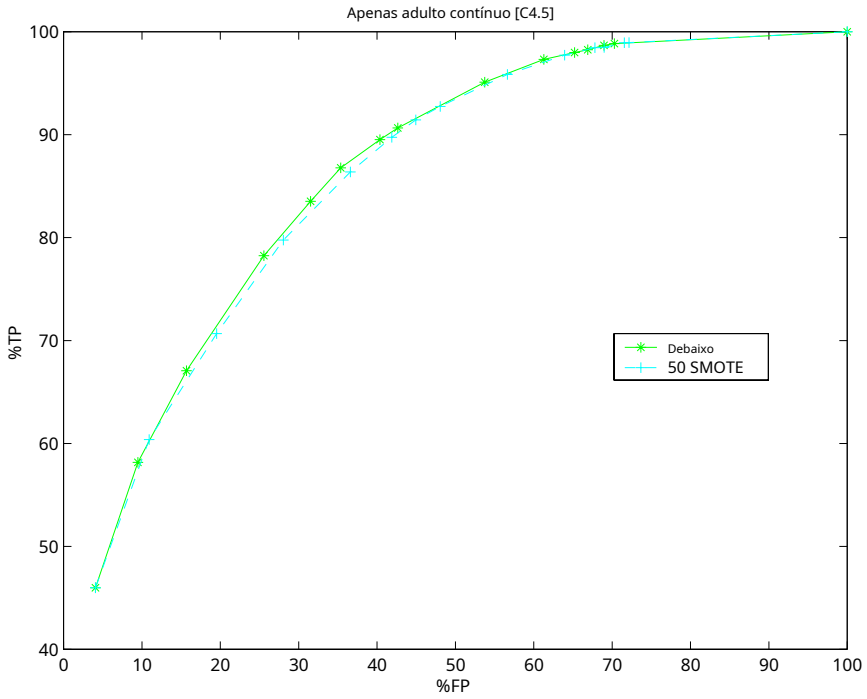

Figura 28: Adulto com apenas traços contínuos. A sobreposição de SMOTE-C4.5 e Under-C4.5 também é observado neste cenário.

entre os valores de recursos correspondentes para todos os vetores de recursos é criado. A distancia δ entre dois valores de característica correspondentes é definido como segue.

$$\delta(V_1, V_2) = \sum_{eu=1}^{n} \left| \frac{C_{1eu}}{C_1} - \frac{C_{2eu}}{C_2} \right| \quad (1)$$

Na equação acima, V_1 e V_2 são os dois valores de característica correspondentes. C_1 é o número total de ocorrências do valor do recurso V_1, e C_{1eu} é o número de ocorrências do valor do recurso V_1 para classe eu. Uma convenção semelhante também pode ser aplicada a C_{2eu} e C_2. k é uma constante, geralmente definida como 1. Esta equação é usada para calcular a matriz de diferenças de valor para cada característica nominal no conjunto dado de vetores de características. A Equação 1 fornece uma distância geométrica em um conjunto fixo e finito de valores (Cost & Salzberg, 1993). Custo e VDM modificado de Salzberg omite o termo de peso W_{uma} incluído em δ computação por Stanfill e Waltz, que tem um efeito de fazer δ simétrico. A distância Δ entre dois vetores de características é dada por:

$$\Delta(X, Y) = W_x W_y \sum_{eu=1}^{N} \delta(x_{eu}, y_{eu}) \quad (2)$$

r=1 fornece a distância de Manhattan, e r=2 fornece a distância euclidiana (Cost & Salzberg, 1993). W_x e W_y são os pesos exemplares no VDM modificado. W_y=1 para um novo exemplo (vetor de recursos), e W_x é a tendência para exemplos mais confiáveis (vetores de características) e é calculado como a razão do número de usos de um vetor de características para o número de usos corretos do vetor de características; assim, vetores de recursos mais precisos terão $W_x \approx$ 1. Para o SMOTE-N podemos ignorar esses pesos na equação 2, pois o SMOTE-N não é usado diretamente para fins de classificação. No entanto, podemos redefinir esses pesos para dar mais peso aos vetores de características da classe minoritária que se aproximam dos vetores de características da classe majoritária; assim, fazendo com que essas características da classe minoritária pareçam mais distantes do vetor de características em consideração. Como estamos mais interessados em formar regiões mais amplas, porém precisas, da classe minoritária, os pesos podem ser usados para evitar o povoamento de vizinhos que se aproximem da classe majoritária. Para gerar novos vetores de recursos de classe minoritária, podemos criar novos valores de recursos definidos levando em consideração o voto majoritário do vetor de recursos e sua k vizinhos mais próximos. A Tabela 6.2 mostra um exemplo de criação de um vetor de recurso sintético.

