Relatório Técnico

Descrição de uma Abordagem Híbrida para Aprender com Classes Desbalanceadas Utilizando Algoritmos Genéticos

Claudia Regina Milaré
Gustavo E. A. P. A. Batista
André C. P. L. F. de Carvalho

Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - USP

Setembro de 2010

Sumário

Sumário **ii**

Resumo **iii**

1 Introdução **1**

2 Trabalhos Relacionados **2**

2.1 Classes Desbalanceadas 3

2.2 Combinando Classificadores 4

3 Abordagem Proposta **5**

4 Avaliação Experimental **8**

5 Conclusão e Trabalhos Futuros **15**

Referências **16**

Resumo

Há um interesse crescente na aplicação de algoritmos evolutivos para induzir regras de classificação. Essa abordagem pode ajudar em áreas que métodos clássicos para indução de regras não têm obtido tanto sucesso. Um exemplo é a indução de regras de classificação em domínios desbalanceados. Dados desbalanceados ocorrem quando algumas classes possuem um número bem maior de exemplos se comparado a outras classes. Geralmente, em Aprendizado de Máquina (AM) tradicional os indutores não são capazes de aprender na presença de conjuntos de dados desbalanceados. Estes indutores geralmente classificam todos os exemplos como sendo da classe que possui o maior número de exemplos. Neste relatório é descrita uma abordagem híbrida para resolver o problema de indução de regras de classificação em domínios desbalanceados, bem como os experimentos realizados para avaliá-la. Nesta abordagem híbrida são criados vários conjuntos de dados balanceados com todos os exemplos da classe minoritária e uma amostra randômica de exemplos da classe majoritária. Esses conjuntos de dados balanceados são fornecidos a sistemas de AM tradicionais, que produzem como saída conjuntos de regras. Os conjuntos de regras são combinados em um repositório de regras e um algoritmo evolutivo é utilizado para selecionar algumas regras deste repositório para construir um classificador. A abordagem descrita possui vantagem em relação a métodos de *under-sampling* desde que reduz a quantidade de informação descartada, e possui vantagem em relação a métodos de *over-sampling*, desde que evita *overfitting*. Esta abordagem foi experimentalmente analizada e os resultados dos experimentos mostram uma melhora no desempenho de classificação medido com a área abaixo da curva ROC (*Receiver Operating Characteristic*).

1 Introdução

Há um interesse crescente na aplicação de algoritmos evolutivos para induzir regras de classificação. Essa abordagem pode ajudar em áreas nas quais os métodos clássicos para indução de regras não têm obtido tanto sucesso. Um exemplo é a indução de regras de classificação em domínios desbalanceados.

Dados desbalanceados ocorrem quando algumas classes possuem um número bem maior de exemplos se comparado a outras classes. Geralmente, em Aprendizado de Máquina (AM) tradicional os indutores não são capazes de aprender na presença de conjuntos de dados desbalanceados. Estes indutores geralmente classificam todos os exemplos como sendo da classe que possui o maior número de exemplos. Entretanto, em muitos domínios, as classes minoritárias são as classes de maior interesse, as quais são atribuídos os custos mais altos. Por exemplo, na detecção de transações fraudulentas em cartões de crédito e na telefonia Phua et al. (2004), no diagnóstico de doenças raras Cohena et al. (2006) e na predição de eventos climáticos Bucene (2008), classificar erroneamente exemplos da classe minoritária é mais caro do que classificar erroneamente um exemplo da classe majoritária. Um classificador que simplesmente classifica todos os exemplos como sendo da classe majoritária é inútil.

Em Milaré et al. (2009b) é proposta uma abordagem híbrida para resolver o problema de indução de regras de classificação em domínios desbalanceados. Nessa abordagem, o problema de aprender com classes desbalanceadas é visto como um problema de busca, e portanto, um algoritmo evolutivo é utilizado para melhorar a busca no espaço de hipóteses. A abordagem proposta cria vários conjuntos de dados balanceados com todos os exemplos da classe minoritária e uma amostra randômica de exemplos da classe majoritária. Esses conjuntos de dados balanceados são fornecidos a sistemas de AM tradicionais, que produzem como saída conjuntos de regras. Os conjuntos de regras são combinados em um repositório de regras e um algoritmo evolutivo é utilizado para selecionar algumas regras desse repositório para construir um classificador.

O método proposto em Milaré et al. (2009b) utiliza a técnica *under-sampling* para criar vários conjuntos de dados balanceados. *Under-sampling* elimina os exemplos da classe majoritária para criar conjuntos de dados balanceados, e portanto, pode descartar dados úteis que poderiam ser importantes para o processo de indução. A abordagem híbrida não possui esta limitação pois muitas amostras de dados são criadas e, portanto, aumenta a probabilidade de

todos os dados serem utilizados no aprendizado. Outra técnica utilizada para tratar dados desbalanceados é *over-sampling*. *Over-sampling* artificialmente aumenta o número de exemplos da classe minoritária. O maior problema desta técnica é que ela supostamente aumenta a probabilidade de ocorrência de *overfitting*, já que faz muitas cópias dos exemplos da classe minoritária. O método proposto evita *overfitting* pois limita o número de regras de cada classificador. Os classificadores criados possuem aproximadamente o mesmo número de regras dos classificadores individuais induzidos a partir dos dados balanceados.

A abordagem híbrida proposta em Milaré et al. (2009b) foi experimentalmente analizada sobre alguns conjuntos de dados desbalanceados utilizando inicialmente dois sistemas de AM bastante conhecidos, C4.5Rules Quinlan (1988) e Ripper Cohen (1995) e os resultados foram publicados em Milaré et al. (2010). Posteriormente, os experimentos foram estendidos com a utilização do CN2 Clark and Niblett (1989) e os resultados foram publicados em Milaré et al. (2009a). A principal métrica utilizada para avaliar os resultados foi a área abaixo da curva ROC (*Receiver Operating Characteristic*), a AUC (*Area Under the ROC Curve*) Fawcett (2004). O uso da métrica AUC é altamente recomendado em experimentos com classes desbalanceadas, pois métricas que dependem dos custos e distribuição das classes, tal como acurácia e taxa de erro, podem enganar em domínios desbalanceados.

O objetivo deste relatório técnico é descrever a abordagem proposta em Milaré et al. (2009b), bem como os experimentos realizados para a sua avaliação.

Este trabalho está organizado da seguinte forma: na Seção 2 são apresentados alguns trabalhos relacionados; na Seção 3 é descrita a abordagem proposta; na Seção 4 são descritos os experimentos realizados; e finalmente, na Seção 5 algumas conclusões e trabalhos futuros são apresentados.

2 Trabalhos Relacionados

Nesta seção são descritos alguns trabalhos relacionados a classes desbalanceadas e métodos para combinar classificadores.

2.1 Classes Desbalanceadas

A precisão de classificação de muitos algoritmos de AM é altamente afetada pela distribuição dos exemplos entre as classes. Como muitos sistemas de aprendizado são projetados para trabalhar com conjuntos de dados balanceados, eles geralmente falham na indução de classificadores capazes de predizer a classe minoritária.

A investigação de alternativas para trabalhar de forma eficiente com problemas que envolvem classes desbalanceadas é uma área importante de pesquisa, pois conjuntos de dados desbalanceados podem ser encontrados em diversos domínios. Por exemplo, na detecção de fraudes em chamadas telefônicas e em transações de cartão de crédito Fawcett and Provost (1997); Stolfo et al. (1997); Phua et al. (2004), o número de transações legítimas é geralmente muito maior do que o número de transações fraudulentas; em análise de risco em seguro Pednault et al. (2000), poucos clientes requisitam o prêmio de seguro em um determinado intervalo de tempo; e em marketing direto Ling and Li (1998), o retorno é geralmente muito pequeno (em torno de 1%) para a maioria das campanhas de marketing.

Muitos trabalhos têm analizado o problema de aprender sobre conjunto de dados desbalanceados (por exemplo, Pazzani et al. (1994); Ling and Li (1998); Kubat and Matwin (1997); Fawcett and Provost (1997); Kubat et al. (1998a); Japkowicz and Stephen (2002); Batista et al. (2004); Weiss (2004)). Entre as principais estratégias utilizadas por estes trabalhos, três abordagens se destacam:

- Aplicar diferentes custos para classificações incorretas: os custos mais altos são aplicados às classes minoritárias;
- *Under-sampling*: balancear artificialmente os dados de treinamento eliminando exemplos da classe majoritária;
- *Over-sampling*: balancear artificialmente os dados de treinamento replicando exemplos da classe minoritária.

A abordagem híbrida proposta em Milaré et al. (2009b) utiliza *under-sampling* como um passo intermediário para criar muitos conjuntos de dados balanceados. Outros trabalhos utilizam uma abordagem semelhante para tratar classes desbalanceadas. Por exemplo, Chan and Stolfo

(1998) divide os exemplos da classe majoritária em muitos subconjuntos não sobrepostos com aproximadamente o mesmo número de exemplos da classe minoritária. Cada subconjunto é combinado com os exemplos da classe minoritária para formar conjuntos de dados balanceados que são fornecidos a um algoritmo de aprendizado. Os classificadores obtidos são integrados utilizando *stacking* Wolpert (1992). Uma abordagem similar é proposta em Liu et al. (2009), em que Adaboost Freund and Schapire (1997) integra a saída de muitos classificadores induzidos a partir de conjuntos de dados balanceados tratados com *under-sampling*.

A abordagem híbrida proposta difere de trabalhos previamente publicados, pois o interesse é criar classificadores simbólicos, isto é, classificadores que podem ser facilmente interpretados por humanos. Embora *ensembles* possam ser construídos a partir de diversos classificadores simbólicos individuais, o classificador final não pode ser considerado um classificador simbólico, pois este classificador não pode ser facilmente interpretado.

2.2 Combinando Classificadores

Muitos trabalhos na literatura descrevem alternativas diferentes para combinar classificadores. Uma aboradagem direta para combinar classificadores é utilizar *ensembles* Opitz and Maclin (1999); Dietterich (1997b). Um *ensemble* é composto por um conjunto de classificadores individuais cujas predições são combinadas para determinar a classe a que pertence um novo exemplo. Geralmente, um *ensemble* é mais preciso que seus classificadores individuais. Apesar do ganho em desempenho, que geralmente é obtido quando se utiliza *ensembles*, a combinação de classificadores simbólicos resulta em um classificador final não simbólico. Dois exemplos bem conhecidos de *ensemble* são Bagging e Booting.

Bagging Breiman (1996) é a técnica mais antiga e simples para criar um *ensemble* de classificadores. Essa técnica utiliza voto majoritário para combinar predições de classificadores individuais e aplica a classe mais frequentemente predita como a classificação final.

Diferente de Bagging, em Boosting Schapire (1990), cada exemplo de treinamento é associado a um peso. Este peso está relacionado com a taxa de acerto da hipótese induzida para aquele exemplo particular. Uma hipótese é induzida por iteração e os pesos associados com cada exemplo deve ser modificado.

Uma segunda abordagem para combinar classificadores é integrar o conhecimento gerado

por diferentes classificadores em uma única base de conhecimento e, então, utilizar um método de seleção de regras para criar um classificador. Em Prati and Flach (2005), os autores propuseram um algoritmo denominado ROCCER para selecionar regras baseado no desempenho das regras sobre o espaço ROC. Outra técnica que utiliza uma abordagem semelhante é o algoritmo GARSS Batista et al. (2006). GARSS utiliza um algoritmo evolutivo para selecionar regras que maximizam a AUC. Ambos, ROCCER e GARSS, são utilizados no contexto de classificação associativa. Outros trabalhos que utilizam algoritmo evolutivo para seleção de regras que combinam conhecimento de uma grande base de conhecimento podem ser encontrados em Ghosh and Nath (2004); Bernardini et al. (2008).

A metodologia geral de induzir diversos classificadores de várias amostras de dados e integrar o conhecimento destes classificadores em um classificador final foi inicialmente proposta por Fayyad, Djorgovski, e Weir Fayyad et al. (1996). Esta metodologia foi implementada no sistema RULER e utilizada no projeto SKICAT, cujo objetivo foi catalogar e analizar objetos de imagens digitalizadas do céu. De acordo com os autores, essa metodologia foi capaz de gerar um conjunto de regras robusto. Além disso, o classificador induzido foi mais preciso do que astrônomos para classificar objetos cósmicos de fotografias. A metodologia adotada no sistema RULER foi estendida no sistema XRULER (eXtended RULER) Baranauskas and Monard (2003) para utilizar conhecimento induzido de diferentes algoritmos.

3 Abordagem Proposta

A abordagem proposta em Milaré et al. (2009b) utiliza um algoritmo evolutivo para selecionar regras com o objetivo de maximizar a AUC para problemas que envolvam classes desbalanceadas. Algoritmos evolutivos são algoritmos de busca baseados na seleção natural e genética Goldberg (1989). Seu funcionamento envolve um conjunto de soluções potenciais (população), geralmente codificadas como uma sequência de bits (cromossomos). A evolução é realizada pela aplicação de um conjunto de transformações (geralmente, os operadores genéticos *crossover* e mutação), e avaliação da qualidade (*fitness*) das soluções.