Seja F1 = ABCDE o vetor de características em consideração e seus 2 vizinhos mais próximos sejam F2 = AFCGN F3 = HBCDN A aplicação do SMOTE-N criaria o seguinte vetor de características: FS = ABCDN

Tabela 7: Exemplo de SMOTE-N

6.3 Aplicação do SMOTE à Recuperação de Informações

Estamos investigando a aplicação do SMOTE para recuperação de informações (IR). Os problemas de IR vêm com uma infinidade de recursos e potencialmente muitas categorias. O SMOTE teria que ser aplicado em conjunto com um algoritmo de seleção de recursos, depois de transformar o documento ou página da Web em um formato de pacote de palavras.

Uma comparação interessante com o SMOTE seria a combinação de Naive Bayes e *Razão de probabilidade*. *Razão de probabilidade*concentra-se em uma classe-alvo e classifica os documentos de acordo com sua relevância para a classe-alvo ou classe positiva. O SMOTE também se concentra em uma classe de destino criando mais exemplos dessa classe.

7. Resumo

Os resultados mostram que a abordagem SMOTE pode melhorar a precisão dos classificadores para uma classe minoritária. O SMOTE fornece uma nova abordagem para sobreamostragem. A combinação de SMOTE e subamostragem tem um desempenho melhor do que a subamostragem simples. O SMOTE foi testado em uma variedade de conjuntos de dados, com vários graus de desequilíbrio e quantidades variáveis de dados no conjunto de treinamento, fornecendo assim um banco de testes diversificado. A combinação de SMOTE e subamostragem também funciona melhor, com base na dominação no espaço ROC, do que variando as taxas de perda no Ripper ou variando as classes prioritárias no Naive Bayes Classifier: os métodos que poderiam lidar diretamente com a distribuição de classes distorcidas. O SMOTE força o aprendizado focado e introduz um viés em relação à classe minoritária. Apenas para Pima — o conjunto de dados menos distorcido — o classificador Naive Bayes tem um desempenho melhor que o SMOTE-C4.5. Além disso, apenas para o conjunto de dados de óleo o Under-Ripper tem um desempenho melhor do que o SMOTE-Ripper. Para o conjunto de dados Can, SMOTE-*classificador*e abaixo-*classificador*As curvas ROC se sobrepõem no espaço ROC. Para todos os outros conjuntos de dados SMOTE-*classificador*tem um desempenho melhor do que Under-*classificador*, Índice de Perdas e Naive Bayes. De um total de 48 experimentos realizados, SMOTE-*classificador*não tem o melhor desempenho apenas para 4 experimentos.

A interpretação de por que a sobreamostragem minoritária sintética melhora o desempenho, enquanto a sobreamostragem minoritária com substituição não é bastante direta. Considere o efeito nas regiões de decisão no espaço de características quando a sobreamostragem minoritária é feita por replicação (amostragem com substituição) versus a introdução de exemplos sintéticos. Com a replicação, a região de decisão que resulta em uma decisão de classificação para a classe minoritária pode realmente se tornar menor e mais específica à medida que as amostras minoritárias na região são replicadas. Este é o oposto do efeito desejado. Nosso método de sobreamostragem sintética funciona para fazer com que o classificador construa regiões de decisão maiores que contenham pontos de classe minoritários próximos. As mesmas razões podem ser aplicáveis ao SMOTE ter um desempenho melhor do que a sinistralidade do Ripper e Naive Bayes; esses métodos, no entanto, ainda estão aprendendo com as informações fornecidas no conjunto de dados, embora com informações de custo diferentes. O SMOTE fornece mais amostras relacionadas de classes minoritárias para aprender, permitindo assim que um aluno esculpa regiões de decisão mais amplas, levando a uma maior cobertura da classe minoritária.

Agradecimentos

Esta pesquisa foi parcialmente apoiada pelo Departamento de Energia dos Estados Unidos através do Programa de Descoberta de Dados do Sandia National Laboratories ASCI VIEWS, número do contrato

DE-AC04-76DO00789. Agradecemos a Robert Holte por fornecer o conjunto de dados de derramamento de óleo usado em seu artigo. Agradecemos também a Foster Provost por esclarecer seu método de uso do conjunto de dados Satimage. Gostaríamos também de agradecer aos revisores anônimos por seus vários comentários e sugestões perspicazes.