Como previamente descrito, a abordagem é baseada na técnica *under-sampling*. Entretanto, para reduzir a probabilidade de perda de informação quando alguns exemplos são descar-

tados, a abordagem cria diversos conjuntos de treinamento. Dado um conjunto de treinamento $\mathcal{T} = \mathcal{T}^+ \cup \mathcal{T}^-$, no qual $\mathcal{T}^+$ é o conjunto de exemplos positivos (minoritário), e $\mathcal{T}^-$ é o conjunto de exemplos negativos (majoritário), n amostras randômicas $\mathcal{T}_1^-, \ldots, \mathcal{T}_n^-$ são criadas de $\mathcal{T}^-$. Cada amostra randômica $\mathcal{T}_i^-$ é uma amostra sem reposição de $\mathcal{T}^-$ e possui o mesmo número de exemplos do conjunto de exemplos positivos, isto é, $|\mathcal{T}_i^-| = |\mathcal{T}^+|$, $1 \leq i \leq n$.

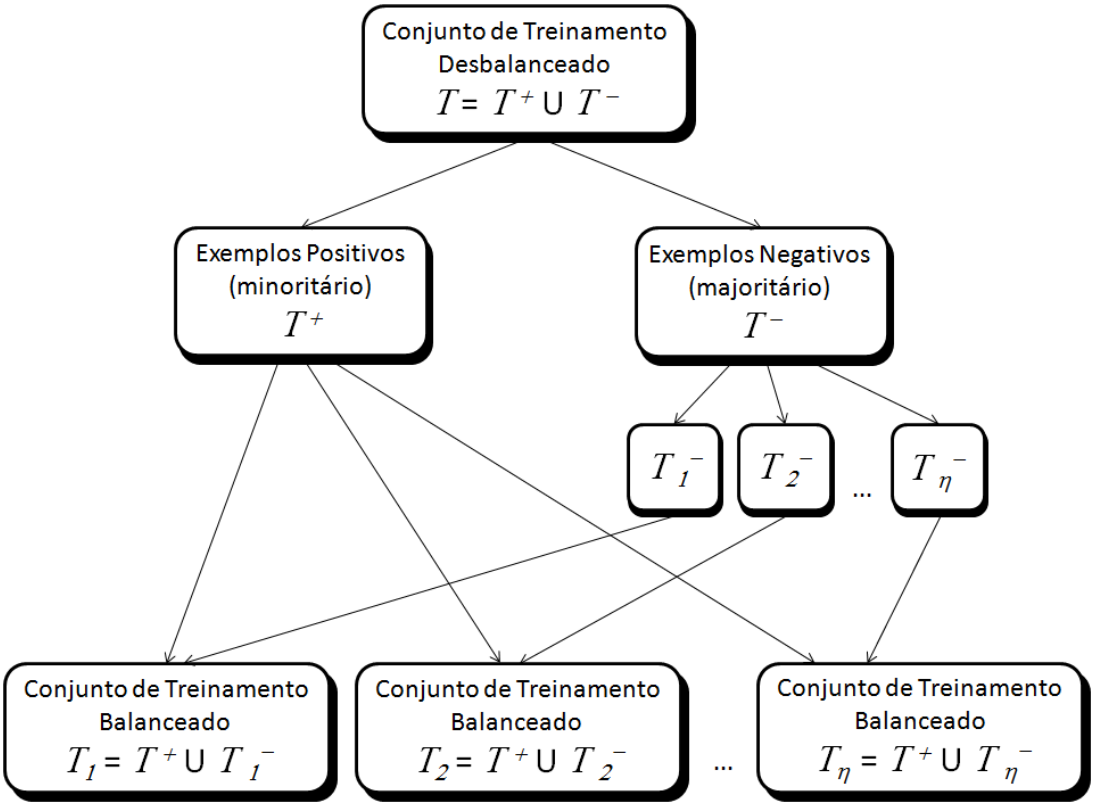

Figura 1: Abordagem utilizada para criar os conjuntos de treinamento balanceados.

No total, n conjuntos de treinamento balanceados T_i são criados pela junção de $\mathcal{T}^+$ com cada $\mathcal{T}_i^-$, isto é, $\mathcal{T}_i = \mathcal{T}^+ \cup \mathcal{T}_i^-$, $1 \leq i \leq n$, como pode ser observado na Figura 1. O valor do parâmetro n nos experimentos foi definido igual a 100. Os conjuntos de regras foram induzidos de cada conjunto de treinamento $\mathcal{T}_i$, como mostrado na Figura 2. As regras de todos os conjuntos de regras de cada indutor foram integradas em um único repositório de regras, e as regras repetidas foram descartadas. Nos experimentos realizados para avaliar a abordagem proposta, os algoritmos de AM utilizados inicialmente para induzir os conjuntos de regras foram C4.5Rules Quinlan (1988) e Ripper Cohen (1995) e posteriormente os experimentos foram estendidos com a utilização do CN2 Clark and Niblett (1989).

O repositório de regras é fornecido como entrada ao algoritmo evolutivo. Uma chave primária (um número natural) é associada com cada regra do repositório. Portanto, cada regra

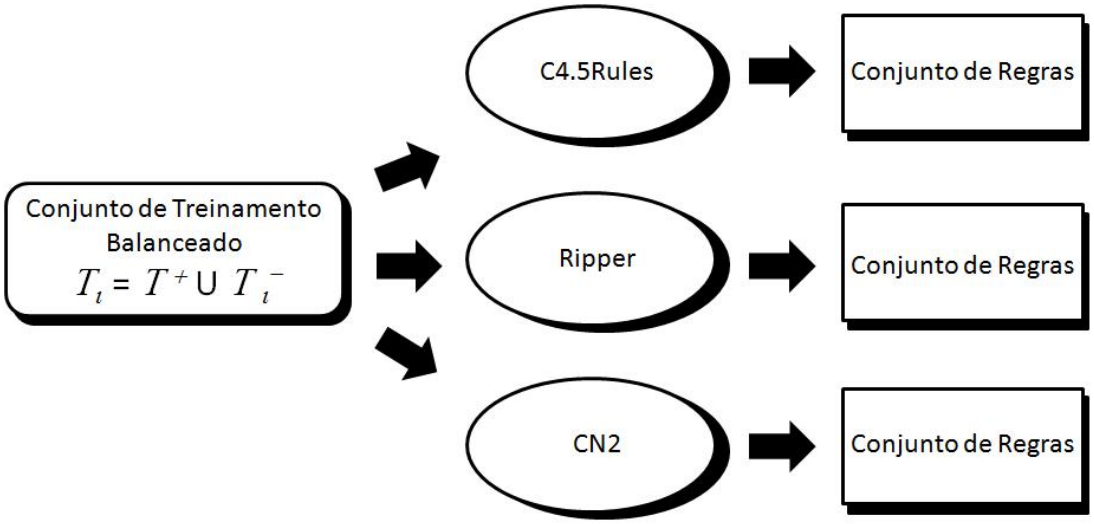

Figura 2: Indução dos conjuntos de regras.