Apêndice A. Gráficos ROC para Conjunto de Dados de Petróleo

As figuras a seguir mostram diferentes conjuntos de curvas ROC para o conjunto de dados de petróleo. A Figura 29 (a) mostra as curvas ROC para o conjunto de dados de Petróleo, conforme incluído no texto principal; A Figura 29(b) mostra as curvas ROC sem o casco convexo ROC; A Figura 29(c) mostra os dois cascos convexos, obtidos com e sem SMOTE. O casco convexo ROC mostrado por linhas tracejadas e estrelas na Figura 29(c), foi calculado incluindo Under-C4.5 e Naive Bayes na família de curvas ROC. O casco convexo ROC mostrado pela linha sólida e pequenos círculos na Figura 29(c) foi calculado incluindo 500 SMOTE-C4.5, Under-C4.5 e Naive Bayes na família de curvas ROC. O casco convexo ROC com SMOTE domina o casco convexo ROC sem SMOTE, portanto o SMOTE-C4.5 contribui com classificadores mais otimizados.

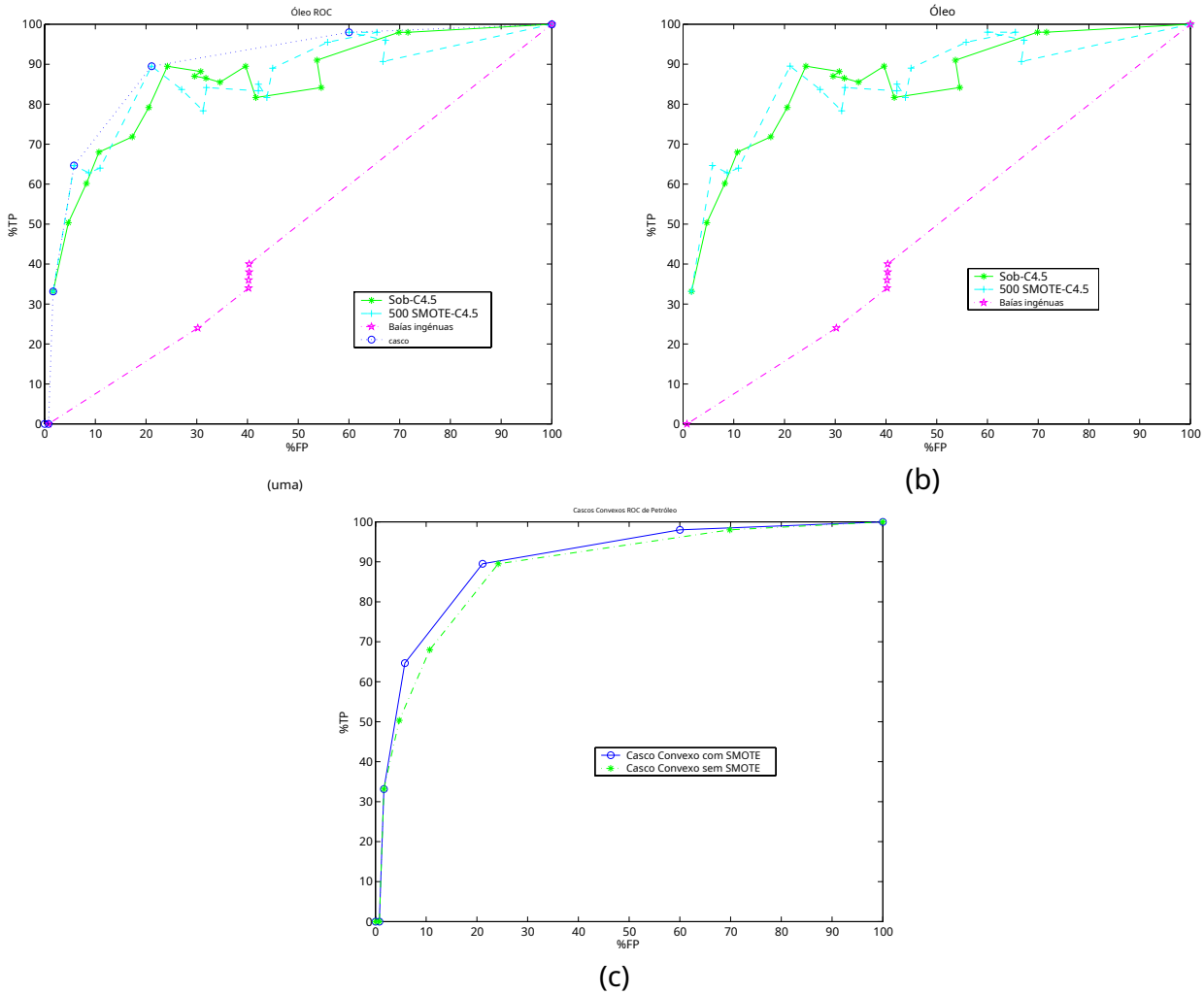

Figura 29: Curvas ROC para o Conjunto de Dados de Petróleo. (a) Curvas ROC para SMOTE-C4.5, Under-C4.5, Naive Bayes e seu casco convexo ROC. (b) Curvas ROC para SMOTE-C4.5, Under-C4.5 e Naive Bayes. (c) cascos convexos ROC com e sem SMOTE.