pode ser acessada independentemente por esta chave. O método utiliza a abordagem Pittsburgh Smith (1980) para codificar classificadores como cromossomos. A abordagem de Pittsburgh é caracterizada por utilizar um classificador como um indivíduo Freitas (2002). Portanto, a população é um conjunto de classificadores. Então, neste trabalho, um vetor de chaves é utilizado para representar um cromossomo, isto é, um conjunto de regras é interpretado como um classificador, como é mostrado na Figura 3. Na implementação utilizada, a população inicial é randomicamente composta por 40 cromossomos. A função de avaliação utilizada é a métrica AUC medida sobre os exemplos de treinamento. O método de seleção é a seleção proporcional, no qual o número de vezes que um cromossomo é esperado reproduzir é proporcional ao seu *fitness*. Um operador de *crossover* simples foi aplicado com probabilidade de 0.4. O operador de mutação altera o valor de elementos do cromossomo randomicamente selecionados, ou seja, uma regra randomicamente selecionada do cromossomo é trocada por outra regra também randomicamente selecionada da base de regras. O operador de mutação foi aplicado com probabilidade de 0.1. As taxas com que os operadores de mutação e *crossover* são aplicados foram escolhidas baseadas na experiência prévia com algoritmos evolutivos Milaré et al. (2004). O número de gerações é limitado em 20. Finalmente, a implementação utiliza um operador de elitismo para reposição da população. De acordo com este operador, o melhor cromossomo de cada população é preservado para a próxima geração.

Como mencionado anteriormente, cada cromossomo representa um classificador. Tipicamente, a abordagem de Pittsburg permite indivíduos de tamanhos variáveis que podem ter seus

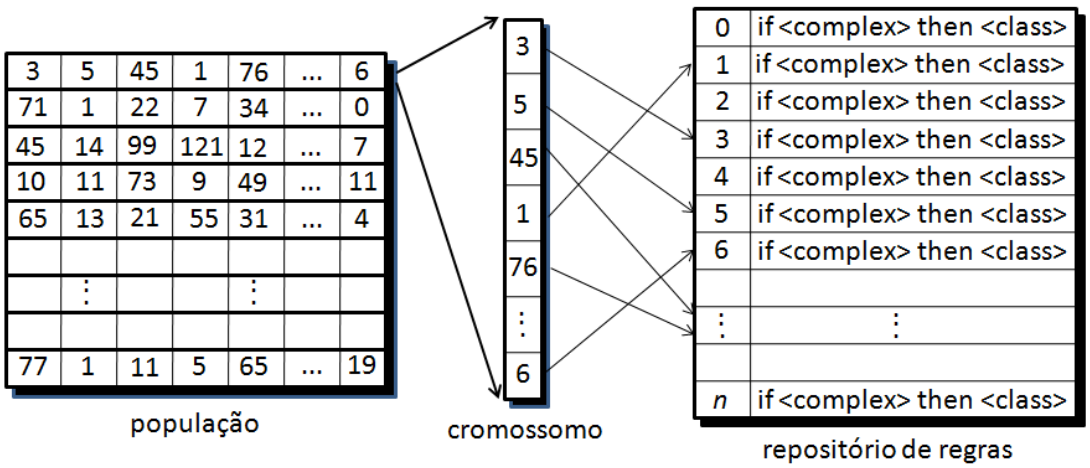

Figura 3: Abordagem utilizada para codificar os classificadores como cromossomos.

tamanhos modificados pela aplicação de *crossover* de dois pontos. Na abordagem proposta, o número de regras de cada cromossomo foi fixado para evitar *overfitting*. Portanto, na implementação realizada é permitido apenas *crossover* simples, ou seja, de apenas um ponto. Como a função de *fitness* é a métrica AUC sobre o conjunto de treinamento e se fosse permitido que o cromossomo crescesse durante a execução do algoritmo evolutivo, os cromossomos poderiam possuir muitas regras e não generalizar bem. Nos experimentos, o número de regras de cada cromossomo difere para cada conjunto de dados. Este número é aproximadamente o número médio de regras obtidas pelos classificadores induzidos sobre cada conjunto $\mathcal{T}_i$. Quando este número é grande, optou-se por limitar o tamanho do cromossomo em no máximo 40 regras por questões de tempo de processamento.

4 Avaliação Experimental

Como descrito anteriormente, alguns experimentos foram realizados, utilizando os sistemas de AM C4.5Rules Quinlan (1988), Ripper Cohen (1995) e CN2 Clark and Niblett (1989), para avaliar a abordagem híbrida proposta. Os experimentos foram realizados sob nove conjuntos de dados de *benchmark* coletados do repositório UCI Asuncion and Newman (2007), e três conjuntos de dados do "mundo real": Mammography Chawla et al. (2002); Oil-spill Kubat et al. (1998b); and Hoar-frost Bucene (2008). Esses conjuntos de dados são relacionados a problemas de classificação de diferentes domínios de aplicação. Na Tabela 1 são resumidas

as principais características dos conjuntos de dados. Essas características são: Identificador – identificador do conjunto de dados utilizado no texto; #Exemplos – número total de exemplos; #Atributos(quanti., quali.) – número de atributos e número de atributos quantitativos e qualitativos; Classes (min., maj.) – rótulo das classes minoritária e majoritária; e Classes % (min., maj.) – porcentagem das classes minoritária e majoritária. Os conjuntos de dados na Tabela 1 estão listados em ordem crescente de grau de desbalanceamento. Conjuntos de dados com mais de duas classes foram transformados em problemas de classificação binário tornando uma das classes como classe minoritária (como indicado na coluna Classes) e concatenando as outras classes como classe majoritária.

Tabela 1: Descrição dos conjuntos de dados.

Identificador	#Exemplos	#Atributos (quanti., quali.)	Classes (min., maj.)	Classes % (min., maj.)
CMC	1473	9 (2, 7)	(1, restante)	(42.73%, 57.27%)
Pima	768	8 (8, 0)	(1, 0)	(34.89%, 65.11%)
Yeast	1484	8 (8, 0)	(NUC, remaining)	(28.90%, 71.10%)
Blood	748	4 (4, 0)	(1, 0)	(24.00%, 76.00%)
Vehicle	946	18 (18, 0)	(van, restante)	(23.89%, 76.11%)
Flare	1066	10 (2, 8)	(C-class, restante)	(17.07%, 82.93%)
Page-blocks	5473	10 (10, 0)	(restante, text)	(10.22%, 89.78%)
Satimage	6435	36 (36, 0)	(4, restante)	(9.73%, 90.27%)
Hoar-frost	3044	236 (200, 36)	(positive, negative)	(6.11%, 93.89%)
Oil-Spill	937	48 (48, 0)	(2, 1)	(4.38%, 95.62%)
Abalone	4177	8 (7, 1)	(15, restante)	(2.47%, 97.53%)
Mammography	11183	6 (6, 0)	(2, 1)	(2.32%, 97.68%)

Cada um dos doze conjuntos de dados foi dividido em 10 pares de conjuntos de treinamento e teste utilizando amostragem randômica estratificada. Os exemplos de treinamento são 75% do conjunto de dados original e o conjunto de teste 25%. Dentro de cada partição, 100 conjuntos de dados balanceados foram criados com todos os exemplos da classe minoritária e uma amostra randômica dos exemplos da classe majoritária, como descrito previamente.