Referências

Blake, C., & Merz, C. (1998). Repositório UCI de Bancos de Dados de Machine Learning http://www.ics.uci.edu/~maprendiz/~MLRepository.html. Departamento de Informação e Ciências da Computação, Universidade da Califórnia, Irvine.

Bradley, AP (1997). O Uso da Área Sob a Curva ROC na Avaliação de Algoritmos de Aprendizado de Máquina.*Reconhecimento de padrões*,*30(6)*, 1145-1159.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, P. (2000). SMOTE: Minoria Sintética Técnica de sobreamostragem. Dentro*Conferência Internacional de Sistemas Computacionais Baseados no Conhecimento*, pp. 46–57. Centro Nacional de Tecnologia de Software, Mumbai, Índia, Allied Press.

Chawla, N., & Hall, L. (1999). Modificando o MUSTAFA para capturar dados salientes. Tecnologia representante ISL-99-01, Universidade do Sul da Flórida, Ciência da Computação e Eng. Departamento

Cohen, W. (1995a). Aprendendo a classificar textos em inglês com métodos ILP. Dentro*Continuar-5º Workshop Internacional de Programação em Lógica Indutiva*, pp. 3–24. Departamento de Ciência da Computação, Kaholieke Universiteit Leuven.

Cohen, WW (1995b). Indução de regra efetiva rápida. Dentro*Proc. 12º Congresso Internacional ência em Aprendizado de Máquina*, pp. 115–123 Lake Tahoe, CA. Morgan Kaufmann.

Cohen, WW, & Singer, Y. (1996). Métodos de Aprendizagem Sensíveis ao Contexto para Categoria de Texto rização. Em Frei, H.-P., Harman, D., Schäuble, P., & Wilkinson, R. (Eds.),*Anais do SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pp. 307–315 Zurique, CH. ACM Press, Nova York, EUA.

Cost, S., & Salzberg, S. (1993). Um algoritmo de vizinho mais próximo ponderado para aprender com Características Simbólicas.*Aprendizado de máquina*,*10*(1), 57-78.

DeRouin, E., Brown, J., Fausett, L., & Schneider, M. (1991). Treinamento de Rede Neural em Classes desigualmente representadas. Dentro*Sistemas Inteligentes de Engenharia Através de Redes Neurais Artificiais*, pp. 135–141 Nova York. Imprensa ASME.

Domingos, P. (1999). Metacusto: Um método geral para tornar os classificadores sensíveis ao custo. Dentro*Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164 San Diego, CA. Imprensa ACM.

Drummond, C., & Holte, R. (2000). Representando explicitamente o custo esperado: uma alternativa à Representação do ROC. Dentro*Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 198–207 Boston. ACM.

Duda, R., Hart, P., & Stork, D. (2001).*Classificação de padrões*. Wiley-Interscience.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Algoritmo de Aprendizagem Indutiva Ritmos e Representações para Categorização de Texto. Dentro*Anais da Sétima Conferência Internacional sobre Gestão da Informação e do Conhecimento.*, pp. 148–155.

Ezawa, K., J., Singh, M., & Norton, S., W. (1996). Bayesiano Orientado a Objetivos de Aprendizagem Redes para Gestão de Riscos em Telecomunicações. Dentro*Anais da Conferência Internacional sobre Aprendizado de Máquina, ICML-96*, pp. 139–147 Bari, Itália. Morgan Kauffmann.

Fawcett, T., & Provost, F. (1996). Combinando mineração de dados e aprendizado de máquina para eficiência Perfil de usuário efetivo. Dentro*Anais da 2ª Conferência Internacional sobre Descoberta de Conhecimento e Mineração de Dados*, pp. 8–13 Portland, OR. AAAI.

Ha, TM, & Bunke, H. (1997). Off-line, Reconhecimento Numérico Manuscrita por Perturbação Método.*Análise de padrões e inteligência de máquina*,*19/5*, 535-539.