Esses conjuntos de dados balanceados foram fornecidos aos sistemas de AM C4.5Rules, Ripper e CN2. Estes algoritmos foram executados com seus parâmetros *default*. As regras de classificação geradas para cada conjunto de treinamento e para cada um dos algoritmos de AM foram combinadas em um repositório de regras. Um algoritmo evolutivo foi utilizado para selecionar um subconjunto de regras e construir um classificador a partir do repositório de regras, como mostrado na Figura 4.

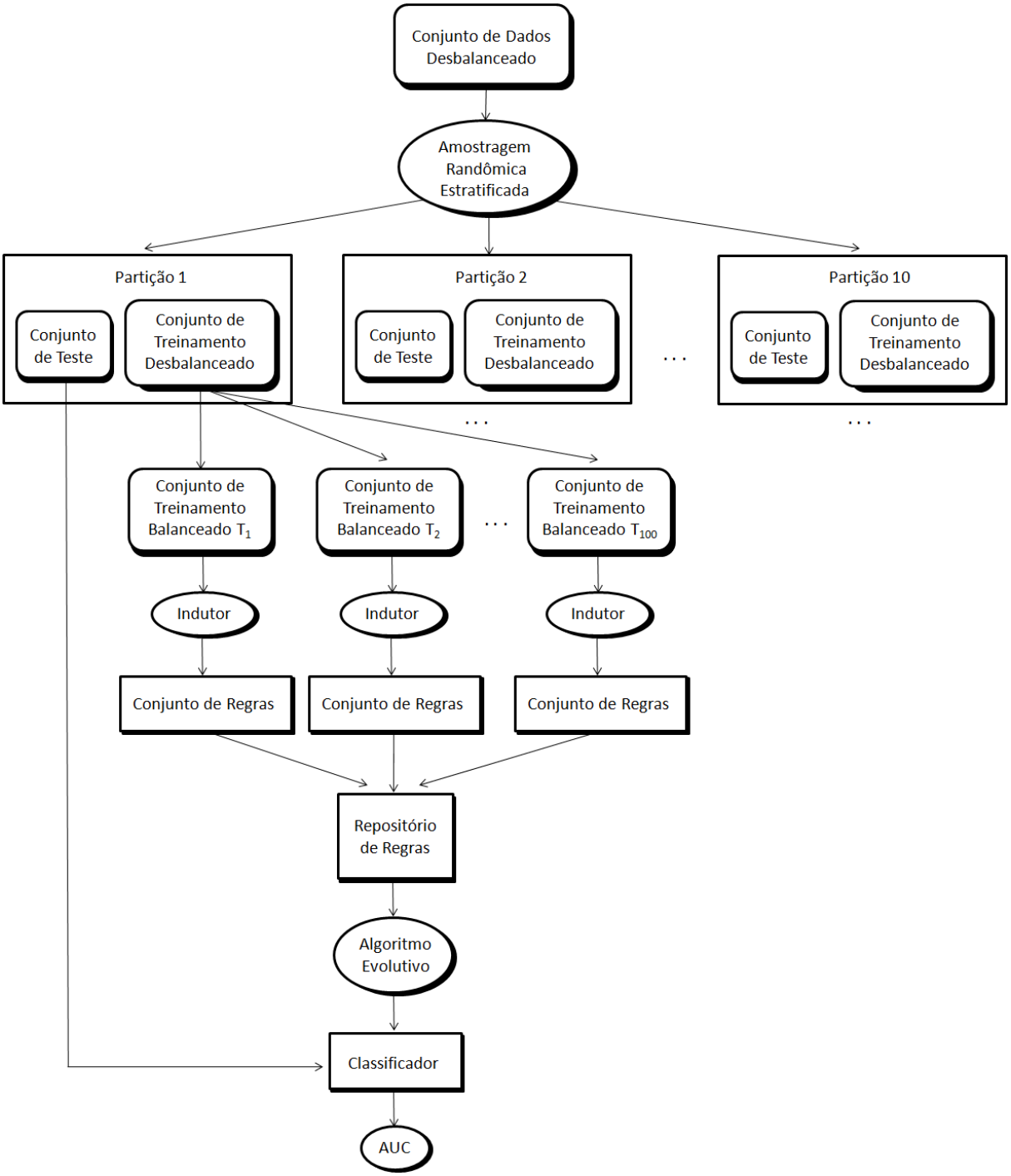

Figura 4: Experimento realizado.

Como descrito na Seção 3, o tamanho do cromossomo, isto é, o número de regras de um classificador individual foi definido como o tamanho médio dos classificadores gerados pelos indutores C4.5Rules, Ripper e CN2. Entretanto, quando o tamanho médio era muito grande, fato que ocorreu para os conjuntos de regras gerados pelo indutor CN2, o tamanho dos indivíduos (cromossomos) foi limitado a no máximo 40 por questões de tempo de processamento. A Tabela 2 mostra o tamanho dos cromossomos utilizados pelo algoritmo evolutivo para cada combinação

de conjunto de dados e indutor.

Tabela 2: Tamanho dos cromossomos.

Conjunto de Dados	C4.5Rules	Ripper	CN2
CMC	20	8	20
Pima	8	4	20
Yeast	10	4	20
Blood	4	4	20
Vehicle	12	6	12
Flare	8	4	20
Page-blocks	15	10	20
Satimage	20	6	40
Hoar-frost	10	10	24
Oil-spill	6	4	10
Abalone	20	6	40
Mammography	10	6	20

Na Tabela 3 são apresentados os resultados obtidos com o indutor C4.5Rules. Todos os resultados são valores médios de AUC calculados sobre os 10 pares de conjuntos de treinamento e teste. Os resultados estão divididos em três colunas. Na coluna C4.5Rules são apresentados os resultados obtidos com o indutor C4.5Rules executado sobre os dados desbalanceados; na coluna Under-sampling são apresentados os resultados obtidos pelo C4.5Rules sobre o conjunto de dados balanceado obtido pela técnica *under-sampling*; e, na coluna EA-C4.5Rules são apresentados os resultados obtidos pela abordagem híbrida proposta. Os melhores resultados estão em negrito. Como pode ser observado, EA-C4.5Rules apresenta valor médio de AUC mais alto para oito dos doze conjuntos de dados.

Tabela 3: Valor da AUC com o indutor C4.5Rules.

Data Set	C4.5Rules	Under-sampling	EA-C4.5Rules
CMC	**71.78 (01.39)**	70.39 (01.97)	68.77 (02.87)
Pima	74.28 (04.13)	75.80 (01.40)	**77.06 (03.52)**
Yeast	74.98 (02.09)	73.69 (02.20)	**76.47 (02.83)**
Blood	**71.22 (02.71)**	69.04 (03.13)	71.01 (03.58)
Vehicle	**96.96 (01.62)**	95.89 (01.61)	96.22 (02.33)
Flare	72.09 (03.44)	**72.70 (04.15)**	69.71 (04.33)
Page-blocks	97.26 (01.42)	96.97 (00.84)	**97.56 (01.09)**
Satimage	86.27 (02.69)	87.96 (01.33)↓	**90.94 (01.30)** ↑
Hoar-frost	86.37 (04.82)	89.58 (02.24)	**92.08 (02.98)**
Oil-spill	77.89 (10.48)	**84.03 (04.31)**	**84.03 (08.06)**
Abalone	77.93 (02.23)	78.31 (02.27)	**80.95 (00.91)**
Mammography	87.20 (03.70)	89.43 (02.55)	**90.30 (02.87)**

Para verificar se a diferença dos resultados obtidos entre as abordagens (C4.5Rules, Under-

sampling e EA-C4.5Rules) é significativa ou não com 95% de confiabilidade, foi utilizado o teste de hipótese *k-fold cross-validated pairet t* Dietterich (1997a). Neste teste comparou-se os resultados obtidos pela abordagem EA-C4.5Rules e C4.5Rules, EA-C4.5Rules e Under-sampling e Under-sampling e C4.5Rules. Na Tabela 3 o símbolo ↑ ao lado de uma abordagem representa um resultado significativo em relação à abordagem que possui ao lado um símbolo ↓.

Pode-se observar na Tabela 3 que há apenas um resultado significativo. Para o conjunto de dados Satimage, a abordagem EA-C4.5Rules obteve um resultado significativo com 95% de confiabilidade em relação à abordagem Under-sampling. Isto significa que a abordagem EA-C4.5Rules é melhor do que a abordagem Under-sampling apenas para o conjunto de dados Satimage.

Na Tabela 4 são apresentados os resultados obtidos com o indutor Ripper. Como descrito previamente para a tabela anterior, na coluna Ripper é mostrado os resultados obtidos pelo indutor Ripper sobre os dados desbalanceados; na coluna Under-sampling são apresentados os resultados obtidos pelo Ripper sobre o conjunto de dados balanceado obtido pela técnica *under-sampling*; e, na coluna EA-Ripper são apresentados os resultados obtidos pela abordagem híbrida proposta. Pode ser observado que EA-Ripper apresenta o valor de AUC mais para todos os conjuntos de dados.

Tabela 4: Valor da AUC com o indutor Ripper.

Data Set	Ripper	Under-sampling	EA-Ripper
CMC	68.64 (02.27)	68.73 (03.18)	**70.34 (02.08)**
Pima	69.98 (02.21) ↓	74.37 (04.16)	**76.94 (02.21)** ↑
Yeast	65.99 (02.13) ↓	69.71 (02.39)	**74.01 (02.55)** ↑
Blood	63.34 (03.69) ↓	67.87 (02.11)	**73.42 (04.32)** ↑
Vehicle	92.21 (02.55)	92.94 (03.06)	**95.94 (01.54)**
Flare	56.94 (02.25) ↓ ⇓	68.98 (03.63) ⇑	**69.40 (04.37)** ↑
Page-blocks	92.41 (01.31) ↓ ⇓	95.10 (00.80) ⇑ ↓	**96.71 (00.47)** ↑
Satimage	75.81 (01.60) ↓ ⇓	86.97 (01.92) ⇑ ↓	**91.11 (01.11)** ↑
Hoar-frost	80.97 (03.76) ↓ ⇓	88.87 (02.32) ⇑	**92.05 (02.93)** ↑
Oil-spill	68.05 (06.80)	77.68 (08.59)	**82.66 (07.27)**
Abalone	59.91 (02.49) ↓ ⇓	75.30 (03.88) ⇑	**81.06 (02.84)** ↑
Mammography	78.20 (02.68) ↓ ⇓	88.73 (01.82) ⇑	**91.97 (02.68)** ↑

Para verificar se a diferença dos resultados obtidos entre as abordagens (Ripper, Under-sampling e EA-Ripper) é significativa ou não com 95% de confiabilidade, foi utilizado o teste de hipótese *k-fold cross-validated pairet t* Dietterich (1997a). Neste teste comparou-se os resultados

obtidos pela abordagem EA-Ripper e Ripper, EA-Ripper e Under-sampling e Under-sampling e Ripper.

Na Tabela 4 os símbolos $\uparrow$ e $\Uparrow$ ao lado da abordagem representa um resultado significativo em relação à abordagem que possui ao lado um símbolo $\downarrow$ ou $\Downarrow$. Por exemplo, pode ser observado pela Tabela 4 que, para o conjunto de dados Satimage, há o símbolo $\uparrow$ ao lado do resultado obtido pela abordagem EA-Ripper e há o símbolo $\downarrow$ ao lado dos resultados obtidos pelas abordagens Under-sampling e Ripper. Isto significa que a abordagem EA-Ripper obteve um resultado estatiticamente melhor, com com 95% de confiabilidade, em relação às abordagens Under-sampling e Ripper. Ainda, para o mesmo conjunto de dados, o símbolo $\Uparrow$ ao lado do resultado obtido pela abordagem Under-sampling e o símbolo $\Downarrow$ ao lado do resultado obtido pela abordagem Ripper representa que a abordagem Under-sampling obteve um resultado estatisticamente melhor, com com 95% de confiabilidade, em relação à abordagem Ripper.

Pode-se observar na Tabela 4 que a abordagem EA-Ripper obteve dois resultados significativos em relação à abordagem Under-sampling para os conjuntos de dados Page-blocks e Satimage. Ainda, a abordagem EA-Ripper obteve nove resultados significativos em relação à abordagem Ripper para os conjuntos de dados Pima, Yeast, Blood, Flare, Page-blocks, Satimage, Hoar-frost, Abalone e Mammography. A abordagem Under-sampling obteve seis resultados significativos em relação à abordagem Ripper para os conjuntos de dados Flare, Page-blocks, Satimage, Hoar-frost, Abalone e Mammography.

Na Tabela 5 são apresentados os resultados obtidos com o indutor CN2. Na coluna CN2 são apresentados os resultados obtidos com o indutor CN2 executado sobre os dados desbalanceados; na coluna Under-sampling são apresentados os resultados obtidos pelo CN2 sobre o conjunto de dados balanceado obtido pela técnica *under-sampling*; e, na coluna EA-CN2 são apresentados os resultados obtidos pela abordagem híbrida. Os melhores resultados estão em negrito. Como pode ser observado, EA-CN2 e Under-sampling apresentam cada um valor médio de AUC mais alto para cinco dos conjuntos de dados, e CN2 para dois conjuntos de dados. No entanto, EA-CN2 apresenta melhores resultados para conjuntos de dados com mais alto grau de desbalanceamento.