Hall, L., Mohney, B., & Kier, L. (1991). O Estado Eletrotopológico: Informações da Estrutura no Nível Atômico para Gráficos Moleculares.*Jornal de Informação Química e Ciência da Computação*,*31*(76).

Japkowicz, N. (2000). O Problema do Desequilíbrio de Classes: Significado e Estratégias. Dentro*Pró-Ceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Indutive Learning*Las Vegas, Nevada.

Kubat, M., Holte, R., & Matwin, S. (1998). Aprendizado de Máquina para Detecção de Petróleo Derramamentos em Imagens de Radar de Satélite.*Aprendizado de máquina*,*30*, 195-215.

Kubat, M., & Matwin, S. (1997). Lidando com a maldição dos conjuntos de treinamento desequilibrados: um Seleção Lateral. Dentro*Anais da XIV Conferência Internacional de Aprendizado de Máquina*, pp. 179–186 Nashville, Tennessee. Morgan Kaufmann.

Lee, S. (2000). Replicação barulhenta na classificação binária distorcida.*Estatísticas Computacionais e Análise de Dados*,*34*.

Lewis, D., & Catlett, J. (1994). Amostragem de Incerteza Heterogênea para Aprendizado Supervisionado ing. Dentro*Anais da XI Conferência Internacional de Aprendizado de Máquina*, pp. 148–156 São Francisco, CA. Morgan Kaufmann.

Lewis, D., & Ringuette, M. (1994). Uma comparação de dois algoritmos de aprendizagem para texto Categorização. Dentro*Anais do SDAIR-94, 3º Simpósio Anual de Análise de Documentos e Recuperação de Informações*, pp. 81-93.

Ling, C., & Li, C. (1998). Mineração de dados para problemas e soluções de marketing direto. Dentro *Anais da Quarta Conferência Internacional sobre Descoberta de Conhecimento e Mineração de Dados (KDD-98)*Nova York, NY. Imprensa AAAI.

Mladenić, D., & Grobelnik, M. (1999). Seleção de recursos para distribuição de classe não balanceada e Naive Bayes. Dentro*Anais da 16ª Conferência Internacional de Aprendizado de Máquina.*, pp. 258–267. Morgan Kaufmann.

O'Rourke, J. (1998).*Geometria Computacional em C*. Cambridge University Press, Reino Unido.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Redução Custos de classificação incorreta. Dentro*Anais da XI Conferência Internacional sobre Aprendizado de Máquina*São Francisco, CA. Morgan Kauffmann.

Provost, F., & Fawcett, T. (2001). Classificação robusta para ambientes imprecisos.*Maaprendizagem chinesa*,*42/3*, 203-231.

Provost, F., Fawcett, T., & Kohavi, R. (1998). O Caso Contra a Estimativa de Precisão para Comparar Algoritmos de Indução. Dentro*Anais da XV Conferência Internacional de Aprendizado de Máquina*, pp. 445–453 Madison, WI. Morgan Kauffmann.

Quinlan, J. (1992).*C4.5: Programas para Aprendizado de Máquina*. Morgan Kaufmann, San Mateo, CA.

Solberg, A., & Solberg, R. (1996). Uma avaliação em larga escala de recursos para automação Detecção de Derramamentos de Petróleo em Imagens ERS SAR. Dentro*Simpósio Internacional de Geociências e Sensoriamento Remoto*, pp. 1484–1486 Lincoln, NE.

Stanfill, C., & Waltz, D. (1986). Rumo ao Raciocínio Baseado na Memória.*Comunicações de o ACM*,*29*(12), 1213-1228.

Swets, J. (1988). Medindo a Precisão de Sistemas de Diagnóstico.*Ciência*,*240*, 1285-1293.

Tomek, I. (1976). Duas modificações da CNN.*Transações IEEE em Sistemas, Homem e Cibernética*,*6*, 769-772.

Turney, P. (1996). Bibliografia sensível ao custo. http://ai.iit.nrc.ca/bibiliographies/cost-sensitivo.html.

van Rijsbergen, C., Harper, D., & Porter, M. (1981). A seleção de bons termos de pesquisa. *Processamento e Gerenciamento de Informações*,*17*, 77-91.

Woods, K., Doss, C., Bowyer, K., Solka, J., Priebe, C., & Kegelmeyer, P. (1993). Comparação Avaliação tiva de Técnicas de Reconhecimento de Padrões para Detecção de Microcalcificações em Mamografia.*Jornal Internacional de Reconhecimento de Padrões e Inteligência Artificial*,*7(6)*, 1417-1436.