Nos experimentos realizados com o indutor CN2 não foi utilizado o conjunto de dados Yeast, como pode ser observado na Tabela 5. Isto porque os valores de AUC para os conjuntos

de regras obtidos pelo CN2 foram em torno de 50.00, ou seja, não foi possível aprender.

Tabela 5: Valor da AUC com o indutor CN2.

Conjunto de Dados	CN2	Under-sampling	EA-CN2
CMC	64.29 (02.35) ⇑	**65.33 (01.31)** ↑	58.33 (01.63) ↓ ⇓
Pima	**78.81 (02.39)**	78.77 (02.95)	77.64 (02.59)
Blood	63.37 (04.12)	**66.05 (04.81)**	65.90 (02.90)
Vehicle	95.97 (02.43)	96.56 (01.60)	**96.65 (02.17)**
Flare	61.16(04.19)	**66.63 (03.67)**	65.32 (02.81)
Page-blocks	96.12 (01.11)	95.42 (00.82)	**97.08 (00.91)**
Satimage	90.59 (00.83	**90.67 (00.67)**	89.97 (01.53)
Hoar-frost	91.08 (03.70)	92.59 (02.24)	**92.81 (03.20)**
Oil-Spill	**84.25 (07.07)**	82.83 (05.76)	81.73 (09.22)
Abalone	65.95 (02.41)	61.56 (06.69)	**69.94 (02.74)**
Mammography	87.05 (08.67)	90.60 (03.19)	**91.57 (01.77)**

Da mesma forma como descrito anteriormente, o teste de hipótese *k-fold cross-validated pairet t* para verificar se a diferença dos resultados obtidos entre as abordagens (CN2, Under-sampling e EA-CN2) é significativa ou não, com 95% de confiabilidade. Na Tabela 5 o símbolo ↑ e ⇑ ao lado da abordagem representa um resultado significativo em relação à abordagem que possui ao lado um símbolo ↓ ou ⇓ .

Pode ser observado na Tabela 5 que a abordagem Under-sampling obteve um resultado significativo em relação à abordagem EA-CN2 para o conjunto de dados CMC. A abordagem CN2 também apresentou um resultado significativo em relação à abordagem EA-CN2 para o conjunto de dados CMC.

Para analizar se há diferença estatisticamente significante entre as abordagens também foi realizado o teste de Friedman[1]. O teste de Friedman foi executado com três hipóteses nulas diferentes para comparar o desempenho das seguintes abordagens:

1. C4.5Rules, C4.5Rules com *under-sampling* e EA-C4.5Rules;
2. Ripper, Ripper com *under-sampling* e EA-Ripper;
3. CN2, CN2 com *under-sampling* e EA-CN2.

Quando a hipótese nula é rejeitada pelo teste de Friedman, com 95% de confiabilidade, pode-se proceder com um teste *pos-hoc* para detectar quais diferenças entre as abordagens são

[1]O teste de Friedman é um teste não paramétrico equivalente ao teste de ANOVA para múltiplas comparações. Ver Demšar (2006) para uma discussão mais detalhada e completa a respeito de testes estatísticos em AM.

significantes. Para isto, foi executado o teste Bonferroni-Dunn para comparações múltiplas como um teste de controle.

Em relação ao desempenho da abordagem EA-C4.5Rules, a primeira hipótese nula não foi rejeitada pelo teste de Friedman. Portanto, não é possível apontar qualquer diferença entre C4.5Rules, C4.5Rules com *under-sampling* e EA-C4.5Rules. Este resultado é claramente um resultado negativo para a aboradagem EA-C4.5Rules. Entretanto, uma análise mais detalhada dos resultados mostra que para a maioria dos conjuntos de dados mais desbalanceados, a abordagem EA-C4.5Rules apresentou bom desempenho. Considerando os conjuntos de dados com classe minoritária abaixo de 10% (aproximadamente) do número total de casos, isto é, os conjuntos de dados Flare, Page-blocks, Satimage, Hoar-frost, Oil-spill, Abalone and Mammography, EA-C4.5Rules sempre apresenta os valores mais altos para AUC. Há somente uma exceção, o conjunto de dados Oil-spill, em que EA-C4.5Rules e Under-sampling apresentam o mesmo valor AUC, mas Under-sampling possui variância menor.

A segunda hipótese nula foi rejeitada pelo teste de Friedman com 95% de confiança. Na Figura 5 são mostrados os resultados do teste Bonferroni-Dunn utilizando a abordagem EA-Ripper como teste de controle. Bonferroni-Dunn indica que a abordagem EA-Ripper é melhor do que as abordagens Ripper e Ripper aliado a *under-sampling* com 95% de confiança.

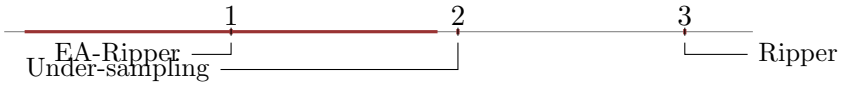

Figura 5: Resultados do teste Bonferroni-Dunn para Ripper. A linha espessa marca o intervalo de uma diferença crítica com 95% de confiabilidade. A diferença crítica é 0.91.

Em relação ao desempenho da abordagem EA-CN2, a terceira hipótese nula não foi rejeitada pelo teste de Friedman. Portanto, não é possível apontar qualquer diferença entre CN2, CN2 com *under-sampling* e EA-CN2.

5 Conclusão e Trabalhos Futuros

Neste relatório técnico foi descrita a abordagem híbrida proposta em Milaré et al. (2009b) bem como os experimentos realizados para avaliá-la publicados em Milaré et al. (2009a, 2010). A abordagem proposta descrita busca resolver o problema de indução de regras de classificação

em conjuntos de dados desbalanceados. Esta abordagem combina indutores simbólicos de AM e algoritmos evolutivos. Nesta abordagem é utilizado um algoritmo evolutivo para realizar uma busca mais extensiva sob o espaço de hipótese.

Na avaliação experimental utilizando os indutores C4.5Rules e Ripper os resultados obtidos foram bastante promissores. A abordagem EA-Ripper apresentou resultados estatisticamente significantes comparado às abordagens Ripper e Ripper com *under-sampling* para o teste de Friedman. A Abordagem EA-C4.5Rules não obteve resultados estatisticamente significantes quando comparado às abordagens C4.5Rules e C4.5Rules com *under-sampling*, mas mesmo assim apresentou bons resultados.

Quando o indutor CN2 foi utilizado para gerar classificadores de conjuntos de dados balanceados, os resultados obtidos não foram tão promissores quanto os resultados obtidos utilizando os indutores C4.5rules e Ripper para gerar os classificadores dos conjuntos de dados balanceados. Para alguns conjuntos de dados utilizados nos experimentos, os classificadores gerados pelo CN2 eram muito grandes, às vezes com mais de 100 regras. Nestes casos, optou-se por um tamanho menor dos cromossomos (classificadores) do algoritmo evolutivo. Portanto, os classificadores encontrados pela abordagem híbrida, para muitos conjuntos de dados, são bem menores do que os classificadores gerados pelo CN2 e pelo método Under-sampling, o que pode ter contribuído com os resultados não tão bons obtidos pela abordagem híbrida.

Como trabalhos futuros pretendemos investigar novas métricas para compor regras para gerar um classificador. Estas métricas indicam quais regras devem disparar nos casos em que múltiplas regras cobrem um exemplo. Estas métricas possuem uma influência direta sobre o desempenho dos classificadores e novas métricas, projetadas para classes desbalanceadas, podem potencialmente melhorar a classificação sobre conjuntos de dados desbalanceados.

Referências

Asuncion, A. and D. Newman (2007). UCI machine learning repository. 8

Baranauskas, J. A. and M. C. Monard (2003). Combining symbolic classifiers from multiple inducers. *Knowledge Based Systems 16*(3), 129–136. Elsevier Science. 5

Batista, G. E. A. P. A., C. R. Milaré, R. C. Prati, and M. C. Monard (2006). A comparison of methods for rule subset selection applied to associative classification. *Inteligencia Artificial* (32), 29–35. 5

Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard (2004). A study of the behavior of several

methods for balancing machine learning training data. *SIGKDD Explorations Newsletter 6*(1), 20–29. Special issue on Learning from Imbalanced Datasets. 3

Bernardini, F. C., R. C. Prati, and M. C. Monard (2008). Evolving sets of symbolic classifiers into a single symbolic classifier using genetic algorithms. In *International Conference on Hybrid Intelligent Systems*, Washington, DC, USA, pp. 525–530. IEEE Computer Society. 5

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140. 4

Bucene, L. C. (2008). Mineração de Dados Climáticos para Alertas de Geada e Deficiência Hídrica. PhD Thesis, FEAGRI/UNICAMP. 1, 8

Chan, P. K. and S. J. Stolfo (1998). Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In *International Conference on Knowledge Discovery and Data Mining*, pp. 164–168. 3

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357. 8

Clark, P. and T. Niblett (1989). The CN2 Induction Algorithm. *Machine Learning 3*, 261–284. 2, 6, 8

Cohen, W. (1995). Fast effective rule induction. In *International Conference on Machine Learning*, pp. 115–123. 2, 6, 8

Cohena, G., M. Hilariob, H. Saxc, S. Hugonnetc, and A. Geissbuhler (2006). Learning from imbalanced data in surveillance of nosocomial infection. *Intelligent Data Analysis in Medicine 37*(1), 7–18. 1

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7*, 1–30. 14

Dietterich, T. G. (1997a). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation 10*(7), 1895–1924. 12

Dietterich, T. G. (1997b). Machine Learning Research: Four Current Directions. Technical report, Oregon State University. 4

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Labs. 2

Fawcett, T. and F. J. Provost (1997). Adaptive fraud detection. *Journal of Data Mining and Knowledge Discovery 1*(3), 291–316. 3

Fayyad, U. M., G. Piatetsky-Shapiro, and P. Smyth (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM 39*(11), 27–34. 5

Freitas, A. A. (2002). *Data Mining and Knowledge Discovery with Evolutionary Algorithms*. Springer-Verlag. 7

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119–139. 4

Ghosh, A. and B. Nath (2004). Multi-objective rule mining using genetic algorithms. *Information Sciences 163*(1-3), 123–133. 5

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley. 5

Japkowicz, N. and S. Stephen (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis 6*(5), 429–449. 3

Kubat, M., R. Holte, and S. Matwin (1998a). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning 30*, 195–215. 3

Kubat, M., R. C. Holte, and S. Matwin (1998b). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning 30*(2-3), 195–215. 8

Kubat, M. and S. Matwin (1997). Addressing the course of imbalanced training sets: One-sided selection. In *International Conference in Machine Learning*, pp. 179–186. Morgan Kaufmann. 3

Ling, C. X. and C. Li (1998). Data mining for direct mining: Problems and solutions. In *International Conference on Knownledge Discovery and Data Mining*, pp. 73–79. 3

Liu, X. Y., J. Wu, and Z. H. Zhou (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39*(2), 539–550. 4

Milaré, C. R., G. E. A. P. A. Batista, and A. C. P. L. F. Carvalho (2009a). Avaliação de uma abordagem híbrida para aprender com classes desbalanceadas: Resultados experimentais com o indutor cn2. In *IV Congresso da Academia Trinacional de Ciências*. 2, 15

Milaré, C. R., G. E. A. P. A. Batista, and A. C. P. L. F. Carvalho (2009b). A hybrid approach to learn with imbalanced classes using evolutionary algorithms. In *Proc. 9th International Conference Computational and Mathematical Methods in Science and Engineering (CMMSE)*, Volume II, pp. 701–710. 1, 2, 3, 5, 15

Milaré, C. R., G. E. A. P. A. Batista, and A. C. P. L. F. Carvalho (2010). A hybrid approach to learn with imbalanced classes using evolutionary algorithms. *Logic Journal of the IGPL*. 2, 15

Milaré, C. R., G. E. A. P. A. Batista, A. C. P. L. F. Carvalho, and M. C. Monard (2004). Applying genetic and symbolic learning algorithms to extract rules from artificial neural neworks. In *Proc. Mexican International Conference on Artificial Intelligence*, Volume 2972 of *LNAI*, pp. 833–843. Springer-Verlag. 7

Opitz, D. and R. Maclin (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research 11*, 169–198. 4

Pazzani, M., C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk (1994). Reducing misclassification costs. In *International Conference in Machine Learning*, pp. 217–225. 3

Pednault, E. P. D., B. K. Rosen, and C. Apte (2000, March). Handling imbalanced data sets in insurance risk modeling. Technical Report RC-21731, IBM Research Report. 3

Phua, C., D. Alahakoon, and V. Lee (2004). Minority report in fraud detection: Classification of skewed data. *SIGKDD Explorations Newsletter 6*(1), 50–59. 1, 3

Prati, R. C. and P. A. Flach (2005). ROCCER: An algorithm for rule learning based on ROC analysis. In *International Joint Conference on Artificial Intelligence (IJCAI'2005)*, pp. 823–828. 5

Quinlan, J. R. (1988). *C4.5 Programs for Machine Learning*. Morgan Kaufmann Publishers, CA. 2, 6, 8

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning 5*(2), 197–227. 4

Smith, S. F. (1980). *A learning system based on genetic adaptive algorithms*. Ph. D. thesis, Pittsburgh, PA, USA. 7

Stolfo, S. J., D. W. Fan, W. Lee, A. L. Prodromidis, and P. K. Chan (1997). Credit card fraud detection using meta-learning: Issues and initial results. In *AAAI-97 Workshop on AI Methods in Fraud and Risk Management*, pp. 83–90. 3

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *SIGKDD Explorations 6*(1), 7–19. 3

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks 5*, 241–259. 4